



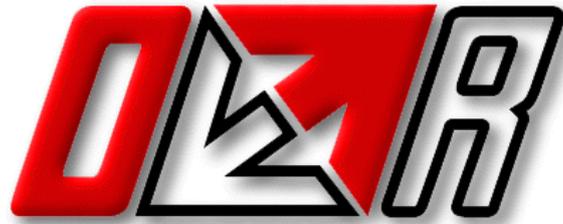
THE OPERATIONS RESEARCH SOCIETY OF NEW ZEALAND

46th ANNUAL CONFERENCE

10 – 11 December 2012

**Victoria University of Wellington
New Zealand**

Sponsored by
The Science Faculty, Victoria University of Wellington;
the Optima Corporation; Concept Consulting;
Orbit Systems; Derceto; Hoare Research Software;
Department of Engineering Science, University of Auckland



The Operations Research Society of New Zealand gratefully acknowledges the generous support of the following conference sponsors:



THE SCIENCE FACULTY, VICTORIA UNIVERSITY OF WELLINGTON



THE OPTIMA CORPORATION
Young Practitioner Prize Sponsor





ORBIT SYSTEMS



Sustainable Water Distribution

DERCETO



HOARE RESEARCH SOFTWARE



DEPARTMENT OF ENGINEERING SCIENCE, UNIVERSITY OF AUCKLAND

Operational Research Society of New Zealand (Inc.).
P.O. Box 6544, Wellesley Street, Auckland, New Zealand
www.orsnz.org.nz

Preface

The papers in this volume form the Proceedings of the 46th Annual Conference of the Operations Research Society of New Zealand (ORSNZ) held 10–11 December 2012 at Victoria University of Wellington, New Zealand.

Welcome to ORSNZ'12, hosted by the School of Mathematics, Statistics and Operations Research at Victoria University of Wellington. Special thanks to all those who have made this conference possible, including all the committee members below and especially the very hard working chairs, Mark Johnston and Stefanka Chukova. The invaluable assistance of Andrew Mason and Andrea Raith is also warmly appreciated.

As is evident in these proceedings, the conference programme covers a wide range of topics in Operations Research. The plenary speakers this year are Paul Reid from New Zealand Post here in Wellington and Grant Read from the University of Canterbury.

We are most grateful for the support received from our sponsors: the Science Faculty, Victoria University of Wellington; the Optima Corporation; Concept Consulting; Orbit Systems; Derceto; Hoare Research Software; the Department of Engineering Science, University of Auckland.

We hope you enjoy your time in Wellington, both at the conference and in taking advantage of all the city has to offer.

John Haywood
ORSNZ'12 Conference Committee
December 2012

Conference Committee:	Mark Johnston (Co-Chair) Stefanka Chukova (Co-Chair) John Haywood Nokuthaba Sibanda Yuichi Hirose
Proceedings Editor:	John Haywood
Administrative Support:	Ginny Whatarau and Kelly Shen

**The 46th ORSNZ Conference
10-11 December 2012,
Victoria University of Wellington, New Zealand
Scientific Programme**

Day 1: Monday 10 December 2012	
Concept Consulting Welcome Breakfast 8:30 – 10:00 (MC Foyer) and Registration	
Opening 10:00 – 10:15 (MCLT101)	
Paul Reid, <i>The Role of Operations Research in Determining Your Strategic Direction</i> Plenary talk 10:15 – 11:10 (MCLT101) Chair: M. Johnston	
Tea/Coffee Break 11:10 – 11:30 (MC Foyer)	
Room 1 (MCLT101) Topic: The Optima Corporation YPP session I Chair: A. Raith	
11:30 – 11:45	[47] Rosemary Read, Shane Dye and Grant Read. <i>Generalized CDDP for Reservoir Management</i>
11:45 – 12:00	[26] Michael Leon, Golbon Zakeri and Anthony Downward. <i>Simulating FTR Strategy in New Zealand Electricity Market</i>
12:00 – 12:15	[29] Kasper Tofte and Troels Martin Range. <i>A Time Indexed Model for the Elective Surgery Scheduling Problem</i>
12:15 – 12:30	[55] Matthew Crowder and Andrew Mason. <i>Districting for the New Zealand Census: MIP-Heuristic Approaches</i>
12:30 – 12:45	[36] Oliver Hinder. <i>Optimizing Clothing Catalogues for EziBuy</i>
Room 2 (MCLT102) ANSYSYS Workshop on Systems Thinking, Systems Modelling and Systems Practice Chair: B. Cavana	
11:30 – 12:00	[61] Marjan van Den Belt. <i>Mediated Modelling to Support Spatial Planning: Population Change, Inequality and City Attractiveness in Wellington</i>
12:00 – 12:20	[62] David Rees, Robert Cavana and Jacqueline Cumming. <i>Using Cognitive Mapping and Qualitative System Dynamics to Develop a Theory of Implementation in Primary Health Care</i>
12:20 – 12:45	Workshop on Systems Thinking, Systems Modeling and Systems Practice: <i>Discussion time</i>
Lunch 12:45 – 1:30 (MC Foyer)	
Room 1 (MCLT101) Topic: The Optima Corporation YPP session II Chair: M. Johnston	
1:30 – 1:45	[42] Simon Bull, Andrew Mason and Andrea Raith. <i>Scheduling Families of Jobs on Multiple Identical Machines to Minimize Total Tardiness</i>
1:45 – 2:00	[28] Olga Perederieieva. <i>Solving Bi-objective Traffic Assignment Based on Time Surplus Maximisation</i>
2:00 – 2:15	[30] Jingze Du, Matthias Ehrgott and Andrea Raith. <i>Optimal Delivery of Arc Modulated Radiation Therapy in Cancer Treatment</i>
2:15 – 2:30	[13] Keith Ruddell and Andrea Raith. <i>An Aggregational Approach to the Traffic Assignment Problem</i>
2:30 – 2:45	[22] Simon Anastasiadis and Stefanka Chukova. <i>Modeling Technology Adoption Decisions where Farmers are Resistant to Change</i>
2:45 – 3:00	[21] Simon Kristiansen and Thomas Stidsen. <i>Elective Course Student Sectioning at Danish High Schools</i>

Room 2 (MCLT102) ANZSYS Workshop on Systems Thinking, Systems Modelling and Systems Practice		Chair: Su-Wuen Ong
1:30 – 1:50	[8] Shanie Atkinson and Michael Shayne Gary. <i>Dynamics of Mergers & Acquisitions Integration</i>	
1:50 – 2:10	[39] Su-Wuen Ong, Robert Cavana and Mondher Sahli. <i>A Qualitative System Dynamics Analysis of Airline Safety in New Zealand</i>	
2:10 – 2:30	[64] Parvathy Muraleedharan and Arun Elias. <i>Modelling Offshore Outsourcing of Software Testing Services: A Telecom New Zealand Case Study</i>	
2:30 – 2:50	[35] Rodney Scott, Robert Cavana and Donald Cameron. <i>Evaluating the Impact of Systems Thinking Workshops on Strategy Implementation in a Government Department</i>	
2:50 – 3:00	Workshop on Systems Thinking, Systems Modeling and Systems Practice: <i>Discussion time</i>	
Tea/Coffee Break 3:00 – 3:30 (MC Foyer)		
Room 1 (MCLT101) Topic: The Optima Corporation YPP session III		Chair: J. Haywood
3:30 – 3:45	[32] Quan Zhou and Tava Olsen. <i>Developing a Rotation Scheme to Reduce Expiration for the Reserve Medical Supply</i>	
3:45 – 4:00	[24] Zhengliang Liu, Matthias Ehrgott and Andrea Raith. <i>Linear Optimization over the Nondominated Set of a Multiobjective Linear Programming Problem</i>	
4:00 – 4:15	[33] Kuan-Min Lin, John Simpson, Giuseppe Sasso, Andrea Raith and Matthias Ehrgott. <i>An Application of Data Envelopment Analysis to External Radiotherapy Treatment Planning</i>	
4:15 – 4:30	[57] Thiranja Babarenda Gamage, Martyn Nash and Poul Nielsen. <i>Optimal Design of Experiments to Determine Mechanical Properties of Soft Bodies</i>	
4:30 – 4:45	[31] Salah Al-Chanati, Golbon Zakeri and Anthony Downward. <i>Simulator for Electricity Related Investments</i>	
4:45 – 5:00	YPP Judges: <i>Discussion time</i>	
Room 2 (MCLT102) ANZSYS Workshop on Systems Thinking, Systems Modelling and Systems Practice		Chair: J. Velez-Castiblanco
3:30 – 3:50	[9] Miles Yang, Michael Shayne Gary and Philip Yetton. <i>Organizational Goals, Feedback Effects, and Performance</i>	
3:50 – 4:10	[10] Jorge Velez-Castiblanco. <i>Intervention as Language Games</i>	
4:10 – 4:30	[43] John Cody, Robert Cavana and David Pearson. <i>Limits to Collective Action – Development of an Evolutionary Game Model</i>	
4:30 – 4:50	[48] Robert Cavana, Kala Retna and Arthur Ahimbisibwe. <i>Structural Equation Modelling of Undergraduate Management Students' Perceptions of Feedback in a New Zealand University</i>	
4:50 – 5:00	Workshop on Systems Thinking, Systems Modeling and Systems Practice: <i>Discussion time</i>	
AGM 5:05 – 6:00 (MCLT101)		
Conference Dinner: James Cook Hotel, 147 The Terrace 6:30 – 10:30		

The 46th ORSNZ Conference
10-11 December 2012,
Victoria University of Wellington, New Zealand
Scientific Programme

Day 2: Tuesday 11 December 2012	
Room 1 (MCLT101) Topic: Stochastic OR	
Chair: I. Ziedins	
8:30 – 8:50	[40] Sima Varnosafaderani and Stefanka Chukova. <i>Modeling Repairs of Systems with a Bathtub-Shaped Failure Rate Function</i>
8:50 – 9:10	[4] Richard Arnold, Stefanka Chukova and Yu Hayakawa. <i>Inference for Multicomponent Systems with Dependent Failures</i>
9:10 – 9:30	[56] Ilze Ziedins. <i>Accumulating Priority Queues: A New Priority Scheme for Hospital Queues?</i>
9:30 – 9:50	[51] Aioporn Sophonsriduk. <i>The Application of Linear Programming to Select the Lowest Cost and Optimized Quality of Textile Wet Process</i>
Room 2 (MCLT102) Topic: Case Studies	
Chair: M. Ehrgott	
8:30 – 8:50	[7] Kevin Ross and Michael Freimer. <i>Contribution Margin Optimisation at Fonterra</i>
8:50 – 9:10	[45] Michelle Goodall and Grant Robinson. <i>Future Focused Network Modelling at New Zealand Post</i>
9:10 – 9:30	[67] Ali Broadbent. <i>Stakeholder Engagement in Capital Budgeting at Counties Manukau District Health Board</i>
9:30 – 9:50	[52] Guillermo Cabrera, Manuel Chica, Matthias Ehrgott and Andrew Mason. <i>Mathematical Programming and Metaheuristic Approaches Applied to Biological-Based Fluence Map Optimization in Radiotherapy</i>
Tea/Coffee Break 9:50 – 10:20 (MC Foyer)	
Grant Read, Economics and Operations Research: A Past, Present and Future Duality	
Plenary talk 10:20 – 11:20 (MCLT101) Chair: Shane Dye	
Room 1 (MCLT101) Topic: Special Session on Energy and Resource Markets Chair: J.Raffensperger	
11:20 – 11:40	[37] Anthony Downward, Golbon Zakeri, Zabin Farishta and Faisal Wahid. <i>Use of Hydro Resources for Irrigation and Electricity Production</i>
11:40 – 12:00	[49] Shane Dye, Grant Read, Rosemary Read and Stephen Starkey. <i>An Evaluation Tool for Reservoir Management</i>
12:00 – 12:20	[38] Indrajana Mahakalanda, Shane Dye, Grant Read and John Raffensperger. <i>Intra-period Market Clearing for a Multi-Use Catchment via CDDP</i>
12:20 – 12:40	Special Session on Energy and Resource Markets: <i>Discussion time</i>
Room 2 (MCLT102) Topic: Multi-Objective Optimization	
Chair: M. Stiglmayr	
11:20 – 11:40	[20] Michael Stiglmayr. <i>On the Multicriteria Linear Bottleneck Assignment Problem</i>
11:40 – 12:00	[27] Andrea Raith, Siamak Moradi, Matthias Ehrgott and Michael Stiglmayr. <i>Exploring Bi-objective Column Generation</i>
12:00 – 12:20	[3] Siamak Moradi, Matthias Ehrgott and Andrea Raith. <i>The Linear Bi-Objective Multi-Commodity Minimum Cost Flow Problem</i>
12:20 – 12:40	[25] Matthias Ehrgott, Maryam Hassanasab and Andrea Raith. <i>Data Envelopment Analysis without Linear Programming</i>
Lunch 12:40 – 1:30 (MC Foyer)	

Room 1 (MCLT101) Topic: Special Session on Energy and Resource Markets Chair: M. Young	
1:30 – 1:50	[66] Martin Young. <i>A Pragmatic Approach to Optimization of Water Distribution Networks</i>
1:50 – 2:10	[34] John Raffensperger and Darren Lumbroso. <i>Notes on a UK Water Market Demonstration</i>
2:10 – 2:30	[5] Vladimir Krichtal and Conrad Edwards. <i>HVDC Roles in the Economic Operation of the New Zealand Electricity Market</i>
2:30 – 2:50	[65] Bhujanga Chakrabarti and Ramesh Rayudu. <i>Review of Modelling for LMPs in Electricity Markets</i>
2:50 – 3:10	Special Session on Energy and Resource Markets: <i>Discussion time</i>
Room 2 (MCLT102) Topic: Computational OR Chair: D. Ryan	
1:30 – 1:50	[19] Christian Rolf. <i>Cloud Computing for Operations Research</i>
1:50 – 2:10	[54] Andrew Mason. <i>SolverStudio for Excel</i>
2:10 – 2:30	[53] David Ryan. <i>It is Time to Enjoy the Best of Both Worlds</i>
2:30 – 2:50	[50] Petros Hadjicostas. <i>Using L1-Regression to Estimate a Monotone Two-Piece Linear Relationship Between Two Angular Variables</i>
2:50 – 3:10	[11] Jason Markham and Nebojsa Djorovic. <i>A Simulation Model of Military Pilot Training</i>
Tea/Coffee Break 3:10 – 3:40 (MC Foyer)	
Room 1 (MCLT101) Topic: Scheduling and Optimization Chair: V. Mabin	
3:40 – 4:00	[59] Antony Phillips, Matthias Ehrgott and David Ryan. <i>Efficient Timetable Modifications in University Course Timetabling</i>
4:00 – 4:20	[60] Sarad Venugopalan and Oliver Sinnen. <i>Bi-Linear Reductions for the Multi-processor Scheduling Problem with Communication Delays using Integer Linear Programming</i>
4:20 – 4:40	[1] Maryam Mirzaei and Victoria Mabin. <i>Project Management: a Comparison of Three Popular Approaches</i>
4:40 – 5:00	[17] Oddo Zhang, Andrew Mason and Andy Philpott. <i>Simulation Optimisation for Ambulance Redeployment</i>
Room 2 (MCLT102) Topic: Case Studies Chair: T. Liddle	
3:40 – 4:00	[16] Craig MacLeod. <i>The Art and Science of Matchmaking (as it Relates to Badminton)</i>
4:00 – 4:20	[18] Desi Adhariani, Nick Sciulli and Bob Clift. <i>Everything is Gonna be Just Fine: The Corporate Governance Practices from Feminist Perspective Using the Optimization Method</i>
4:20 – 4:40	[12] Natashia Boland, Martin Savelsbergh and Mohsen Reisi. <i>Demand Driven Throughput Assessment for Hunter Valley Coal Chain</i>
4:40 – 5:00	[44] Anthony Downward, Yeong Fatt Thai and Golbon Zakeri. <i>Maximizing the Size of a Diamond, Cut from a Given Rough Stone</i>
Closing 5:00 – 5:05 (MCLT101)	

Author Index

Adhariani, Desi	3
Ahimbisibwe, Arthur	67
Al-Chanati, Salah	13
Anastasiadis, Simon	15
Arnold, Richard	25
Atkinson, Shanie	30
Babarenda Gamage, Thiranja	40
Boland, Natashia	44
Broadbent, Ali	46
Bull, Simon	47
Cabrera, Guillermo	57
Cameron, Donald	337
Cavana, Robert	67, 87, 258, 307, 337
Chakrabarti, Bhujanga	77
Chica, Manuel	57
Chukova, Stefanka	15, 25, 362
Clift, Bob	3
Cody, John	87
Crowder, Matthew	88
Cumming, Jacqueline	307
Djorovic, Nebojsa	220
Downward, Anthony	13, 97, 98, 176
Du, Jingze	108
Dye, Shane	118, 211, 297
Edwards, Conrad	157
Ehrgott, Matthias	57, 108, 119, 185, 194, 238, 278, 288
Elias, Arun	248
Farishta, Zabin	98
Freimer, Michael	320
Gary, Michael Shayne	30, 387
Goodall, Michelle	129
Hadjicostas, Petros	139
Hassanasab, Maryam	119
Hayakawa, Yu	25
Hinder, Oliver	148
Krichtal, Vladimir	157

Kristiansen, Simon	166
Leon, Michael	176
Lin, Kuan-Min	185
Liu, Zhengliang	194
Lumbroso, Darren	287
Mabin, Victoria	228
MacLeod, Craig	204
Mahakalanda, Indrajanaka	211
Markham, Jason	220
Mason, Andrew	47, 57, 88, 223, 398
Mirzaei, Maryam	228
Moradi, Siamak	238, 288
Muraleedharan, Parvathy	248
Nash, Martyn	40
Nielsen, Poul	40
Olsen, Tava	400
Ong, Su-Wuen	258
Pearson, David	87
Perederieieva, Olga	268
Phillips, Antony	278
Philpott, Andy	398
Raffensperger, John	211, 287
Raith, Andrea	47, 108, 119, 185, 194, 238, 288, 326
Range, Troels Martin	354
Rayudu, Ramesh	77
Read, Grant	2, 118, 211, 297
Read, Rosemary	118, 297
Rees, David	307
Reid, Paul	1
Reisi, Mohsen	44
Retna, Kala	67
Robinson, Grant	129
Rolf, Christian	317
Ross, Kevin	320
Ruddell, Keith	326
Ryan, David	278, 336
Sahli, Mondher	258
Sasso, Giuseppe	185
Savelsbergh, Martin	44
Sciulli, Nick	3
Scott, Rodney	337

Simpson, John	185
Sinnen, Oliver	378
Sophonsridsuk, Aioporn	338
Starkey, Stephen	118
Stidsen, Thomas	166
Stiglmayr, Michael	288, 348
Thai, Yeong Fatt	97
Tofte, Kasper	354
van Den Belt, Marjan	361
Varnosafaderani, Sima	362
Velez-Castiblanco, Jorge	371
Venugopalan, Sarad	378
Wahid, Faisal	98
Yang, Miles M.	387
Yetton, Philip W.	387
Young, Martin	397
Zakeri, Golbon	13, 97, 98, 176
Zhang, Oddo	398
Zhou, Quan	400
Ziedins, Ilze	410

Table of Contents

The Role of Operations Research in Determining Your Strategic Direction	1
<i>Paul Reid</i>	
Economics and Operations Research: A Past, Present and Future Duality	2
<i>Grant Read</i>	
Everything is Gonna be Just Fine: The Corporate Governance Practices from Feminist Perspective Using the Optimisation Method	3
<i>Desi Adhariani, Nick Sciulli and Bob Clift</i>	
Simulator for Electricity Related Investments	13
<i>Salah Al-Chanati, Golbon Zakeri and Anthony Downward</i>	
Modeling Technology Adoption Decisions where Farmers are Resistant to Change	15
<i>Simon Anastasiadis and Stefanka Chukova</i>	
Inference for Multicomponent Systems with Dependent Failures	25
<i>Richard Arnold, Stefanka Chukova and Yu Hayakawa</i>	
Dynamics of Mergers Acquisitions Integration	30
<i>Shanie Atkinson and Michael Shayne Gary</i>	
Optimal Design of Experiments to Determine Mechanical Properties of Soft Bodies	40
<i>Thiranjana Babarenda Gamage, Martyn Nash and Poul Nielsen</i>	
Demand Driven Throughput Assessment for Hunter Valley Coal Chain	44
<i>Natashia Boland, Martin Savelsbergh and Mohsen Reisi</i>	
Stakeholder Engagement in Capital Budgeting at Counties Manukau District Health Board	46
<i>Ali Broadbent</i>	
Scheduling Families of Jobs on Multiple Identical Machines to Minimize Total Tardiness ..	47
<i>Simon Bull, Andrew Mason and Andrea Raith</i>	
Mathematical Programming and Metaheuristic Approaches Applied to Biological-Based Fluence Map Optimization in Radiotherapy	57
<i>Guillermo Cabrera, Manuel Chica, Matthias Ehrgott and Andrew Mason</i>	
Structural Equation Modelling of Undergraduate Management Students Perceptions of Feedback in a New Zealand University	67
<i>Robert Cavana, Kala Retna and Arthur Ahimbisibwe</i>	
Review of Modelling for LMPs in Electricity Markets	77
<i>Bhujanga Chakrabarti and Ramesh Rayudu</i>	
Limits to Collective Action Development of an Evolutionary Game Model	87
<i>John Cody, Robert Cavana and David Pearson</i>	
Districting for the New Zealand Census: MIP-Heuristic Approaches	88
<i>Matthew Crowder and Andrew Mason</i>	
Maximizing the Size of a Diamond, Cut from a Given Rough Stone	97
<i>Anthony Downward, Yeong Fatt Thai and Golbon Zakeri</i>	

Use of Hydro Resources for Irrigation and Electricity Production	98
<i>Anthony Downward, Golbon Zakeri, Zabin Farishta and Faisal Wahid</i>	
Optimal Delivery of Arc Modulated Radiation Therapy in Cancer Treatment	108
<i>Jingze Du, Matthias Ehrgott and Andrea Raith</i>	
An Evaluation Tool for Reservoir Management	118
<i>Shane Dye, Grant Read, Rosemary Read and Stephen Starkey</i>	
Data Envelopment Analysis without Linear Programming	119
<i>Matthias Ehrgott, Maryam Hassanasab and Andrea Raith</i>	
Future Focused Network Modelling at New Zealand Post	129
<i>Michelle Goodall and Grant Robinson</i>	
Using L1-Regression to Estimate a Monotone Two-Piece Linear Relationship Between Two Angular Variables	139
<i>Petros Hadjicostas</i>	
Optimizing Clothing Catalogues for EziBuy	148
<i>Oliver Hinder</i>	
HVDC Roles in the Economic Operation of the New Zealand Electricity Market	157
<i>Vladimir Krichtal and Conrad Edwards</i>	
Elective Course Student Sectioning at Danish High Schools	166
<i>Simon Kristiansen and Thomas Stidsen</i>	
Simulating FTR Strategy in New Zealand Electricity Market	176
<i>Michael Leon, Golbon Zakeri and Anthony Downward</i>	
An Application of Data Envelopment Analysis to External Radiotherapy Treatment Planning	185
<i>Kuan-Min Lin, John Simpson, Giuseppe Sasso, Andrea Raith and Matthias Ehrgott</i>	
Linear Optimization over the Nondominated Set of a Multiobjective Linear Programming Problem	194
<i>Zhengliang Liu, Matthias Ehrgott and Andrea Raith</i>	
The Art and Science of Matchmaking (as it Relates to Badminton)	204
<i>Craig MacLeod</i>	
Intra-period Market Clearing for a Multi-Use Catchment via CDDP	211
<i>Indrajanaka Mahakalanda, Shane Dye, Grant Read and John Raffensperger</i>	
A Simulation Model of Military Pilot Training	220
<i>Jason Markham and Nebojsa Djorovic</i>	
SolverStudio for Excel	223
<i>Andrew Mason</i>	
Project Management: A Comparison of Three Popular Approaches	228
<i>Maryam Mirzaei and Victoria Mabin</i>	
The Linear Bi-Objective Multi-Commodity Minimum Cost Flow Problem	238
<i>Siamak Moradi, Matthias Ehrgott and Andrea Raith</i>	

Modelling Offshore Outsourcing of Software Testing Services: A Telecom New Zealand Case Study	248
<i>Parvathy Muraleedharan and Arun Elias</i>	
A Qualitative System Dynamics Analysis of Airline Safety in New Zealand	258
<i>Su-Wuen Ong, Robert Cavana and Mondher Sahli</i>	
Solving Bi-objective Traffic Assignment Based on Time Surplus Maximisation	268
<i>Olga Perederieieva</i>	
Efficient Timetable Modifications in University Course Timetabling	278
<i>Antony Phillips, Matthias Ehrgott and David Ryan</i>	
Notes on a UK Water Market Demonstration	287
<i>John Raffensperger and Darren Lumbroso</i>	
Exploring Bi-objective Column Generation	288
<i>Andrea Raith, Siamak Moradi, Matthias Ehrgott and Michael Stiglmayr</i>	
Generalized CDDP for Reservoir Management	297
<i>Rosemary Read, Shane Dye and Grant Read</i>	
Using Cognitive Mapping and Qualitative System Dynamics to Develop a Theory of Implementation in Primary Health Care	307
<i>David Rees, Robert Cavana and Jacqueline Cumming</i>	
Cloud Computing for Operations Research	317
<i>Christian Rolf</i>	
Contribution Margin Optimisation at Fonterra	320
<i>Kevin Ross and Michael Freimer</i>	
An Aggregational Approach to the Traffic Assignment Problem	326
<i>Keith Ruddell and Andrea Raith</i>	
It is Time to Enjoy the Best of Both Worlds	336
<i>David Ryan</i>	
Evaluating the Impact of Systems Thinking Workshops on Strategy Implementation in a Government Department	337
<i>Rodney Scott, Robert Cavana and Donald Cameron</i>	
The Application of Linear Programming to Select the Lowest Cost and Optimized Quality of Textile Wet Process	338
<i>Aioporn Sophonsridsuk</i>	
On the Multicriteria Linear Bottleneck Assignment Problem	348
<i>Michael Stiglmayr</i>	
A Time-Indexed Model for the Elective Surgery Scheduling Problem	354
<i>Kasper Tofte and Troels Martin Range</i>	
Mediated Modelling to Support Spatial Planning: Population Change, Inequality and City Attractiveness in Wellington	361
<i>Marjan van Den Belt</i>	

Modeling Repairs of Systems with a Bathtub-Shaped Failure Rate Function	362
<i>Sima Varnosafaderani and Stefanka Chukova</i>	
Intervention as Language Games	371
<i>Jorge Velez-Castiblanco</i>	
Bi-Linear Reductions for the Multiprocessor Scheduling Problem With Communication Delays Using Integer Linear Programming.....	378
<i>Sarad Venugopalan and Oliver Sinnen</i>	
Organizational Goals, Feedback Effects, and Performance.....	387
<i>Miles M. Yang, Michael Shayne Gary and Philip W. Yetton</i>	
A Pragmatic Approach to Optimization of Water Distribution Networks.....	397
<i>Martin Young</i>	
Simulation Optimisation for Ambulance Redeployment	398
<i>Oddo Zhang, Andrew Mason and Andy Philpott</i>	
Developing a Rotation Scheme to Reduce Expiration for the Medical Reserve Supply	400
<i>Quan Zhou and Tava Olsen</i>	
Accumulating Priority Queues: A New Priority Scheme for Hospital Queues?.....	410
<i>Ilze Ziedins</i>	

PLENARY TALK: DAY 1

The Role of Operations Research in Determining Your Strategic Direction

Paul Reid

Group General Manager: Innovation and Technology
New Zealand Post Group
Wellington, New Zealand

Abstract

Over the last 20 years, Paul Reid has worked in a number of companies that rely heavily on Operations Research techniques to determine their strategic direction, as well as for day to day operational decisions. Paul will share his insights into the use of Operations Research in forestry, aviation, weather and logistics. He will also outline some of the challenges ahead for New Zealand Post.

PLENARY TALK: DAY 2

Economics and Operations Research: A Past, Present and Future Duality

E Grant Read
Department of Management
University of Canterbury
New Zealand
grant.read@canterbury.ac.nz

Abstract

Economics and Operations Research sprang from common roots, and this has been particularly evident in the energy sector, where large scale optimisation models have been applied to many “economic” planning and policy problems. At Canterbury, Operations Research even grew up in an Economics department. Over the years, though, the disciplines seemed to grow apart. Operations Researchers sometimes claimed that they were the ones who really knew how to solve the big problems economists could only wave their hands at. But, as they developed tools to solve ever larger planning problems, economists became increasingly wary of the whole concept of large scale planning. There is still a strong link between the two disciplines, though, with OR traditionally focussed on the “primal” problem of finding efficient solutions, and Economics focussed more on the “dual” problem, of efficient price interactions. And there is also a very large overlap between the disciplines, in practice, with “OR” models increasingly used to clear markets, and to model and guide “economic” behaviour of market participants. Indeed, some of us are not sure whether we are really OR people, or economists. Here we try to draw some lessons from a survey of fruitful interaction between the disciplines in the New Zealand electricity sector, over the years, and speculate a little about future trends.

Everything is Gonna be Just Fine: The Corporate Governance Practices from Feminist Perspective Using the Optimisation Method

Desi Adhariani
Victoria University, Melbourne, Australia
and University of Indonesia
desi.adhariani@vu.edu.au

Nick Sciulli
Victoria University, Melbourne, Australia

Bob Clift
Victoria University, Melbourne, Australia

Abstract

The purpose of this research is to **project** the financial condition of a company using the feminist ethics of care principles. One company is chosen to explore the application of the feminist perspective in corporate governance practices, even though the company does not claim to apply the feminist perspective. BHP Billiton is selected as the sample as it is renowned as one big company in Australia with wide social and environmental responsibility. A projection of financial condition of the company is performed as a tool to describe how a company can use its resources while achieving the objective to satisfy the stakeholders. The output is the balance sheet and income statement projections, from which several financial ratios are then computed. The projection is developed using the **simple optimisation method** (linear programming). The result shows that the financial condition of the company in the future is stable and sustainable as expected by the theory of feminist ethics of care.

1. Background

A large proportion of mainstream literature in accounting and finance conceptualises the definition of a firm as a “nexus of contracts” with conflicts of interest among the contracting parties (Coase 1937). One of the conflicts of interest occurs between shareholders and managers, or what is articulated by Jensen and Meckling (1976) as the agency problems, in terms of competing and conflicting claims, problem solving using rules and laws, and the rights and obligations measurement in legitimacy and the power dimension (Machold, Ahmed & Farquhar 2008). Implicitly or explicitly, the moral reasoning within this discourse is a masculinist view according to Gilligan (1982). Under this perspective, which is called the ethics of rights, the agency problem due to the conflicting interests among parties in a company, can be reduced but not really eradicated at the roots. The corporate collapses that occur repeatedly is a strong

evidence that the mode of operating using rules, laws and a power framework are a short-term remedy without sufficient guarantee of long term success in term of the sustainability of a business.

Many reasons can be attributed to such deficiency but it can be traced back to the ineffective corporate governance represented by the lack of ethical commitment in running a business. This research presents an argument that the lack of ethical conduct could also be because the ethics principles applied in companies are only based on the ethics of rights perspectives yet ignoring the other types of ethics that can become the complements.

The adequacy of applying the ethics of rights perspective only gives rise to some writings which try to view governance from the feminist ethics perspective, which is called the ethics of care. While the rights perspective emphasises the rules and respect for the rights of others, the care point of view stresses the responsibility, relationship, concern, care, continued attachment, sacrifice and the avoidance of hurting another (Reiter 1997).

Traditionally, a different moral emphasis can be traced back to gender stereotypes. A study conducted by Bampton and Maclagan (2009) confirms that women are more inclined to follow the ethics of care than men and make them react differently to business ethics issues compared with men. Nevertheless, ‘the difference between masculinist and feminist perspectives is not exclusively and sharply defined along sexual lines’ (Machold, Ahmed & Farquhar 2008, p. 668). Moreover, Velasquez (1998, p.126) emphasises that ‘caring is not the task of women, but a moral imperative for both men and women’. With this idea, a feminine firm which values connectedness and relationships in its vision and mission can generate the bonds of trust from its stakeholders and, hence, overcome the inefficiencies of its masculine counterpart.

2. Aim of the Project

The main aim of this research is to design a model to project the financial condition of a company which integrate good corporate governance principles using a feminist perspective. The feminist theory applied is restricted to the ethics of care that is most closely associated with the research work by Carol Gilligan (1982) which has driven many research in this area afterwards (Bampton & Maclagan 2009; French & Weis 2000; Reiter 1997).

3. Literature Review

The ethics of care is a feminist critique to rebalance the other beliefs called the ethics of rights/ethics of justice or separative model (Reiter 1997). The work of Gilligan (1982) is performed as a response to the observations made by developmental psychologist Kohlberg (1981) who found that women scored lower on the test of moral development. Gilligan argued that the result might be biased since Kohlberg’s theory was developed using exclusively male samples; therefore, she introduced a different perspective of female moral discourse which is labelled the ethics of care, as the opposite to the ethics of justice attributed to males. The differences between the two are displayed in Table 1.

Table 1. Ethics of Care vs Ethics of Rights

No.	Ethics of care	Ethics of rights
1	Achieved through perception of one's self as connected to others	Achieved through process of separation and individuation of self from others
2	Moral dilemmas contextual	Moral dilemmas universal
3	Dilemmas solved through inductive thinking	Dilemmas solved through application of abstract or formal thinking
4	Development through stages is sequential and hierarchical	Development through stages is invariantly sequential and hierarchical
5	Principle of moral responsibility is reflected in the voices of women	Principle of moral responsibility is universal
6	Distinguished by an emphasis on attachments, issues of self-sacrifice and selfishness, and consideration of relationships as primary	Distinguished by an emphasis on separateness, issues of rules and legalities, and consideration of the individual as primary

Source: (Brabeck 1993)

Based on those differences, several arguments are attributed to the ethics of care which will be used in this project:

1. The ethics of care emphasises on the responsibility, relationships and connection to others; while the ethics of justice stresses on rules and on the respect for the rights of others.
In corporate governance research, prioritising relationships as suggested by the ethics of care is translated into the stakeholders' objective of a company.
2. In the ethics of rights, adherence to universal principles is sought to solve moral dilemmas; while the ethics of care suggests more contextual responses to moral issues (case by case, or concentrate on particular situation). The problem resolutions in the ethics of care include concern, care, continued attachment, responsibility, sacrifice, and the avoidance of hurting another.
3. This research does not take into account the fifth difference which states that "principle of moral responsibility is reflected in the voices of women" because later research after Gilligan's breaking away from the ethics of rights point out that the "different voice" of moral reasoning is not necessarily attributed to women (French & Weis 2000; Simola 2003). An ethics of care extends to men as well and is influenced by social, political and economic contexts as can be ascertained in Hofstede's findings (1998; 2005). The inclusiveness of feminist theory that does not just apply to women can be viewed as an attempt to provide alternative ways of theorising and to encourage researchers to explicitly put forth values implied in "previously-determined-to-be value-free" research.
4. Since building a good relationship is primary, acknowledgement and effective management of relationships is the core of the ethics of care.
5. The feminist ethics envisages the importance of relationship maintenance that goes beyond rules and regulations. Since maintaining relationships to other parties named stakeholders are important then it is further translated in to several activities which are now known as CSR activities.
6. Go beyond rules and regulations means that the CSR activities conducted by a company are not only be intended to fulfil the regulations from government or

international pressures; rather, they are intended to maintain good relationships with stakeholders. This means, the nature of the activities should not only be mandatory or compulsory but also, more importantly, **voluntary, or discretionary responsibilities** (Carroll 1979).

4. Research Question

Based on the literature review presented in the previous section, an optimisation model is developed using the feminist ethics of care perspective. It is aimed at answering this research question:

“How is the financial condition of the company if the feminist ethics of care is applied in its corporate governance practices?”

5. Approach and Methodology

5.1 Introduction to an Optimisation Approach

Mathematical programming, also known as optimisation, is an approach used to find the best possible solution (the optimal or most efficient way) of using limited resources to achieve certain definitive objectives (Ragsdale 2001). This project utilizes an optimisation approach to develop the financial projection model of a company, because it can clearly incorporate the objectives statement of a company and the constraints that the company faces; and hence, it fits within the aim of this project. Specifically, this research uses the linear programming method, which is one of the deterministic models, with the assumption that all controllable and uncontrollable variables are known.

5.2 Research Design

5.2.1 Sample Selection

This project is a case study of one particular company in Australia. The company chosen must be the one that has the data which fit the model.

A natural resources company which has a primary listing on the Australian Securities Exchange (ASX) in Australia, is chosen as the sample. This company is selected because the nature of its business creates externalities to social community and environment, and hence, the approach that this company take to manage those problems is interesting to be investigated. This company (Company X) provides explanation on its sustainability activities in a separate report which is published annually.

Data from the five years 2006-2010 are used to project the financial condition in 2011-2015. Several data inputs of this company can be found in Table 2.

5.2.2 Data Sources and Collection

This project uses secondary data from annual reports including financial statements.

5.2.3 Research Approach

This study involves two significant parts in order to answer the research question.

1. Part 1: The optimisation model, and the simulation of the model under feminist ethical perspective is developed and conducted.
2. Part 2: Several financial ratios are calculated based on the resulting proforma financial statements to project the financial condition of the company.

Table 2. Financial Data of Company X

US\$ million	2005	2006	2007	2008	2009	2010	Applied Number	2011	2012	2013	2014	2015
Revenues (Net sales revenue plus other income)	39,886	47,962	59,991	50,762	53,212			57,838	62,865	68,330	74,269	80,725
Revenue growth	0.202	0.251	-0.154	0.048			0.087					
Operating costs (payments to suppliers, contractors etc)	17,988	19,936	26,358	23,877	22,306			29,400	34,030	39,390	45,593	52,773
Operating Costs/Sales Revenue	0.5595	0.5047	0.4432	0.4755	0.4225		0.4811					
Employee wages and benefits (Expenditure on wages and benefits of the employee workforce and not future commitments)		2,982	3,311	4,360	4,345	4,830		5,230	6,054	7,007	8,111	9,388
Employee wages and benefits/Sales Revenue		0.0927	0.0838	0.0733	0.0865	0.0915	0.0856					
Payments to providers of capital: Dividends to all shareholders	1,423	1,936	2,271	3,135	4,563	4,618		4,674	4,730	4,788	4,846	4,904
Dividends growth	0.3605	0.1730	0.3804	0.4555	0.0121							
Payments to providers of capital: interest payments made to providers of loans		626	601	722	589	496						
Payments to government (Gross taxes and royalties)	2,784	5,341	6,061	8,121	7,940	6,892		9,103	10,537	12,196	14,117	16,340
Payments to government/Sales Revenue		0.1661	0.1335	0.1365	0.1581	0.1305	0.1490					
Community investments (voluntary contributions of funds in the broader community)	57.4	81.3	103	141	198	200						
Environmental expenditure (including R&D, Site rehabilitation, environmental monitoring, and other environment expenditure such as environmental impact assessment and training.)	267	309	288									
Environmental expenditure/EBIT	0.0315	0.0211	0.0157									
Accounts payable (Trade & other payables)	4,053	4,724	6,774	5,619	6,467			7,308	6,459	9,792	11,334	13,118
Accounts payable (Trade & other payables)/Assets	0.0835	0.0812	0.0893	0.0713	0.0728		0.0796					
Suppliers/contractors related expenses	16,384	19,230	25,577	24,805	22,326							
Inventory	2,732	3,296	4,971	4,821	5,334							
Interest on loans	516	557	722	574	484			899	1,069	1,270	1,510	1,794
Interest bearing liabilities	7,648	9,291	9,234	15,325	15,573			16,151	19,172	22,786	27,081	32,186
Growth of Interest Bearing Liabilities	0.2148	-0.0061	0.6596	-0.1143								
Prearranged take down of BL							-0.0061					
Interest bearing liabilities after prearranged take down								15,490	13,408	13,326	13,245	13,164
Interest on prearranged take down liabilities								752	747	743	738	734
Interest on loans/Interest bearing liabilities	0.0675	0.0600	0.0782	0.0375	0.0357	0.0557						
Number of ordinary shares (not in US\$)	3,495,949,933	3,357,503,573	3,358,359,496	3,358,359,496	3,358,359,496			3,359,367,004	3,360,374,814	3,361,382,916	3,362,391,341	3,363,400,659
Growth of number of ordinary shares	-0.0396	0.0003	0.0000	0.0000			0.0003					
Sales Revenue	32,153	39,498	59,473	50,211	52,798			61,113	70,738	81,878	94,772	109,688
Sales Growth	0.2284	0.5057	-0.1557	0.0515			0.1575					
Assets	48,516	58,168	75,889	78,770	88,852			91,783	106,237	122,968	142,334	164,750
Assets/Sales Ratio	1.5089	1.4727	1.2760	1.5688	1.6829		1.5019					
EBIT	8,481	14,671	18,401	24,145	12,160	20,031		23,831	27,584	31,928	36,956	42,776
EBIT/Sales ratio	0.4563	0.4659	0.4060	0.2422	0.3794	0.3899						
Depreciation expense	2,236	2,395	3,585	3,852	4,732			4,863	5,645	6,553	7,608	8,831
Property, Plant and Equipment (Net Book value)	30,983	36,705	47,392	49,032	55,576			64,517	74,896	86,945	100,933	117,171
PP&E growth	0.1846	0.2095	0.0359	0.1353			0.1609					
Depreciation expense/PP&E	0.0722	0.0652	0.0757	0.0786	0.0851	0.0754						
Sustainable activities in current and non-current Provision account	5,427	6,436	7,273	8,103	8,592			9,645	10,828	12,156	13,846	15,319
Provision growth	0.1659	0.1300	0.1141	0.0603			0.1226					

US\$ million	2005	2006	2007	2008	2009	2010	Applied Number	2011	2012	2013	2014	2015
Research and Development expenses	76	189	244	156	65			186	215	249	288	334
Total expenses	22,403	26,352	35,976	38,640	33,295			41,178	47,663	55,169	63,858	73,915
Total expenses/Sales Revenue	0.6988	0.6872	0.6049	0.7696	0.6306		0.6738					
R&D expenses/Total expenses	0.0034	0.0064	0.0068	0.0040	0.0020		0.0045					
Total fines for breaching environmental regulation	0.4795	0.0374	0.1176	0	0.0351							
Preferred dividend	17	1	1	1	1	0						
Exploration and evaluation expense (excluding impairment) in Operating Section of Cash Flow Statement	561	528	859	1,009	1,030			1,030	1,201	1,390	1,609	1,863
Exploration and evaluation expense (excluding impairment) /Total Exp	0.0250	0.0200	0.0239	0.0261	0.0309		0.0252					
Trade and other receivables	3,831	4,689	9,901	5,133	6,543			7,853	9,089	10,521	12,178	14,095
Trade and other receivables/Total Assets	0.0790	0.0806	0.1291	0.0654	0.0736		0.0856					
Inventories	2,732	3,296	4,971	4,821	5,334			5,502	6,368	7,371	8,532	9,876
Inventories / Total Assets	0.0563	0.0567	0.0655	0.0612	0.0600	0.0599						
Interest coverage=EBIT/Interest	28	33	33	21	41	21						
Dividend (c)	0.18	0.24	0.35	0.41	0.42							
Dividend growth	0.3056	0.4894	0.1714	0.0244			0.0244					
EPS (cents)	173.2	229.5	275.3	105.6	228.6							
Dividend payout ratio: Dividend per share/EPS	0.1039	0.1024	0.1271	0.3883	0.1837			645	747	865	1,001	1,158
Royalty-related taxes	425	341	723	495	451							
Royalty-related taxes/Total expenses	0.0190	0.0129	0.0201	0.0128	0.0135	0.0157						

6. Optimisation Model Development

This study develops a quantitative financial optimisation model based on Carleton's linear programming model (1970), that will be referred to as the "traditional" model because it was developed using the "masculinist" view of finance and accounting theory.

The optimisation model can be found in Table 3. Since the model is intended to be used for long range financial planning, then the planning period used in this project is five years.

Table 3. Optimisation Model

Optimisation Model	Explanation
<p>Objective Function Maximize $\sum_{t=1}^5$ <i>economic value retained</i></p> <p>= Max $\sum_{t=1}^5$ <i>economic value generated and distributed</i></p> <p>=Max $\sum_{t=1}^5$ <i>Revenues – Operating costs – Employees wages and benefits – Payments to providers of capital – Payments to government – Community investments – Environmental contributions</i></p> <p>=Max $\sum_{t=1}^5$ <i>Revenue – Payments to suppliers, contractors, etc – Wages and benefits to employees – Shareholder dividends – Interest payments – Gross taxes and royalties – Voluntary contributions of funds in the broader community – Total environmental expenditure</i></p>	<p>The economic value retained concept is advised by the GRI (Global Reporting Initiative)-G3 to be disclosed by a company in its sustainability report/section in Annual Report.</p>
<p>Decision Variables</p> <p><i>Dividend_t</i> Δ <i>Long term debt_t</i> <i>Payment to Government_t</i> <i>Community Contributions_t</i> <i>Environmental expenditure_t</i> <i>Available for Commons_t</i></p>	<p>The Available for Commons variable is the decision variable in the constraints, which is defined as the available earnings to be distributed to the common stockholders.</p>
<p>Constraints</p>	
<p>$Revenue_t = \alpha_t Revenue_{t-1}$</p>	<p>It is assumed that a company will prefer to maintain a stable growth of revenues represented by α_t</p>
<p>$Operating\ costs_t = \beta_t Operating\ costs_{t-1}$</p>	<p>β_t is defined as the ratio of Operating costs/Revenue</p>
<p>$Employee\ wages\ and\ benefits_t = \gamma_t Employee\ wages\ and\ benefits_{t-1}$</p>	<p>γ_t is defined as the ratio of the Employee Wages and benefits /Revenue</p>
<p>$Payments\ to\ government_t = Gross\ taxes_t + Other\ Payments_t$ $Payments\ to\ government_t = (tax\ rate \times Profit\ before\ tax_t) + Other\ Payments_t$ $Payments\ to\ government_t = 30\%(EBIT_t - i_tLTD_t - i'_t\Delta LTD_t) + Other\ Payments_t$</p>	<p>i is the interest rate applicable to the existing long term debt</p>

Optimisation Model	Explanation
	i' is the interest rate applicable to the positive changes in long term debt (or interest bearing liabilities)
<p><i>Available for Common_t = After tax Profit_t - Preferred Dividend_t - Special Adjustment_t</i></p> <p><i>After tax Profit_t = (1-tax rate) (EBIT_t + Δe PPE_t - Δa PPE_t - Σ_{t=1}⁵ i x Existing Liabilities - Σ_{t=1}⁵ i' x ΔInterest bearing liabilities</i></p>	<p>Δe is the company's depreciation rate of assets</p> <p>Δa is the tax-reported accelerated depreciation</p>
<p>Provision amounts related with CSR activities in Balance Sheet should be growing in the average of provision growth in the last 5 years</p> <p><i>Provision_t = (1+ε) Provision_{t-1}</i></p>	
<p>Research and development expense for CSR activities should be stable in the average proportion of (R&D expenses / Total expenses) for the last 5 years</p> <p><i>R&D expenses_t = ε Total expenses_t</i></p>	<p>Even though the amounts are not material, but this constraint is intended to ensure that a company is continuing to search new ways to better interact with its social community and environment.</p>
<p>The ratio of Total fines to the amount of Community contribution and environmental expenditure should be less than the minimum ratio in the last five years</p> <p><i>$\frac{\text{Total fines}_t}{\text{Comm.contr} + \text{Env.exp}} \leq X$</i></p>	
<p>Community contribution and environmental expenditure should be at least certain percentage of EBIT</p> <p><i>Community contribution_t ≥ X% EBIT_t</i></p> <p><i>Environmental expenditure_t ≥ X% EBIT_t</i></p>	
<p>The interest coverage should be at least equal the minimum interest coverage multiplier in the last five year period.</p> <p><i>$\frac{\text{EBIT}_t}{(\sum_{t=1}^5 i \times \text{Existing Liabilities} + \sum_{t=1}^5 i' \times \Delta \text{Interest bearing Liabilities})} \geq X$</i></p>	
<p>The growth of dividend payout should be at least stable at certain minimum dividend growth rate</p> <p><i>D_t - θ_t D_{t-1} ≥ 0</i></p>	
<p>Payout restriction of dividend payment (lower and upper bound)</p> <p><i>D_t ≥ δ₁ AFC_t</i></p> <p><i>D_t ≤ δ₂ AFC_t</i></p>	

The result of the application of the model using Solver software in Excel is given in Table 4.

Table 4. Solver Result

	2011	2012	2013	2014	2015
Dividend	7,624	8,859	10,292	11,952	13,874
Interest Bearing Liabilities	380	734	856	996	1,188
Payment to Government	7,564	8,787	10,209	11,716	13,461
Community Contribution	238	276	319	370	428
Environmental Expenditure	477	552	639	739	856
Available For Commons	19,547	22,715	26,390	30,647	35,575
Target Cell (Max): Coeff Obj = 68,080					

Using the optimal output generated, the Pro Forma Balance Sheet and Other financial data and related ratios for Company X are constructed as follows:

Table 5. Proforma Balance Sheet

Item	As of Year-End (US\$ Million)				
	2011	2012	2013	2014	2015
<i>Assets</i>	91,783	106,237	122,968	142,334	164,750
Trade and other receivables	7,853	9,089	10,521	12,178	14,095
Inventories	5,502	6,368	7,371	8,532	9,876
Property, Plant and Equipment (Net Book Value)	64,517	74,896	86,945	100,933	117,171
<i>Liabilities</i>					
Accounts payable (Trade & other payables)	7,308	8,459	9,792	11,334	13,118
Interest bearing liabilities	380	734	856	996	1,188
<i>Equity</i>	84,094	97,044	112,321	130,005	150,444

Table 6. Other Financial Data and Related Ratios

Item	Years											
	2005	2006	2007	2008	2009	2010	Applied Number	2011	2012	2013	2014	2015
Revenues (Net sales revenue plus other income)		39,686	47,962	59,991	50,762	53,212		57,838	62,865	68,330	74,269	80,725
Revenue growth		0.202	0.251	-0.154	0.048		0.087					
Operating costs (payments to suppliers, contractors etc)		17,988	19,936	26,358	23,877	22,306		29,400	34,030	39,390	45,593	52,773
Operating Costs/Sales Revenue		0.5595	0.5047	0.4432	0.4755	0.4225	0.4811					
Employee wages and benefits (Expenditure on wages and benefits of the employee workforce and not future commitments)		2,982	3,311	4,360	4,345	4,830		5,230	6,054	7,007	8,111	9,388
Employee wages and benefits/Sales Revenue		0.0927	0.0838	0.0733	0.0865	0.0915	0.0856					
Payments to providers of capital: Dividends to all shareholders	1,423	1,936	2,271	3,135	4,563	4,618		7,624	8,859	10,292	11,952	13,874
Dividends growth	0.3605	0.1730	0.3804	0.4555	0.0121		0.0121	0.6509	0.1620	0.1618	0.1613	0.1608
Payments to government (Gross taxes and royalties)	2,784	5,341	6,061	8,121	7,940	6,892		7,564	8,787	10,209	11,716	13,461
Payments to government/Sales Revenue		0.1661	0.1535	0.1365	0.1581	0.1305	0.1490					
Community investments (Voluntary contributions of funds in the broader community)	57.4	81.3	103	141	198	200		238	276	319	370	428
Environmental expenditure (including R&D, Site rehabilitation, environmental monitoring, and other environment expenditure such as environmental impact assessment and training.)	267	309	288	The Company's focus on integrating environmental responsibility into activities means that it is not possible to accurately extract expenditure spent on the environment and, for that reason, it is no longer reported.				477	552	639	739	856
Environmental expenditure/EBIT	0.0315	0.0211	0.0157									
Number of ordinary shares (not in US\$)		3,495,949,933	3,357,503,573	3,358,359,496	3,358,359,496	3,358,359,496		3,359,367,004	3,360,374,814	3,361,382,926	3,362,391,341	3,363,400,059
Growth of number of ordinary shares		-0.0396	0.0003	0.0000	0.0000		0.0003					
Sales Revenue		32,153	39,498	59,473	50,211	52,798		61,113	70,738	81,878	94,772	109,698
Sales Growth		0.2284	0.5057	-0.1557	0.0515		0.1575					
EBIT	8,481	14,671	18,401	24,145	12,160	20,031		23,831	27,584	31,928	36,956	42,776
EBIT/Sales ratio		0.4563	0.4659	0.4060	0.2422	0.3794	0.3899					
Sustainable activities in current and non-current Provision account		5,427	6,436	7,273	8,103	8,592		9,645	10,828	12,156	13,646	15,319
Provision growth		0.1859	0.1300	0.1141	0.0603		0.1226					
Research and Development expenses		76	169	244	156	65		186	215	249	288	334
Total expenses		22,403	26,352	35,976	38,640	33,295		41,178	47,663	55,169	63,858	73,915
Total expenses/Sales Revenue		0.6968	0.6672	0.6049	0.7696	0.6306	0.6738					
R&D expenses/Total expenses		0.0034	0.0064	0.0068	0.0040	0.0020	0.0045					
Total fines for breaching environmental regulation		0.4798	0.0374	0.1176	0	0.0351						
Exploration and evaluation expense (excluding impairment) in Operating Section of Cash Flow Statement		561	528	859	1,009	1,030		1,038	1,201	1,390	1,609	1,863
Exploration and evaluation expense (excluding impairment)/Total Exp		0.0250	0.0200	0.0239	0.0261	0.0309	0.0252					
Dividend payout								7,624	8,859	10,292	11,952	13,874
Available for Commons								19,547	22,715	26,390	30,647	35,575

7. Analysis and Conclusion

The optimal solutions resulting from the optimisation approach as presented above show that Company X is in stable and sustainable condition for the five years ahead. Maximising the stakeholders' wealth as implicitly suggested by the ethics of care will force the company toward the effort to balancing the interests of the stakeholders. One thing to be noticed, however, the interest bearing liabilities are advised to be reduced from previous years; meaning that the company should prepare other sources or alternatives for financing.

References

- Bampton, R. and P. Maclagan. 2009. "Does a care orientation explain gender differences in ethical decision making? A critical analysis and fresh findings." *Business Ethics: A European Review* 18: 179-191.
- Brabeck, M. 1993. "Moral judgement: theory and research on differences between males and females," in M. Larrabee (ed.), *An Ethic of Care*, Routledge, New York, pp. 33-48.
- Carleton, W.T. 1970. "An analytical model for long-range financial planning." *The Journal of Finance* 25: 291-315.
- Carroll, A.B. 1979. "A three-dimensional conceptual model of corporate performance." *The Academy of Management Review* 4: 497-505.
- Clarkson, M.B.E. 1995. "A stakeholder framework for analyzing and evaluating corporate social performance." *The Academy of Management Review* 20: 92-117.
- Coase, R.H. 1937. "The nature of the firm." *Economica* 4: 386-405.
- French, W. and A. Weis. 2000. "An ethics of care or an ethics of justice." *Journal of Business Ethics* 27: 125-136.
- Gilligan, C. 1982. *In a Different Voice, Psychological Theory and Women's Development*, Harvard University Press, Cambridge.
- Hofstede, G. 1998. *Masculinity and Femininity: the Taboo Dimension of National Cultures*, Cross-Cultural Psychology, Sage, California.
- Hofstede, G. and G.J. Hofstede. 2005. *Cultures and Organizations: Software of the Mind*, 2nd edn, McGraw-Hill, New York.
- Jensen, M.C. and W.H. Meckling. 1976. "Theory of the firm: managerial behavior, agency costs and ownership structure." *Journal of Financial Economics* 3: 305-360.
- La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R. Vishny. 2000. "Investor protection and corporate governance." *Journal of Financial Economics* 58: 3-27.
- Machold, S., P. Ahmed and S. Farquhar. 2008. "Corporate governance and ethics: a feminist perspective." *Journal of Business Ethics* 81: 665-678.
- Ragsdale, C.T. 2001. *Spreadsheet Modeling and Decision Analysis*, 3rd edn, Thomson South Western, Ohio.
- Reiter, S. 1997. "The ethics of care and new paradigms for accounting practice." *Accounting, Auditing & Accountability Journal* 10: 299-324.
- Scott, W. 2006. *Financial Accounting Theory*, 4th edn, Pearson Prentice Hall, Toronto.
- Simola, S. 2003. "Ethics of justice and care in corporate crisis management." *Journal of Business Ethics* 46: 351-361.

Simulator for Electricity Related Investments

Salah Al-Chanati, Golbon Zakeri and Anthony Downward
Department of Engineering Science
The University of Auckland
New Zealand
salah.ajr.alchanati@gmail.com g.zakeri@auckland.ac.nz
a.downward@auckland.ac.nz

Extended Abstract

The New Zealand Electricity Market (NZEM) is a wholesale market where electricity is bought, sold and traded between major participants: generators (electricity production stations), distributors (local lines companies), retailers (retail companies which compete to buy wholesale electricity and compete to retail it to consumers), and industrial users.

Electricity is traded through bids/offers and short-/long-term contracts that have financial obligations on both parties involved. In the wholesale market, bids are submitted by purchasers that express the willingness to purchase electricity for a given price, while offers are submitted by generators and describe the amount of electricity they are willing to produce at a given price. Both bids and offers are submitted to a centralised system as price-quantity pairings, also referred to as tranches of an offer stack. The sequence of all these bids and offers together are called the aggregated demand stack and offer stack respectively. An optimisation problem will then be solved by the centralised system to obtain an intersection point between the demand stack and offer stack, known as the “point of dispatch”. At this point, the market clears and electricity will be generated and dispatched to the purchasers at the intersection price per unit of electricity, while the gross payoff that the generator will expect is simply the amount of electricity generated (MWh) multiplied with the clearing price.

Yet, due to seasonal changes (i.e. weather conditions) and its impact on lake levels (electricity supply), as well as the unpredictable demand for electricity by domestic users and business (electricity demand), NZEM participants are exposed to severe financial risks, such as loss of potential revenues (for generators) and additional electricity costs (for distributors, retailers and industrial users).

To address these risks, a risk management programme using hedging techniques and financial contracts take place. This can be used to guarantee a minimum price received for supplying electricity for the generators and a maximum price paid by distributors, retailers and industrial users.

Other energy markets, e.g. PJM (United States) and NEM (Australia), have a more established and complex financial systems composed of different structures and regulations. In comparison, NZEM related financial instruments are currently trading on the Australian Stock Exchange (ASX) and are still relatively new (trading since July 2009). These contracts are the ASX New Zealand Electricity (NZE) Futures Contracts for the Otahuhu (North Island) and Benmore (South Island) nodes. They allow NZEM participants to secure an average price, for a 3-month period into the future, and hedge against any potential price volatilities.

However, whether taking a buy or sell position, both parties need to determine a reasonable price range to pay for these contracts. In addition, they also need to consider the possibility of generating a profit in the process.

The aim of this research project is to develop a simulator that can assist any individual, groups or organisations wanting to incorporate these financial instruments into their portfolios. This will allow users of the *Simulator* to price these contracts without over paying (buyer) or under selling (seller) to hedge against price volatilities, and better manage the risk exposures faced in the NZEM. In addition, users will be able to maximise (minimise) the potential returns (losses) on these financial contracts while hedging against the adverse events mentioned previously.

The *Simulator* undergoes a three stage process: estimating the future electricity prices at each node; estimating the future quarterly contract prices; determining the profit (loss) generated from engaging in a particular ASX NZE Futures Contract.

Electricity price estimation is performed using Bayes' Rule, a Hidden Markov model inference technique. Based on today's electricity price and historical price patterns (between 2000 and 2011), the *Simulator* generates probabilities of likelihood of the type of year we are expected to be in. Then, the average of these predicted 90-92 days of electricity prices associated with the contract period (e.g. January to March 2012 = March 2012 contract) is taken as the predicted contract price. Finally, the predicted contract price is compared with the actual price trading on the ASX. If the ASX contract price is below the simulator's prediction, i.e. we expect the average electricity prices to go up into the future, a buy position takes place, and vice versa for a sell position when the ASX price is above the simulator's prediction.

The *Simulator* has eight prediction, buy position and sell position parameters. The main four parameters, which the user can change, are: *Investment Level* and *Number of Contracts Purchased* (allow the user to decide how much to invest), and *Price Gap* (ASX Price - Simulator Price) and *Percentage of Contracts Sold* (allow the user to decide at what price level and how many contracts to sell).

Ultimately, users will be able to gain an initial understanding of electricity price volatilities, of how the ASX NZE Futures Contracts operate, and enables them to develop an appropriate strategy to match their risk preferences and potentially earn a profit from their desired actions.

Note: the Simulator uses data between 1 January 2000 and 30 September 2012.

Acknowledgments

Special thanks go to Dr Golbon Zakeri, Principle Supervisor, for facilitating this research project and for taking me on board as her Part IV student. My thanks extend to her supervision, guidance, patience and support throughout this research project. Golbon has not only helped me gain an extensive knowledge on Operations Research and its application in the New Zealand Electricity Market, but also gave me a great insight to the benefits the Department of Engineering Science (DES) has to offer to solve real world problems.

Special thanks go to Dr Anthony Downward, Supervisor, for stepping in as the second supervisor, monitoring my progress, reviewing my work from start to finish, and providing useful recommendations in the development of this research project. Tony was a key contributor towards my learning during the seven months working on this research project.

Modeling Technology Adoption Decisions where Farmers are Resistant to Change

Simon Anastasiadis and Stefanka Chukova
School of Mathematics, Statistics and Operations Research
Victoria University of Wellington, New Zealand
mail4starzi@gmail.com
stefanka.chukova@ecs.vuw.ac.nz

Abstract

Nutrient emissions from agricultural land are now widely recognized as key contributors to poor water quality in local lakes, rivers and streams. Regulatory intervention to protect and improve water quality appears to be necessary in many cases. However, farmers attitudes suggest that they are resistant to making the changes that would be required under such regulation.

This study develops a model of farmers resistance to change and their adoption of new mitigation technologies under nutrient trading regulation. We specify resistance to change as a bound on the adoption of new technologies and allow this bound to relax as farmers resistance to change weakens.

Key words: agriculture, inertia, mitigation, technology adoption

1 Introduction

Nutrient emissions from non-point sources, such as agricultural land, are increasing recognized as one of the key contributors to poor water quality. Declining water quality is a serious problem in many developed countries, including New Zealand, and in an increasing number of developing countries (Sutton et al. 2011; Parliamentary Commissioner for the Environment 2006). Numerical modeling of different approaches to improving water quality can help inform the decisions of both policy makers and local stakeholders.

Nutrient trading schemes are discussed in the Economics literature as a cost effective approach to managing water quality. The standard models of nutrient trading assume that emitters are willing to change, emitters respond optimally to regulation, and emitters' decisions are independent of their past decisions and the decisions of other emitters.

However, evidence suggests these assumptions are not a good representation of reality for New Zealand farmers. Farmers have expressed a reluctance to change where it involves the adoption of unfamiliar farm management practices or technologies (see for example Fenemor et al. 2012); they tend to manage their business

with a ten or more year time horizon; and may have incentives to delay the adoption of new practices or technologies in order to capitalize on learning opportunities (Coleman and Sin 2012). Furthermore, there is a well known psychological phenomenon where people and organizations continue a familiar practice, even though a better one is available, until the cost of continuing with their current practice exceeds the cost of change (see for example Ram 1987).

Designing a model to reflect the reality identified above is a challenging task. It involves quantifying something that is difficult to identify and measure, namely: farmers' willingness and motivation to change. Farmers' willingness and motivation to change will depend on time, their past behavior, and the past behavior of others farmers. We use the term inertia to describe the nature of a system or agent (farmers) that slows the speed at which they change. While this term appears to be used across a broad range of literature, modeling of agents' inertia does not appear to be widely established practice. Ram (1987) asserts that it is important to study not only adoption but also resistance to adoption. In this paper, we take a novel approach by specifying a model that explicitly includes farmers' inertia.

Research into the factors that affect the adoption on new agricultural practices and technologies have highlighted the importance of social and professional networks, information, and costs. Meta-analyses by Skinner and Staiger (2005), and Baumgart-Getz, Stalker Prokopy, and Floress (2012) show that studies consistently identified network effects on farmers' learning as a key determinant of their technology adoption. Pannell et al. (2006) note that Australian farmers have an excess of information and are almost never passive in their receipt of information. A similar idea is discussed by Hanna, Mullainathan, and Schwartzstein (2012) who demonstrate a model where farmers fail to learn, not because they do not have data, but because they fail to notice important features of the data they already possess.

A less emphasized determinant of farmers' adoption decisions is farmers' attitudes to farming and to change. Dury et al. (2010) interview farmers in France and identify the following farmer objectives: maximizing profit or income, establishing and maintaining a secure source of income, and reducing or simplifying their workload. Connor et al. (2008) and Ward et al. (2008) give classifications of Australian farmers according to their attitudes. The five key elements of farmers' attitudes they identify are: business focus, innovative, willingness to learn, responsiveness to social influence, and environmental concern.

Ellison and Fudenberg (1993) appear to be the first to use the term inertia with respect to modeling agents' or firms' decisions to adopt new technology. They developed an agent-based model of adoption where agents revise their choice of two technologies in response to further learning. In their model not all agents may revise their technology choice each period, a property they call inertia.

The paper is set out as follows: Section 2 describes the general form of our model. Specific functional forms are discussed in section 3, and section 4 concludes.

2 The Inertia Model

A common approach when modeling adoption decisions is to treat adoption as an irreversible binary decision: in some period firms make a step change from the old technology to the new technology. The key question under this framework is when will firms adopt the new technology? We take an alternative approach, rather

than considering the adoption of a specific technology, we treat the adoption of technologies as a continuous decision. In this section we develop a model where farmers decide how much more to adopt new technologies.

A farmer faced with regulatory pressure to reduce nutrient emissions can either adopt new technologies or can attempt to reduce emissions given their current technologies. The adoption of new technologies will improve the cost effectiveness of any nutrient reductions the farmer makes. Although the optimal response to nutrient regulation must involve the adoption of appropriate technologies (and practices) in the long run, farmers are likely to be resistant to making these changes in the short run as they will involve risk and learning new or unfamiliar activities. We will describe this resistance as inertia. Farmers' inertia will decline with time, as they seek more profitable ways of managing their farms, and as farmers observe their neighbors making changes on their own farms.

Consider a catchment containing several farms. Each farm is managed by a farmer who chooses the level of inputs (animals, fertilizer, labor, capital) to put on their farm and how much to adopt new technologies. Farms have two outputs: their production good (e.g.: meat, milk, or fiber) that results in profit that is collected by the farmer, and nutrient emissions that are a byproduct that results in environmental degradation. In general, more intensive farms (i.e.: farms with higher levels of inputs) generate both more profits and more nutrient emissions.

Suppose the community is concerned about the environmental degradation caused by farms' nutrient emissions. However, this concern is not sufficient to motivate farmers to reduce emissions where these reductions are costly, and hence the community has asked the local government body to intervene in the catchment as a regulator. The aim for regulatory intervention is to lower nutrient emissions in the catchment to acceptable levels.

Farmers can theoretically achieve any level of nutrient emissions, without adopting new technologies, by reducing their use of existing inputs. However, lowering emissions via reducing inputs is costly to the farm business as it has a large impact on profit. New mitigation technologies (and the associated practices) are designed to lower the cost to the farmer of reducing nutrient emissions by making reductions more cost effective. A farmer who adopts some new technology will be able to meet any nutrient target at lower cost (with more profit) than a farmer who adopts no technology. It follows that, under regulation, farmers have an incentive to adopt new technology where the gains from cost effective mitigation exceed the cost of investing in the new technology.

While farmers have incentives to adopt new technologies, they may be slow to adopt due to their inertia. Each period, farmers' inertia will determine what new technologies they are willing to adopt. As farmers' inertia changes over time they will become more willing to adopt new technologies.

The optimal decision in this context could be determined by a benevolent social planner. The social planner's aim is to maximize the combined profit of all farmers, given that nutrient emissions from farming (n_{it}) must not exceed specified levels (\bar{N}_t), and that farmers face limits on their adoption of new technologies. To accomplish this, the social planner chooses the use of technology (m_{it}) and inputs (θ_{it}) for all

farmers ($i = 1, \dots, I$) and time periods ($t = 1, \dots, T$).

$$\max_{\forall i, t: \theta_{it}, m_{it}} \sum_{t=1}^T \sum_{i=1}^I \frac{1}{(1+r)^t} \pi_i(x_{it}, \theta_{it}, m_{it}) \quad (1)$$

$$\text{s.t.: } \forall t: \sum_{i=1}^I n_{it}(x_{it}, \theta_{it}, m_{it}) \leq \bar{N}_t, \quad (2)$$

$$\forall i, t: \Delta m_{it} \leq \Delta \bar{m}_{it}(t, m_{i,t-1}, \{m_{j \neq i, t-1}\}), \quad (3)$$

$$\forall i, t: m_{it}, \theta_{it} \geq 0,$$

where x_{it} gives various exogenous factors (such as prices, land quality, soil type, and climate), r is the discount rate, and \bar{N}_t is the maximum acceptable level of nutrient emissions for the catchment (the cap) in period t . The profit function of farmer i , given by $\pi_i(\cdot)$ in (1), includes the revenue from production, and the cost of new technologies and all inputs.

We refer to (3) as farmers' inertia constraint. $\Delta m_{it} = m_{it} - m_{i,t-1}$ is the new technology adopted by farmer i in period t , and $\Delta \bar{m}_{it}(\cdot)$ gives the maximum amount of new technology farmer i is willing to adopt in period t .

We define farmers' inertia as $\bar{m}_{it} = m_{i,t-1} + \Delta \bar{m}_{it}$. This gives the maximum amount of technology that farmer i is willing to have adopted by the end of period t ($\bar{m}_{it} \geq m_{it}$). A farmer's inertia is binding if $\bar{m}_{it} = m_{it}$. As \bar{m}_{it} increases, farmers inertia weakens, and farmers are willing to adopt more technology.

To reflect reality, we require that profit is initially increasing with inputs (up to some level θ_i^u), nutrient emissions are increasing with inputs, and profit exhibits decreasing marginal returns with respect to inputs.

$$\theta_{it} \in [0, \theta_i^u] \rightarrow \frac{\partial \pi_i}{\partial \theta_{it}} \geq 0, \quad \frac{\partial n_{it}}{\partial \theta_{it}} \geq 0, \quad \frac{\partial^2 \pi_i}{\partial \theta_{it}^2} < 0.$$

This ensures that more inputs on a farm result in higher nutrient emissions, and that there exists a finite, non-negative profit maximizing level of inputs, i.e. given the exogenous inputs (x_i) profit maximizing farmers will seek to use the optimal level of inputs (θ_{it}^*) in order to earn the maximum possible profit ($\pi_i^*(\cdot)$).

It follows that farmers can decrease their nutrient emissions (n_{it}) by reducing their inputs (θ_{it}). However, lowering inputs results in reduced farm profits. The adoption of new technologies improves the cost effectiveness of input reductions. So that the same reduction in inputs (to achieve reductions in nutrient emissions) can be achieved with a smaller reduction in profit. We require that the change in profit with respect to inputs ($\partial \pi_i / \partial \theta_{it}$) is initially decreasing with the adoption of technology (up to some level $m_{it}^u(\theta_{it})$).

$$m_{it} \in [0, m_{it}^u(\theta_{it})] \rightarrow \frac{\partial^2 \pi_i}{\partial m_{it} \partial \theta_{it}} < 0.$$

This ensures that there is a financial incentive to adopt technology and that there exists a finite, non-negative optimal level of technology adoption.

The key contribution of our model is the inclusion of farmers' inertia in the decision problem as given by (3). This extends the standard economic model for profit maximization given an environmental constraint. If farmers' have no inertia

(3) will always be non-binding, and our model reduces to the standard economic model (such as that given by Tietenberg 2006, chapter 2) as follows:

$$\begin{aligned} \max_{\forall i,t:\theta_{it},m_{it}} \quad & \sum_{t=1}^T \sum_{i=1}^I \frac{1}{(1+r)^t} \pi_i(x_{it}, \theta_{it}, m_{it}) \\ \text{s.t.:} \quad & \forall t : \sum_{i=1}^I n_{it}(x_{it}, \theta_{it}, m_{it}) \leq \bar{N}_t \quad , \\ & \forall i, t : m_{it}, \theta_{it} \geq 0 \quad . \end{aligned}$$

One approach to finding a solution for the social planner's decision problem is to set it up as a dynamic program and solve using backwards recursion. However, as the state of the decision problem is defined by the set of all farmers' technology decisions in a given period (m_{it} for all i) the problem rapidly becomes computationally intractable as the number of farmers increases.

To make the decision problem tractable we assume that the social planner is short sighted and optimizes within each time period rather than over all time periods. The social planner's decision problem can then be expressed as follows:

$$\forall t : \max_{\forall i:\theta_{it},m_{it}} \sum_{i=1}^I \pi_i(x_{it}, \theta_{it}, m_{it}) \quad (4)$$

$$\text{s.t.} \quad \forall t : \sum_{i=1}^I n_{it}(x_{it}, \theta_{it}, m_{it}) \leq \bar{N}_t \quad , \quad (5)$$

$$\begin{aligned} \forall i, t : \Delta m_{it} &\leq \Delta \bar{m}_{it}(t, m_{i,t-1}, \{m_{j \neq i, t-1}\}) \quad , \quad (6) \\ \forall i, t : m_{it}, \theta_{it} &\geq 0 \quad . \end{aligned}$$

It is then straightforward to solve the problem using forward recursion. The solution to the revised model, given by (4) to (6), will provide a lower bound on the objective function of the original model, and will be a very good approximation to the original model when the discount rate (r) is large. Note that we drop the discounting factor ($(1+r)^{-t}$) in the above decision problem as it reduces to a multiplicative constant in this formulation.

In practice, it is difficult for a social planner to obtain farm specific information (i.e. to observe $\pi_i(\cdot)$ or $n_{it}(\cdot)$). This hinders their ability to solve the decision problem specified by (4) to (6).

An alternative to solving the social planner's decision problem is to establish a nutrient trading scheme. Under a nutrient trading scheme the regulator provides a supply of allowances equal to \bar{N}_t , with each allowance entitling the owner to emit one unit of nutrients. Farmers can trade allowances between themselves, but must surrender sufficient allowances to cover their emissions at the end of each period. The equilibrium price of allowances (p_t) is determined by the profit that can be earned from one unit of emissions. Given this price, farmers' individual decisions will be identical to the equivalent decisions made by the social planner (this can be demonstrated via the Karush-Kuhn-Tucker conditions).

3 Specific Functional Forms

In this section we propose and justify specific functional forms for farmers' profit (π_i) and inertia functions ($\Delta \bar{m}_{it}$). These are the functional forms we will use in

future analysis on this model.

So far, we have used θ_{it} to represent the quantity of inputs. However, in reality there are many farm inputs (including stock, fertilizer, imported feed, fencing materials, and labor) and the combination of these inputs, not just the quantity, is important. Furthermore the same quantity of nutrient emissions can be achieved using many different combinations of inputs. Incorporating in our model the process by which inputs are converted to profits and emissions is unnecessary and detracts from our focus on farmers' inertia.

For any given farm-level nutrient target a skilled farmer could determine the profit maximizing combination of inputs that will meet this nutrient target. This implies that we do not need to consider the set of all possible combinations of inputs but can focus on the subset that would be used by a skilled farmer to maximize profit. Note that by construction there exists a bijection between this subset and the set of all feasible n_{it} . Hence rather than treating θ_{it} as a decision variable (so n_{it} is determined implicitly) it will be equivalent to treat n_{it} as the decision variable (so θ_{it} is determined implicitly). For our specific functional forms we will treat m_{it} and n_{it} as decision variables. To simplify the notation, we will suppress x_{it} and θ_{it} .

Recall from the previous section, that we require that there exists a finite, non-negative profit maximizing level of inputs, and that the adoption of technology improves the cost effectiveness of reductions in inputs. When we consider n_{it} as a decision variable, these requirements are expressed as follows: profit is initially increasing with nutrient emissions (up to some level n_i^u); marginal profit is decreasing with nutrient emissions; and the change in profit with respect to nutrient emissions is initially decreasing with the adoption of technology (up to some level $m_{it}^u(n_{it})$).

$$n_{it} \in [0, n_i^u] \rightarrow \frac{\partial \pi_i}{\partial n_{it}} \geq 0, \quad \frac{\partial^2 \pi_i}{\partial n_{it}^2} < 0, \quad m_{it} \in [0, m_{it}^u(n_{it})] \rightarrow \frac{\partial^2 \pi_i}{\partial m_{it} \partial n_{it}} < 0 .$$

The first two requirements ensure that higher profit from a farm results in increased emissions, and that there exists a finite, non-negative profit maximizing level of emissions. The third requirement ensures that farmers have a financial incentive to adopt technology, and that there exists a finite, non-negative optimal level of technology adoption.

3.1 Farm profit function

We will express farmers' profits as a function of their nutrient emissions (n_{it}) and technology (m_{it} , measured as the amount of nutrients the adopted technology can cost effectively mitigate). We propose the following functional form.

$$\pi_i(m_{it}, n_{it} | m_{i,t-1}) = a_i n_{it}^2 + b_i n_{it} + c_i + d_i (e_i - n_{it} - m_{it})^2 - q_i (m_{it} - m_{i,t-1}) - p_t n_{it} , \quad (7)$$

where a_i , b_i , c_i and d_i are coefficients ($a_i, d_i < 0$ and $b_i > 0$), e_i is the farm's baseline emissions (at time $t = 0$ before regulation is introduced), $p_t \geq 0$ is the equilibrium price of nutrient allowances (as determined in the nutrient trading scheme), and $q_i \geq 0$ is the cost of new technology to farmer i .

This form extends that used by Anastasiadis et al. (2011), and can intuitively be decomposed into four parts: The first three terms give profit as a function of nutrient emissions; the fourth term provides a measure of the difference between farmers' current nutrient reductions ($e_i - n_{it}$) and the nutrient reductions their

technology can carry out cost effectively (m_{it}); the fifth and sixth terms account for the cost of the farm inputs (additional technology and allowances respectively). We choose a quadratic form for the profit function for its simplicity and because it results in linear (or piece-wise linear) demand functions, as we will see below.

Figure 1 gives an example of this functional form for $\pi_i(m_{it}, n_{it})$ excluding the cost of allowances. The gray line gives farmers' profit if they have no resistance to change. The black lines demonstrate farmers' profit functions given bounds on their adoption of technologies (\bar{m}_{it} equals 10 and 30).

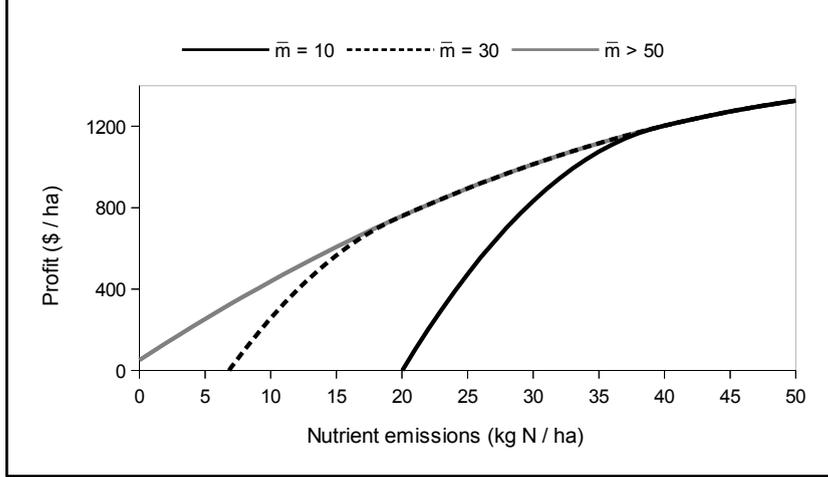


Figure 1. Farm's profit per hectare for levels of inertia

It is clear from figure 1 that farm profit is decreasing as nutrient emissions decrease, and furthermore that profit decreases by more when farmers' inertia is binding. The relaxing of the farmer's inertia constraint and the further adoption of technology improves farmers' profits as it moves the dashed black curve towards the gray curve, and the solid black curve towards the dashed black curve.

Given the functional form of a farm's profit function we can determine a farmer's optimal decision as a function of the price of allowances and their inertia as follows:

Let $m_{it} < \bar{m}_{it}$:

Then $\frac{\partial \pi_{it}}{\partial m_{it}} = 0$ implies $-2d_i(e_i - n_{it} - m_{it}) - q_i = 0$, so:

$$m_{it}^* = e_i - n_{it} + \frac{q_i}{2d_i} \quad (8)$$

And $\frac{\partial \pi_{it}}{\partial n_{it}} = 0$ implies $2a_i n_{it} + b_i + q_i - p_t = 0$, so:

$$n_{it}^* = \frac{p_t - b_i - q_i}{2a_i} \quad (9)$$

Let $m_{it} = \bar{m}_{it}$:

Then $\frac{\partial \pi_{it}}{\partial n_{it}} = 0$ implies $2a_i n_{it} + b_i - 2d_i(e_i - n_{it} - \bar{m}_{it}) - p_t = 0$, so:

$$n_{it}^* = \frac{p_t - b_i + 2d_i(e_i - \bar{m}_{it})}{2a_i + 2d_i} \quad (10)$$

From (8) to (10) it is clear that there are two key reasons why farmers may not adopt technology. First, farmers' inertia (\bar{m}_{it}) in period t may be binding. Second, the benefits of adopting new technologies may not exceed the cost. So the costs of new technologies (q_i) are too high in comparison to the improvements in the cost effectiveness of nutrient reductions ($-d_i$) that come from adopting

For a given allowance price, we can determine the \bar{m}_{it} value such that a farmer's inertia is only just binding. This occurs when $\bar{m}_{it} = e_i - n_{it} + q_i/2d_i$ and nutrient emissions are given by (10). Substituting one into the other gives:

$$\bar{m}_{it} = e_i - \frac{p_t - b_i - q_i}{2a_i} + \frac{q_i}{2d_i}. \quad (11)$$

So for a given allowance price, farmers with \bar{m}_{it} lower than the \bar{m} given by (11) will have a binding inertia constraint and farmers with \bar{m}_{it} higher than \bar{m} given by (11) will have a non-binding inertia constraint.

Given the price at which farmers' inertia becomes binding from (11) we can construct a farmer's demand for allowances as a function of price of allowances. When the price is less than the price that satisfies (11), the farmer's inertia is non-binding and their demand is given by (9). When the price is more than the price that satisfies (11), the farmer's inertia is binding and their demand is given by (10). Figure 2 gives examples of farmers' demand functions that correspond to the example profit functions given in Figure 1.

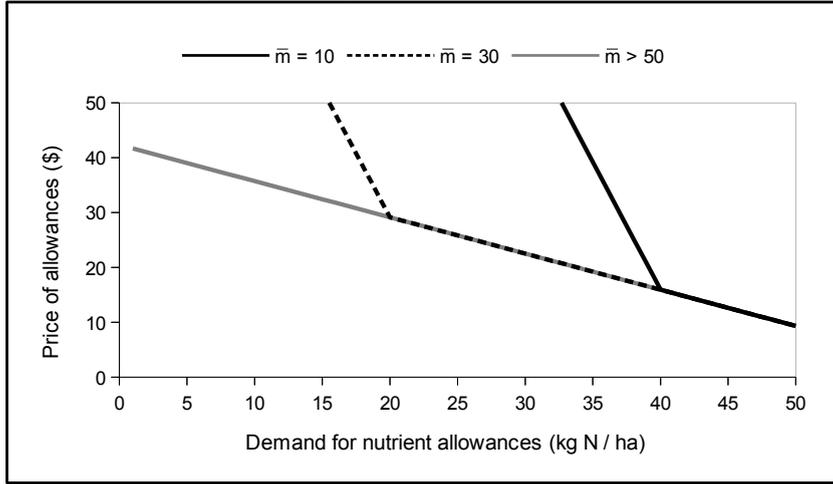


Figure 2. The demand for allowances for different levels of inertia

It is clear from figure 2 that farmers' demand for allowances is decreasing as the price of allowances rises. Furthermore, we observe that farmers' demand for allowances is higher, at any given price, when their inertia is binding. This is to be expected as, from figure 1, the loss of profit from reducing emissions is much higher when a farmer's inertia is binding than when it is non-binding.

3.2 The inertia function

The form of the inertia function $\Delta\bar{m}_{it}(\cdot)$ is independent of the functional form chosen for the farmers' profit functions. We propose the following functional form. We use a linear form for simplicity in the absence of strong preferences for any other shape.

$$\Delta\bar{m}_{it}(\tau_i, m_{i,t-1}, p_{t-1}, \{m_{j \in I_i, t-1}\}) = \begin{cases} 0 & \text{if } g_{i,t} \leq \delta_i \\ g_{i,t} - \delta_i & \text{if } g_{i,t} > \delta_i \end{cases} \quad (12)$$

with $g_{i,t}$ a measure of the pressure on farmer i to adopt new technologies. I_i gives the subset of all farmers (I) that farmer i might learn from, including themselves ($i \in I_i$). We propose the following function form for $g_{i,t}$:

$$g_{i,t} = \alpha_i e_i \tau_i + \beta_i (m_{i,t-1}^* - m_{i,t-\tau_i}) + \gamma_i (\max_j \{m_{j \in I_i, t-1}\} - m_{i,t-\tau_i}), \quad (13)$$

where α_i , β_i and γ_i are coefficients ($\forall i : \alpha_i, \beta_i, \gamma_i \in [0, 1]$), δ_i is a farmer specific threshold, τ_i gives the time since the farmer last adopted new technologies, and $m_{i,t-1}^*$ gives the technology that would have been optimal last period (as determined by equations 8 and 9).

A farmer's inertia depends on the time since they last adopted new technologies (τ_i), the difference between their technology and the technology that would have been optimal last period ($m_{i,t-1}^* - m_{i,t-\tau_i}$), and the difference between their technology and the most technology used by any comparable farmer ($\max_j \{m_{j \in I_i, t-1}\} - m_{i,t-\tau_i}$).

We use α_i to capture how often the farmer i replaces or updates their existing technologies. Farmers with low α values will be more likely to have long delays between changes in technology, perhaps because they are capital constrained. We interpret $\alpha_i = 1$ as indicating that a farmer considers updating all their existing technologies every year, and $\alpha_i = 0$ as indicating that a farmer never considers updating their existing technologies.

We use β_i to capture the business focus of the farm ($\beta_i \in [0, 1]$). Farmers with high β values will be more likely to adopt additional technology when their current technologies are not optimal, this may be because they are more comfortable innovating. We interpret $\beta_i = 1$ as indicating that a farmer is always motivated to adopt technology in order to earn more profit, and $\beta_i = 0$ as indicating that a farmer is never motivated to adopt technology in order to earn more profit.

We use γ_i to capture a farmer's willingness and ability to learn from other farmers ($\gamma_i \in [0, 1]$). Farmers with high γ values will be more likely to adopt technology when other farmers have already done so, this may be because they are receptive to being social influenced or because they are resistant to adopting new technologies until they have observed these being used on other farms. We interpret $\gamma_i = 1$ as indicating that a farmer is continually seeking to learn from their networks, and $\gamma_i = 0$ as indicating that a farmer never seeks to learn from their networks.

We use δ_i to capture overall resistance to change ($\delta_i \geq 0$). Farmers with high δ values will be less likely to adopt new technologies, this may be because they have low environmental concern or prefer traditional methods of farming.

4 Moving Forward

We have specified a model for farmers' adoption of new technology in response to nutrient regulation. Given the functional forms for our model, future work will demonstrate the range of possible behaviors that the model can produce. In addition we should like to compare the inertia model against the simple model that ignores inertia. This should enable us to estimate the cost of reaching a given nutrient target when farmers short term response to regulation is less cost effective than their eventual response.

References

- Anastasiadis, Simon, Marie-Laure Nauleau, Suzi Kerr, Tim Cox, and Kit Rutherford. 2011. "Does complex hydrology require complex water quality policy? NManager simulations for Lake Rotorua." Motu working paper 11-14, Motu Economic and Public Policy Research.

- Baumgart-Getz, Adam, Linda Stalker Prokopy, and Kristin Floress. 2012. "Why farmers adopt best management practice in the United States: A meta-analysis of the adoption literature." *Journal of Environmental Management* 96:17–25.
- Coleman, Andrew, and Isabelle Sin. 2012. "The adoption of environmentally friendly technologies in agriculture." Technical Report, Motu Economic and Public Policy Research. Motu Note.
- Connor, Jeffery D., John Ward, Craig Clifton, Wendy Proctor, and Darla Hatton MacDonald. 2008. "Designing, testing and implementing a trial dryland salinity credit trade scheme." *Ecological Economics* 67:574–588.
- Dury, J., F. Garcia, A. Reynaud, O. Therond, and JE. Bergez. 2010. "Modelling the complexity of the cropping plan decision-making." Edited by David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, and Filatova T., *International Congress on Environmental Modelling and Software Modelling for Environments Sake*.
- Ellison, Glenn, and Drew Fudenberg. 1993. "Rules of thumb for social learning." *Journal of Political Economy* 101 (4): 612–643.
- Fenemor, Andrew, Jim Sinner, Simon Anastasiadis, Nick Cradock-Henry, Hana Crengle, Simon Harris, John Bright, Suzie Greenhalgh, and Suzi Kerr. 2012. Simulating Market Based Instruments (MBIs) for Water Allocation and Quality in New Zealand. Report prepared for Ministry of Primary Industries.
- Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2012. Learning through noticing: Theory and experimental evidence in farming. NBER Working Paper 18401.
- Pannell, D. J., G. R. Marshall, N. Barr, A. Curtis, F. Vanclay, and R. Wilkinson. 2006. "Understanding and promoting adoption of conservation practices by rural landholders." *Australian Journal of Experimental Agriculture* 46:1407–1424.
- Parliamentary Commissioner for the Environment. 2006. "Restoring the Rotorua Lakes: The Ultimate Endurance Challenge." Technical Report, Parliamentary Commissioner for the Environment, Wellington.
- Ram, S. 1987. "A model of innovation resistance." *Advances in Consumer Research* 14:208–212.
- Skinner, Jonathan, and Douglas Staiger. 2005. "Technology adoption from hybrid corn to beta blockers." Technical Report 11251, NBER working paper series.
- Sutton, M.A., O. Oenema, J. W. Erisman, A. Leip, H. van Grinsven, and W. Winiwarter. 2011. "Too Much of a Good Thing." *Nature* 472:159–161.
- Tietenberg, T. H. 2006. *Emissions Trading: Principles and Practice*. 2nd ed. Resources for the Future Press.
- Ward, J.R., B.A. Bryan, N.D. Crossman, and D. King. 2008. "Evidence based farmer decision models to assess the potential for multiple benefit carbon trading." *International Association for the Study of Commons, conference July 2008*.

Inference for Multicomponent Systems with Dependent Failures

Richard Arnold, Stefanka Chukova
School of Mathematics, Statistics and Operations Research
Victoria University of Wellington, New Zealand
{*richard.arnold, stefanka.chukova*}@msor.vuw.ac.nz

Yu Hayakawa
School of International Liberal Studies
Waseda University, Tokyo, Japan
yu.hayakawa@waseda.jp

Abstract

We present a general approach to inference in Independent Overlapping Subsystem models, where a component's failure time is the time of the earliest failure in all of the subsystems of which it is a part, and each of those subsystems has an independent failure process. We apply this method to observations of an IOS model that associates individual shock processes with sets of overlapping subsystems made up of groupings of components, giving examples for various system configurations (series, parallel, and other arrangements).

Key words: Reliability, Multicomponent Systems

1 Introduction

This short paper explores methods for statistical inference for the model with correlated failures among components in multicomponent systems introduced by Anastasiadis et al. (2010). This model is a member of a general class where correlated component failures are generated by associating each component with one or more subsystems (groups of components). Each subsystem has an independent failure process associated with it, and a separate failure time is randomly generated from the distribution of its failure process. The failure time of each individual component is the earliest failure time of any of the subsystems of which it is a part. We refer to such models as 'Independent Overlapping Subsystem' (IOS) models, and this class includes the multivariate Marshall-Olkin exponential model (MVE) Marshall and Olkin (1967). Full details of the results presented here may be found in Arnold, Chukova, and Hayakawa (2012), and only a brief summary is included below.

There are existing inference procedures for the MVE model, in particular the bivariate exponential (BVE) model. Proschan and Sullo (1976) obtained an estimator which converges to the Maximum Likelihood Estimate (MLE) for large samples.

Bayesian inference for the BVE was presented by Shamseldin and Press (1984) under parallel sampling assuming identical marginals, and by Pena and Gupta (1990) for series and parallel structures. Karlis (2003) derived an Expectation Maximisation (EM) algorithm for ML estimation for MVE and an empirical Bayes method for BVE was given by Hanagal and Ahmadi (2009). Our paper extends and generalises this earlier work.

2 General Inference Framework

We consider a system with m components, labelled $j = 1, \dots, m$. Groups of these components form **subsystems**. We assume that $M \leq 2^m - 1$ of these subsystems have distinct, independent failure processes associated with them. A failure time t_k is associated with each subsystem k according to the failure process of that subsystem. The failure time of **component** j is then the earliest failure time of all the subsystems of which it is a part.

The system is observed for an **observation time** τ_0 , a set of $0 \leq H \leq m$ distinct failure times $0 < T_1 < T_2 < \dots < T_H \leq \tau_0$ are observed, and no repairs are made. At time T_h the components labelled by indices in the set A_h fail. The entire system has a **structure** (e.g. series or parallel if $m = 2$), defined by the collection of minimal cut sets \mathbb{C} . The system either fails at T_H or the system continues to be observed until the observation time τ_0 . Thus The **system censoring time** is $\tau_S = T_H$ if the sytem fails at T_H , otherwise it is the observation time $\tau_S = \tau_0$. An **observation** of the system $(\mathbf{T}, \mathbf{A}, \tau_S)$ thus consists of the (ordered) failure time H -tuple (T_1, \dots, T_H) , the collection of H **failure sets** (A_1, \dots, A_H) , and the system censoring time τ_S .

We assume that each subsystem has a specified absolutely continuous survival function $\bar{F}_k(t|\Psi_k)$, a density $f_k(t|\Psi) = g_k(t|\Psi_k)\bar{F}_k(t|\Psi_k)$ (for parameters Ψ_k), and that when a subsystem fails all of its member components fail. The likelihood for a set of n observations can then be written

$$L(\Psi|\{y_i\}) = \prod_{i=1}^n \left\{ \left[\sum_{\mathbf{k}_i \in \kappa(\mathbf{A}_i)} \prod_{h=1}^H g(T_{ih}|\Psi_{k_{ih}}) \right] \left[\prod_{k=1}^M \bar{F}(\tau_{ik}|\Psi_k) \right] \right\} \quad (1)$$

Here τ_{ik} is the time that subsystem k was either observed to fail, or the time beyond which failures of that subsystem became unobservable, and $\kappa(\mathbf{A}_i)$ is the set of possible sequences of subsystem failures that are consistent with the observation.

Example. Five component bridge system, arranged as shown in Figure 1, with failure processes associated with the $M = 10$ subsystems $\mathbb{S} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}, \{1, 2, 5\}, \{1, 2, 3, 4, 5\}\}$.

Consider an example observation of $H = 3$ distinct failure times: $(\mathbf{T}, \mathbf{A}, \tau_S) = ((T_1, T_2, T_3), (\{1\}, \{2, 3\}, \{5\}), \tau_S)$. In the bridge the 4-5 path operates until the failure of component 5 at T_3 causes system failure, and so $\tau_S = T_3$. There are four possible sequences of subsystem failure consistent with the observation: $\kappa(\mathbf{A}) = \{(\{1\}, \{2, 3\}, \{5\}), (\{1\}, \{1, 2, 3\}, \{5\}), (\{1\}, \{2, 3\}, \{1, 2, 5\}), (\{1\}, \{1, 2, 3\}, \{1, 2, 5\})\}$ leading to the likelihood

$$L(\Psi|y) = g_1(T_1)[g_{23}(T_2)g_5(T_3) + g_{123}(T_2)g_5(T_3) + g_{23}(T_2)g_{125}(T_3) + g_{123}(T_2)g_{125}(T_3)] \\ \times \bar{F}_1(T_1)\bar{F}_2(T_2)\bar{F}_3(T_2)\bar{F}_4(T_3)\bar{F}_5(T_3)\bar{F}_{12}(T_2)\bar{F}_{23}(T_2)\bar{F}_{123}(T_2)\bar{F}_{125}(T_3)\bar{F}_{12345}(T_3)$$

Bridge
(1&2)|(1&3&5)|(4&5)|(2&3&4)

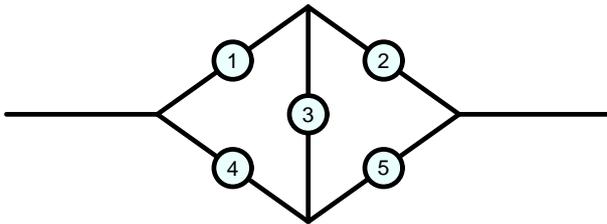


Figure 1: Five component bridge system.

3 Inference Example

We now apply the results of Section 2 to particular cases of the IOS model presented by Anastasiadis et al. (2010) and Anastasiadis et al. (2013).

In this model the log survival function of m_k -component subsystem S_k is

$$\log \bar{F}_k(t) = \mu_{k\ell} \left(-t + \frac{1}{\lambda_k m_k} [1 - e^{-\lambda_k m_k t}] \right) \quad (2)$$

where μ_k is the arrival rate and λ_k the strength of a sequence of shocks which damage the system. There are $2M$ parameters: two for each subsystem $\Psi = \{(\mu_k, \lambda_k)\}_{k=1}^M$. The minimal sufficient statistics are the observations themselves $\{y_i = (\mathbf{T}_i, \mathbf{A}_i, \tau_{Si})\}_{i=1}^n$.

We demonstrate our inference procedure for the complex system shown in Figure 2. This is a distributed computer network with 8 equivalent parallel paths, each with four processing tasks (Figure 2(a)). Each task is itself achieved by 4 parallel processors (Figure 2(b)). There are thus 128 individual components in all, each with its own shock process ($\mu_j = 1, \lambda_j = 1$). Each of the 32 task subsystems (the squares in Figure 2(a)) also has its own subsystem shock process. The first task subsystem (labelled ‘1’ in Figure 2(a)) is the weakest ($\mu_j = 1, \lambda_j = 1.25$), and the remaining task subsystems are stronger ($\mu_j = 1, \lambda_j = 0.125$). We thus have a total of $128+32=160$ subsystems. For inference purposes we assume that the 128 individual components are identical, and that the task subsystems in equivalent positions on each branch are also identical. By setting the parameters of many of the subsystems to be equal to one another in this way we reduce the dimension of parameter space from 320 to 10.

In 100 simulations of this system no system failures are seen up to an observation time of $\tau_0 = 2$. Figure 3(a) shows the distribution of subsystem failure times for 10 of the simulations. Task subsystem failures are indicated with circles. The log likelihood surface for the parameters of the task 1 subsystem (Figure 3(b)) is very flat and shows the negative correlation that exists between the shock arrival rate μ and the shock strength λ .

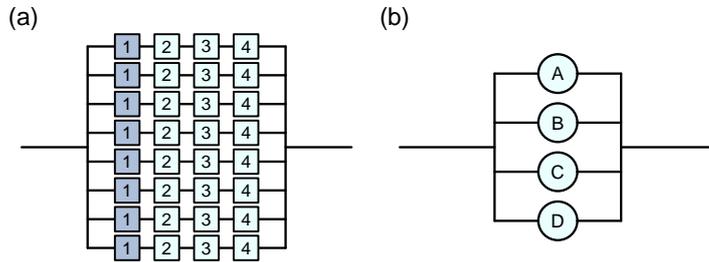


Figure 2: (a) Example 128 component system. Every square in (a) represents a separate four component parallel subsystem as shown in (b). See text for further details.

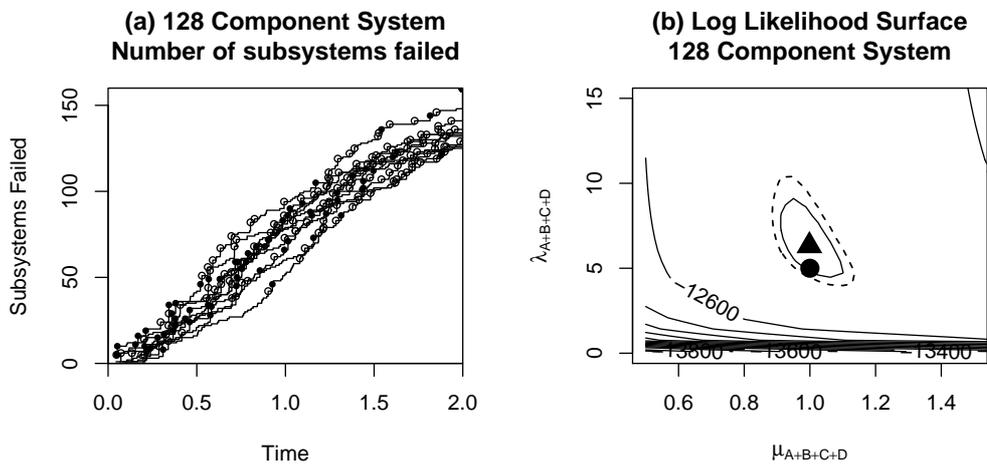


Figure 3: (a) Cumulative frequency distribution of number of subsystems failed for 10 simulations of the 128 component system from Figure 2. Failures of task subsystems are marked with circles (filled circles for task 1 subsystems). (b) Log likelihood surface of the rate and shock strength of the task 1 subsystem.

4 Summary

Our method generalises existing inferential approaches for multicomponent systems by treating systems as assemblies of independent overlapping systems. In complex systems the parameter space dimension may be very high, so simplifying assumptions are needed. In particular, only a subset of possible subsystems should be included in the likelihood, and assumptions can be made about subsystems having identical properties.

References

- Anastasiadis, S., R. Arnold, S. Chukova, and Y. Hayakawa. 2010. “Failures in Multicomponent Systems.” Edited by S. Chukova, J. Haywood, and T. Dohi, *Advanced Reliability Modeling IV, Proceedings of the 4th Asia-Pacific Symposium on Advanced Reliability and Maintenance Modelling*. Taiwan: McGraw Hill, 9–16.
- . 2013. “Multicomponent System with Dependent Failures: Modeling and Simulation.” To appear: *IEEE Transactions on Reliability*.
- Arnold, R., S. Chukova, and Y. Hayakawa. 2012. “Inference for Multicomponent Systems with Dependent Failures.” Edited by H. Yamamoto, C. Qian, L. Cui, and T. Dohi, *Advanced Reliability Modeling V, Proceedings of the 5th Asia-Pacific Symposium on Advanced Reliability and Maintenance Modelling*. Taiwan: McGraw Hill, 9–16.
- Hanagal, D. D., and K. A. Ahmadi. 2009. “Bayesian Estimation of the Parameters of Bivariate Exponential Distributions.” *Communications in Statistics - Simulation and Computation* 38:1391–1413.
- Karlis, D. 2003. “ML Estimation for Multivariate Shock Models via an EM Algorithm.” *Annals of the Institute of Statistical Mathematics* 55:817–830.
- Marshall, A. W., and I. Olkin. 1967. “A Multivariate Exponential Distribution.” *Journal of American Statistical Association* 62:30–44.
- Pena, E. A., and A. K. Gupta. 1990. “Bayes Estimation for the Marshall-Olkin Exponential Distribution.” *Journal of the Royal Statistical Society. Series B* 52:379–389.
- Proschan, F., and P. Sullo. 1976. “Estimating the Parameters of a Multivariate Exponential Distribution.” *Journal of American Statistical Association* 71:465–472.
- Shamseldin, A. A., and S. J. Press. 1984. “Bayesian Parameter and Reliability Estimation for a Bivariate Exponential Distribution.” *Journal of Econometrics* 24:363–378.

Dynamics of Mergers & Acquisitions Integration

Shanie Atkinson and Michael Shayne Gary
Australian School of Business – School of Management
University of New South Wales
Australia
shanie.atkinson@unsw.edu.au

Abstract

Prior research suggests 60-83 percent of merger and acquisition (M&A) transactions fail to deliver the value that initially motivated the acquisition. In many M&A transactions, a substantial amount of the potential value is lost during the integration phase. This paper reports the findings from in-depth fieldwork investigating how the M&A integration process unfolds over time. Data was collected from semi-structured interviews with 17 M&A integration specialists; drawing on more than 190 merger integration experiences. System Dynamics methods were utilised to identify four common patterns of behaviour over time for M&A integration outcomes. Causal loop diagrams were used to structure and analyse the interview data to identify the causal relationships responsible for successful and unsuccessful M&A integrations. The result was an integrated causal diagram of the feedback processes driving the four common dynamic patterns of behaviour. Our findings show that managerial pressure to generate new synergies – after discovering that some of the initially identified synergies are unachievable – can result in ‘synergy fatigue’. This undermines commitment, focus and engagement, which activate a host of reinforcing feedback processes reducing revenue and cost synergy efforts, reversing previously captured synergy benefits, and negatively impacting performance of the ongoing business.

1 Introduction

The majority of corporate merger and acquisition (M&A) transactions fail to deliver post-acquisition improvements in performance that motivated the transaction (King, Dalton, Daily & Covin 2004) and 26% deliver a loss for the acquirer (KPMG 2006). The objective of this research is to provide new theoretical insights into the processes and policies that drive less successful outcomes from the M&A integration process. This research focuses on the integration phase of the M&A process, the processes and policies in M&A integration have been identified as important determinants of success for M&As and contribute to the poor outcomes recorded for value creation from M&A transactions (Cording, Christmann & King 2008, Haspeslagh & Jemison 1991, Jemison & Sitkin 1986, Schweiger & Goulet 2000).

This inductive study involving M&A integration experts investigates the causal relationships and dynamics in the integration process. Data was gathered from semi-structured individual and focus group interviews. The tools of System Dynamics (Sterman 2000) were applied to structure and analyse the data, to identify the patterns of behaviour in outcome measures, and provide insight into the causal dynamics driving the patterns of behaviour.

The outcomes of the research include the identification of four patterns of behaviour over time for common M&A integration outcome measures, the identification of causal relationships between factors in the process, and an integrated causal diagram of the feedback processes driving the four patterns of behaviour¹. The analysis shows that managerial pressure to generate new synergies can result in ‘synergy fatigue’, which activates a number of undesired feedback effects that undermine the value created in the M&A integration process.

2 Prior Research on M&A Integration in Strategic Management

There is a long history of strategic management research in the area of M&A integration (Cording, Christmann & Bourgeois 2002). Jemison and Sitkin (1986) recognised the M&A process as an important factor in determining the success of M&A transactions. The concept was further developed by the influential work of Haspeslagh and Jemison (1991), who distinguished between the initial decision-making process problems and integration process problems of the M&A process. Haspeslagh and Jemison’s (1991) work defined broad management approaches to M&A integration, but did not address the dynamics within the M&A integration process.

Much of the empirical research on the M&A process has focused on the decision-making process problems. However, there is increasing recognition that more work is needed to understand integration process problems. Homburg and Bucerius (2006) highlight the effect of speed of integration on the success of M&A transactions. They found the effect of speed on integration depended on the level of external and internal relatedness of the acquired and acquiring firms.

Cording et al. (2008) found that intermediate goals assist in the achievement of acquisition goals. Cording et al. theorise that intermediate goals “break down the complex causal chain between integration decisions and acquisition performance into more manageable segments” (Cording et al, 2008: p.759). Gaining a deeper understanding of the complex causal chain referred to by Cording et al. is of interest to this research. Cording et al. encourage further research to understand “what actions lead to what outcomes” (p.760), a suggestion that aligns with the objective of this research.

Other research in M&A integration has identified individual factors that affect the results of the transaction for the acquirer. Failure of the M&A integration process may result from difficulties in coordinating activity due to cultural differences and conflict in the process (Weber & Camerer 2003). Also, the level of M&A experience impacts the outcome of transactions and depends on the relatedness of industrial environments. Integration experience in similar industries has positive transfer effects on acquisition performance (Finkelstein & Haleblan 2002). Other research shows that management team turnover has a detrimental effect on post-acquisition performance and the instability has a flow-on effect to employees (Cannella & Hambrick 1993). In addition, prior work shows that negative effects on employees, including employee distress that has been induced by the integration process, can be alleviated by a high level of honest and realistic communication by management (Schweiger & DeNisi 1991, Schweiger, Ivancevich & Power 1987).

Overall, after four decades of research into M&A’s, we have only a piecemeal

¹ A fifth pattern was identified, it occurred where integration was expected but no integration process was initiated. We have chosen not to discuss the fifth pattern in this paper due to the constraints of space and the limited additional insights it provided.

understanding of the dynamics at work in the M&A process (Bower 2001). Strategy research in the area of M&A integration remains fragmented and has not been systematically linked to theory development (Schweiger & Goulet 2000). The evidence from prior findings provides a list of factors that affect the outcome of M&A integrations, but the continued failure of the process hints that there are factors and effects at play in the M&A integration process that interact to deliver unexpected, puzzling outcomes (King, Dalton, Daily & Covin 2004). Research is needed to systematically link managerial processes and policies to successful outcomes throughout the merger integration process (Schweiger & Goulet 2000) and to investigate the dynamics driving the outcomes of the process. These gaps in the existing literature motivated our research questions:

1. What are the dynamic patterns of behaviour and outcomes observed over time in M&A integration?
2. What are the causal relationships responsible for driving those patterns of behaviour? This includes the effects of different integration processes and policies, with an emphasis on understanding the drivers of poor, unsuccessful M&A integrations.
3. What processes and policies lead to poor outcomes for the M&A integration process?

3 Methodological Approach

We apply a grounded theory approach (Strauss & Corbin 1990) to allow theoretical insights to come to light from the analysis. Answering the research questions required gaining knowledge about the most common dynamic patterns of behaviour for M&A integration outcomes and the causal relationships driving those patterns. M&A integration experts were considered to be a credible source for this information. These experts included experienced consultants specializing in M&A integration and also experienced senior managers that were involved in multiple M&A integrations. Data collection techniques included individual and focus group interviews with M&A integration experts.

The initial phase of individual interviews involved 17 M&A integration professionals with over 190 M&A integration experiences and extended over a five-month period. The interviews were semi-structured. Where possible, the interviews were recorded. Each interview was approximately 90 minutes long. Hand written notes were taken during the interview. Interview notes were typed up immediately following each interview. The interview notes attempted to capture the key points from the discussion and reproduce the patterns of behaviour that were drawn by interviewees. In greater than 70% of the interviews, the interview notes were sent to the interviewee and the interviewee was asked if the notes captured the essence of the discussion and diagrams correctly. All agreed that the interview notes accurately reflected the discussion.

Interview recordings were transcribed and the rich data captured in the transcription was coded and analysed. As recommended by Glaser and Strauss (1967), analysis alongside data collection facilitated the use of probing during the interviews. Probing allowed a deeper investigation of the points of interest that were emerging from the analysis (Eisenhardt 1989).

The data were analysed using the tools of System Dynamics (Sterman 2000). Casual loop diagrams were constructed iteratively throughout the period of data collection.

Insights into the dynamics of the process occurred as the causal loop diagrams captured connections between the factors, processes and policies, which arose from the interviews.

Following the conclusion of the individual interviews, two focus group interviews were held. One was held in the Melbourne office and one in the Sydney office of the leading team of M&A integration consultants in Australia. The Melbourne focus group involved 5 experts, all of whom had been previously interviewed as part of the initial stage of interviews. The Sydney focus group involved 7 M&A integration experts of whom two had been previously interviewed in the initial stage of individual interviews.

During the focus group interviews, the findings from the individual interview phase were presented including the patterns of behaviour and the causal loop diagrams. The focus groups were asked if they agreed with the conclusions when they reflected on their professional experience. The outcomes from the focus group interviews included confirmation of the captured patterns of behaviour, suggestions for further patterns, and extensions to the causal loop diagram. The findings and analysis are presented in the following two sections.

4 Findings

Four patterns of behaviour in four outcome measures are commonly observed in M&A integration. These patterns or integration scenarios are shown in Figure 1. The integration scenarios have been labelled: Fulfilled expectations, Over performance, Creep, and Death spiral. The patterns are captured in terms of the outcome measures: Synergy cost savings, Revenue synergies, Customer retention, and Employee engagement and satisfaction. The Creep and Death spiral patterns will be our focus as these reflect the most problematic behaviour, the Fulfilled expectations and Over performance patterns are briefly explained.

The Fulfilled expectations pattern is labelled number 1 in Figure 1 and occurs approximately 20% of the time. In this scenario the integration rolls out as planned and the target outcome measures are achieved as forecast. The Over performance pattern, the line labelled number 2 in Figure 1, occurs when the performance in all of the outcome measures is higher and achieved earlier than expected. This scenario is similar to the Fulfilled expectations scenario but higher due to management involvement in planning pre-deal; this involvement drives high quality planning and commitment and above forecast outcomes. As one expert explained, “The best run processes feel very simple...it is about having people on the hook all the way through...the people who are ultimately responsible for doing the integration”.

The Creep pattern is labelled number 3 in Figure 1 and occurs to some extent in the majority of M&A integrations. Under the Creep scenario, synergies are initially achieved as planned, but then the energy and enthusiasm for synergy initiatives wane, synergy fatigue sets in, tracking stops, and management commitment and focus move away. Synergy fatigue is described by one expert: “...if the synergies that you came up with upfront aren’t right, if your assumptions were bad, you need to go out and find some more synergies, because we still need to realise that. People actually get fatigued and tired of continually trying to find and chase synergies.”

As a result, “creep” occurs in the outcomes delivered, a movement back towards the original position. In the creep scenario once integration plan tracking stops, there is a claw back of synergy cost savings. For example, employees that were made redundant as part of the cost saving plans, are re-employed as contractors. One expert, while

drawing the creep pattern for synergy cost savings, explains that creep arises due to the failure of the governance process: "...what we see on a lot of deals, there's one where it starts really quickly, then it levels out much earlier and doesn't get up, so you end up with this big value gap over here [referring to right hand side of the diagram]...the issue is normally around the process, because they should have been picking up the errors much earlier on...the issues are more governance-related and how you can actually bring that back to running a robust program."

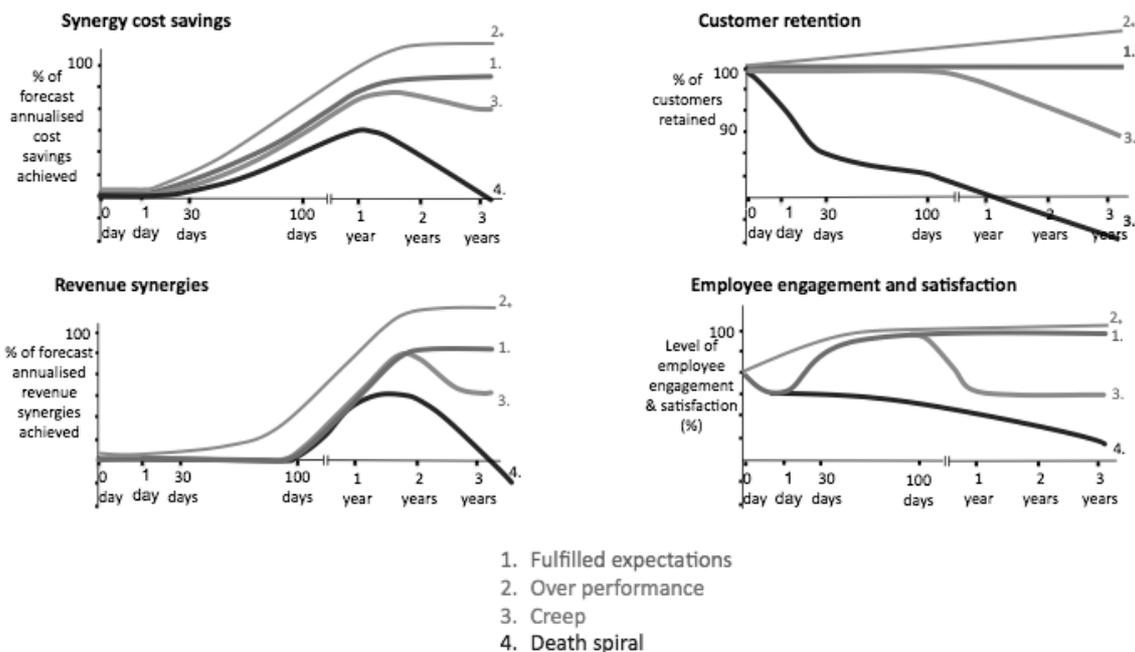


Figure 1. Summary of Integration Scenarios

In the Creep scenario initial efforts to achieve revenue synergies are rewarded by increased revenue, but the focus fades over time and revenue synergies decline. As an example, the initial focus on cross selling opportunities often achieves returns, but the effort devoted to cross selling declines over time and sales people return to their old behaviours. A similar pattern is observed in customer retention. One integration expert explains creep in the customer retention outcome measure: "The pattern actually is normally about very high retention in the early days, because there's a big focus on it. But I actually think that it then tails off over time when expectations actually don't get met. So, the goal on customer retention is pretty simple, it's 'We're going to retain 100% of all our customers throughout'. And there's a really big focus and a big investment program around doing that upfront. That doesn't get sustained for 2 or 3 years, that really gets sustained when you're doing the integration, and the real question then is, 'Are you meeting all your customers' promises 12, 18, 24 months down the track?' And certainly what I've read on the Big Bank case, I know they stopped tracking that after about 12 months. So they put it in place, tracked it, tried to make sure they were as close to this 100% as they could be, but I'm pretty sure they've, and I've read recently in the press, they've started losing customers."

Employee engagement and satisfaction suffers a similar experience under the Creep scenario. With the progress of time the focus on communication to employees wavers,

promises are not fulfilled, synergy fatigue sets in, and employee satisfaction & engagement decline.

The fourth scenario, the Death spiral pattern, occurs when the pressures of the integration are not well managed and they “break the business”. There may be numerous initial causes for pressure to occur, but poor management decisions and processes create the downward spiral. Initial pressures may be the result of poor planning or unachievable synergy forecasts. For example, one expert explained that if business unit management are not involved in the assessment of the synergy numbers, the numbers will be wrong. The result is, “then the first thing they see is that it’s rubbish, and then it has completely detracted from the process”. Poor assessment of synergies may drive low management commitment, especially when management is, “given a KPI that is something he doesn’t believe in.” The result is uncertainty about the operating model, the way in which the business will operate in the future.

Uncertainty leads to declining employee engagement, which in turn undermines productivity and customer service. These feedback effects are exacerbated by rising unintended employee departures. Also, these feedbacks add costs and delays to the integration process that have flow on effects to other employees and customers. Once activated, these feedbacks can cause a downward spiral of the M&A integration. One expert consultant explains the death spiral effect: “it is a cancer...it is debilitating...it creates a negative vibe that impacts value and performance...it is a distraction to everything and people do not want to be there and it is usually your star performers that leave...it is like a death spiral effect and it is hard to get momentum around the business to drive the integration program...and people talk to their customers about it.” The Death spiral pattern is labelled number 4 in Figure 1 and results in declining outcome measures over time.

The analysis of the four patterns of behavior, together with rich descriptions from the interviews, provided insights into the dynamics of the integration process.

5 Analysis and Discussion

Causal loop diagramming was applied to analyse the interview data and to connect the patterns of behaviour to the causal relationships responsible for driving those patterns that arose during the interviews.

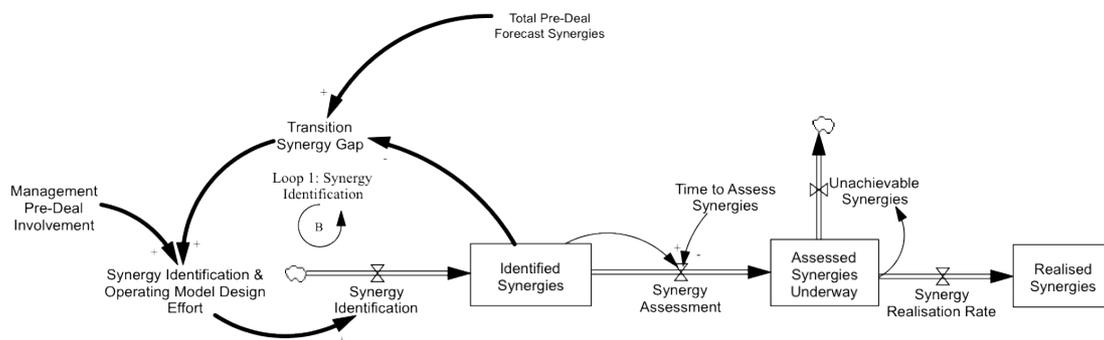


Figure 2. Synergy identification loop

The first causal loop is displayed as Loop 1 in Figure 2 and shows that the search for synergies drives the integration process. The achievement of pre-deal forecast synergies motivates the entire integration process. This pre-deal forecast number is derived from the strategic deal rationale. The search for synergies continues until the “Total Pre-Deal Forecast Synergies” amount is equal to the “Identified Synergies”

amount. Synergy identification effort is facilitated by pre-deal involvement of affected business managers. The importance of management involvement in the pre-deal phase is illustrated in an M&A integration expert’s comment on the reasons for the failure of a M&A transaction, “I think the key failure was the people who did the deal weren’t the people who did the execution”.

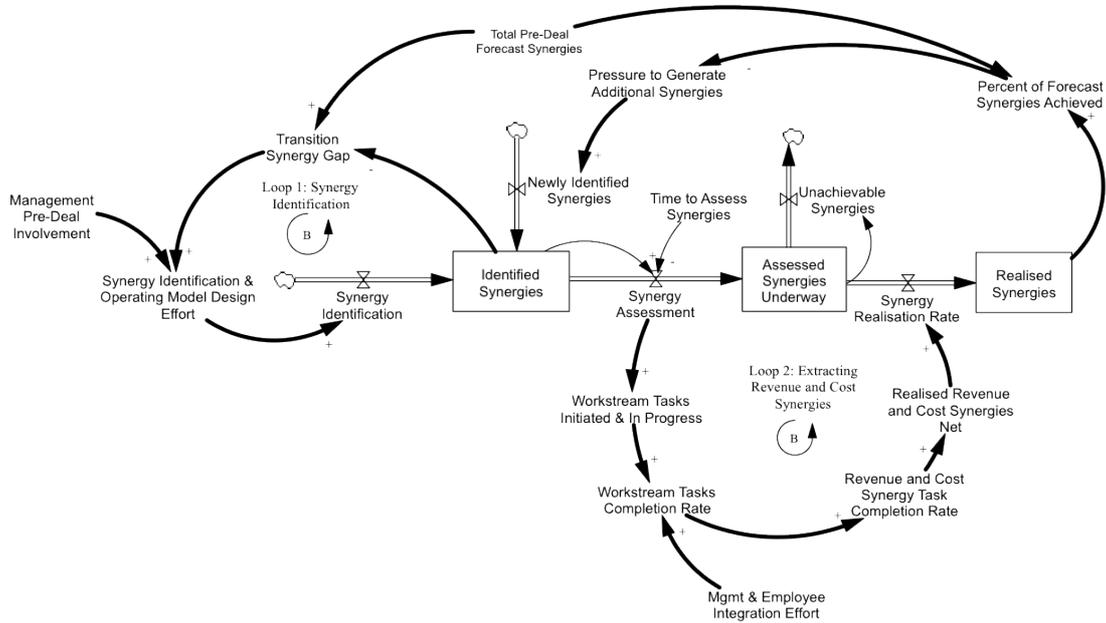


Figure 3. Synergy realisation loop

The second loop reflects the Fulfilled expectations pattern of behaviour and is included as Loop 2 in Figure 3. This loop moves our focus from high level synergy assessment to the next level of detail. To realise synergies, tasks have to be allocated to the relevant ‘work streams’ or divisions within a business. It is only when the work streams finish the tasks, that synergies are realised. These relationships are captured in the bottom half of Loop 2. Some tasks may not be achievable as shown with an arrow out of the stock of “Assessed Synergies Underway” in Figure 3. The relationships shown in the top half of the diagram complete the loop. Specifically, realised synergies are continuously compared with the pre-deal forecast synergies. Until 100% of the forecast pre-deal synergies are achieved, management exerts pressure to generate additional synergies to close the gap.

Figure 4 shows the next set of outer loops, reflecting the Creep pattern of behaviour. When realised synergies are less than pre-deal forecast synergies, there is a search to identify new synergies. Synergy fatigue may occur from the stresses of continually having to find and realise synergies, this stress may initially be driven by unachievable pre-deal forecast synergies. The effect of synergy fatigue is to decrease employee engagement and work stream management commitment and focus. The change in management commitment and focus negatively impacts management and employee integration effort and the work stream task completion rate. Declining employee engagement erodes realised synergies (ie., creep), and potentially decreases profitability of Business As Usual (BAU) – the on-going normal operations of the acquired and acquiring businesses. As a result, lower performance creates pressure to generate new, additional synergies. This dynamic is captured in Loops 3, 4 and 5 in Figure 4. Only

management policies and procedures can control the escalating stress on the system, and limit potentially poor outcomes.

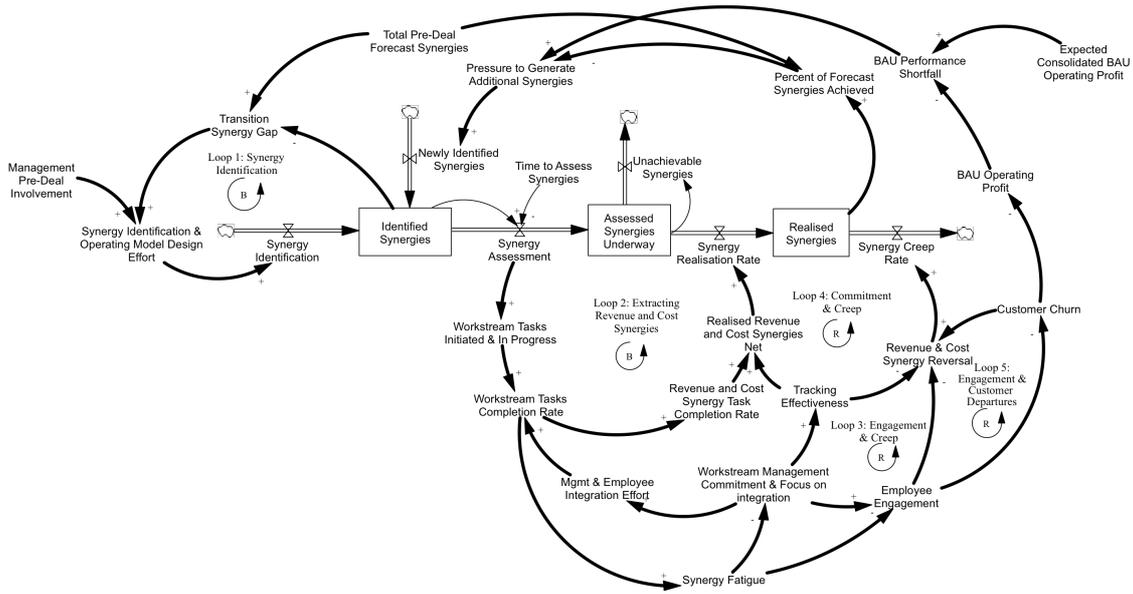


Figure 4. Effect of synergy fatigue

Without appropriate management processes and policies, the Death spiral pattern may occur. The initial effects of synergy fatigue on employee engagement may be exacerbated by increases in unplanned turnover of employees. In turn, rising employee turnover increases uncertainty. This dynamic is displayed in Loops 6 and 7 in Figure 5, the two loops reinforce initial changes in employee engagement. Declining employee engagement may also decrease productivity, which negatively impacts employee performance and further deteriorates profitability of business as usual, this effect is captured in Loop 8 in Figure 5. These feedback effects create more pressure to generate new, additional synergies and further stress the business.

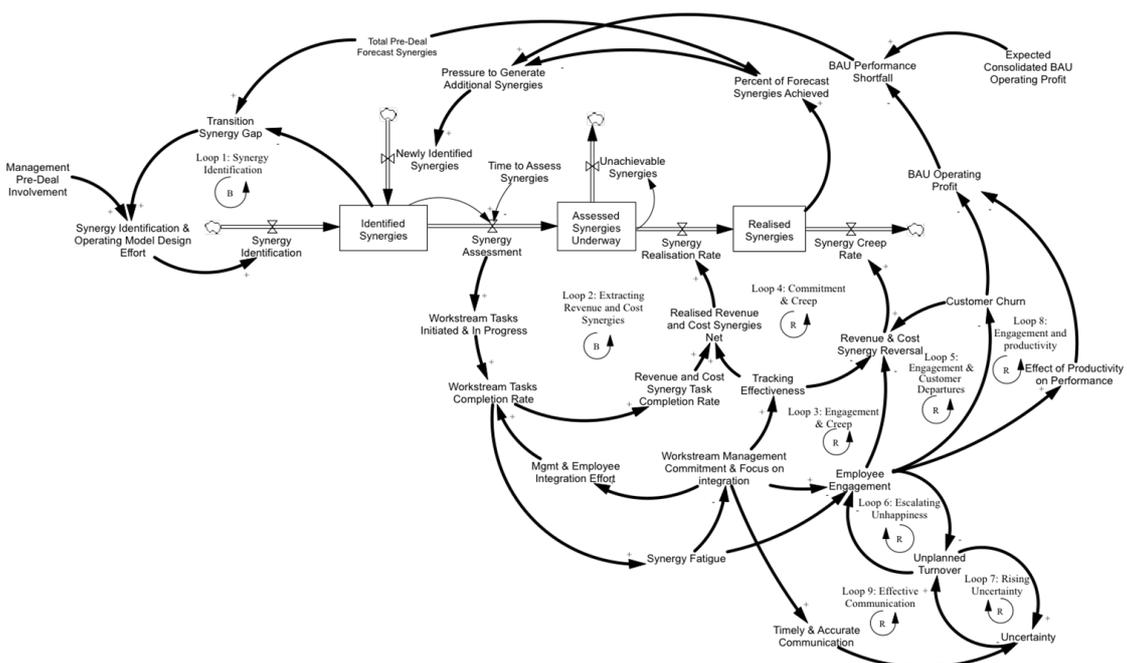


Figure 5. Effect of unplanned turnover, uncertainty, & communication

Timely and accurate communication decreases uncertainty for employees, but poor or inaccurate communication increases uncertainty. An M&A integration consultant explained that, “successful projects have high levels of personal engagement, there needs to be consistency and compliance of script... companies are extremely fragile during periods of uncertainty”. This effect is captured in Loop 9 in Figure 5.

The causal diagram shown in Figure 5 captures a dynamic system of interactions between factors, processes and policies in M&A integration. This set of reinforcing feedback loops exacerbates the effect of even a small stress on the system if not appropriately managed. As a result, this feedback structure can fundamentally change the value created from an M&A transaction from positive to negative returns.

As the causal diagram emerged, the links were reviewed with reference to prior research to determine if prior research supported the connections. The effect of “Unplanned turnover” on “BAU Operating Profit” is supported by the work of Cannella and Hambrick (1993) who found a negative relationship between levels of management turnover and post-acquisition performance. Bergh’s (2001) finding of a positive relationship between management agreement on acquisition goals and expectations of success agrees with the positive causal link between “Management Pre-Deal Involvement” and “Synergy Identification & Operating Model Design Effort”. The link captured between “Timely & Accurate Communication” and “Uncertainty” is supported by a host of studies on employee distress including Bastein (1987), Schweiger, Ivancevich & Power (1987), Schweiger & DeNisi (1991), and Ellis, Rues & Lamont (2009). The higher rates of turnover in firms undergoing M&A integration than firms that are not in an integration process agrees with Krug’s (2003) finding of very high levels of executive departure for two years post-merger. Overall, there are no connections in the causal diagram that are contrary to prior research but there are factors and dynamics that have not been previously addressed.

6 Conclusion

This inductive study involving experts provides new insights into the dynamics of the M&A integration process and improves our understanding about how poor outcomes emerge from M&A transactions. The research involved in-depth fieldwork investigating the causal relationships at work in the M&A integration process. Four common patterns of behaviour or integration scenarios were identified. The causal relationships emerged through analysing the interview data using causal loop diagramming. The integrated causal diagram presents emergent theory through identification of new factors, relationships, feedback and dynamics within the M&A process. This work is an example of how causal diagramming can be successfully applied to capture dynamics and new insights in an implementation process.

This research is unique in that it focuses on the dynamics of the M&A integration process. New theoretical insights provided by this research include the identification of “Pressure to Generate Additional Synergies”, “Synergy Fatigue”, and the feedback effects on realised synergies and BAU operating profit. Our findings show that managerial pressure to generate new synergies – after discovering that some of the initially identified synergies are unachievable – can result in ‘synergy fatigue’. This undermines employee commitment, focus and engagement, which activate a host of reinforcing feedback processes reducing revenue and cost synergy efforts, reversing previously captured synergy benefits, and negatively impacting performance of the ongoing business. The limitation of this research is that the theoretical relationships

have not been confirmed with direct reference to case study data, our reference has been indirect through the stories and experiences of M&A integration experts. This limitation highlights the opportunity for case study research to further investigate and confirm the connections captured in the causal loop diagram.

We hope this research alerts management to the potential loss in value that may result from pursuing pre-deal synergy forecasts to the detriment of other factors in the M&A integration process.

7 References

- Bastien, D.T. 1987. Common Patterns of Behavior and Communication in Corporate Mergers and Acquisitions. *Human Resource Management*. **26**(1): 17-33.
- Bergh, D.D. 2001. Executive retention and acquisition outcomes: A test of opposing views on the influence of organisational tenure. *Journal of Management*. **27**: 603-622.
- Bower, J.L. 2001. Not All M&A's Are Alike - and That Matters. *Harvard Business Review*. **March**: 93-101.
- Cannella, A.A.J., D.C. Hambrick. 1993. Effects of executive departures on the performance of acquiring firms. *Strategic Management Journal*. **14**: 137-152.
- Cording, M., P. Christmann, L.J.I. Bourgeois. 2002. *A Focus on Resources in M&A Success: A Literature Review and Research Agenda to Resolve Two Paradoxes*. Denver, CO.
- Cording, M., P. Christmann, D.R. King. 2008. Reducing Causal Ambiguity in Acquisition Integration: Intermediate Goals as Mediators of Integration Decisions and Acquisition Performance. *Academy of Management Journal*. **51**(4): 744-767.
- Eisenhardt, K.M. 1989. Building Theories from Case Study Research. *The Academy of Management Review*. **14**(4): 532-550.
- Ellis, K.M., T.H. Rues, B.T. Lamont. 2009. The Effects of Procedural and Informational Justice in the Integration of Related Acquisitions. *Strategic Management Journal*. **30**: 137-161.
- Finkelstein, S., J. Halebian. 2002. Understanding Acquisition Performance: The Role of Transfer Effects. *Organization Science*. **13**(1): 36-47.
- Glaser, G., A. Strauss. 1967. *The discovery of grounded theory: Strategies of qualitative research*. Wiendenfeld and Nicholson, London.
- Haspeslagh, P.C., D.B. Jemison. 1991. *Managing acquisitions: creating value through corporate renewal* The Free Press, New York.
- Homburg, C., M. Bucorius. 2006. Is speed of integration really a success factor of mergers and acquisition? An analysis of the role of internal and external relatedness. *Strategic Management Journal*. **27**: 347-367.
- Jemison, D.B., S.B. Sitkin. 1986. Corporate Acquisitions: A Process Perspective. *Academy of Management Review*. **11**(1): 145-163.
- King, D.R., D.R. Dalton, C.M. Daily, J.G. Covin. 2004. Meta-analyses of post-acquisition performance: Indications of unidentified moderators. *Strategic Management Journal*. **25**: 187-200.
- KPMG. 2006. *The Morning After: Driving for post deal success*. KPMG International, UK.
- Krug, J.A. 2003. Why Do They Keep Leaving? *Harvard Business Review*. **February**: 14-15.
- Schweiger, D.M., A.S. DeNisi. 1991. Communication with employees following a merger: a longitudinal field experiment. *Academy of Management Journal*. **34**: 110-135.
- Schweiger, D.M., P.K. Goulet. 2000. Integrating Mergers and Acquisitions: An International Research Review. *Advances in Mergers and Acquisitions*. **1**: 61-91.
- Schweiger, D.M., J.M. Ivancevich, F.R. Power. 1987. Executive Actions For Managing Human Resources Before And After Acquisition. *Academy of Management Executive*. **1**(2): 127-138.
- Sterman, J.D. 2000. *Business dynamics: systems thinking and modelling for a complex world* McGraw-Hill, New York.
- Strauss, A., J. Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage Publications, Inc., Newbury Park, California.
- Weber, R., C.F. Camerer. 2003. Cultural Conflict and Merger Failure: An Experimental Approach *Management Science*. **49**(4): 400-415

Optimal Design of Experiments to Determine Mechanical Properties of Soft Bodies

T.P. Babarenda Gamage¹, M.P. Nash^{1,2}, P.M.F. Nielsen^{1,2}
Auckland Bioengineering Institute¹, Department of Engineering Science²
The University of Auckland
New Zealand
psam012@aucklanduni.ac.nz

Abstract

There is a growing need for reliable mathematical models that simulate the biomechanical behaviour of soft tissues. Such models can be useful in many medical applications such as dynamical tracking of breast tumours for patient repositioning during radiation therapy. Reliable model predictions typically require mechanical properties of an individual's tissues to be identified. Determining these properties is difficult as it is unclear which mechanical tests should be performed to successfully characterise the mechanical response of the tissues. This can potentially lead to significant variability being introduced in estimating the mechanical properties. This study aimed to develop an optimal design framework to determine specific experiments to perform to improve parameter precision. The utility of the framework is demonstrated for identifying mechanical properties of a soft body using gravity loading.

Key words: Design of Experiments, Parameter estimation

1 Introduction

In a recent study, we investigated the use of gravity as a tool to identify the mechanical properties of a soft body (Babarenda Gamage et al. 2011). This approach was validated using controlled experiments on a two-layered silicone gel cantilever beam measuring 70mm by 30mm by 30mm. Eight variations on the beam orientation, ψ , were laser scanned. Non-linear least squares optimisation was used to determine θ , the unknown stiffness of each layer of the beam by minimizing the sum of squared errors between model predicted deformation and the experimental observations. This is represented as a mean squared error to remove dependence on

the number of surface points scanned,

$$\phi = \frac{\sum_{j=1}^M \sum_{i=1}^{N_j} \|Z_{ij}\|^2}{\sum_{j=1}^M N_j} \quad (1)$$

where $\|Z_{ij}\|$ is the Euclidean distance between the i^{th} laser scanned datapoint and its closest point on the surface of a model in the j^{th} orientation, M is the total number of gravity loaded orientations tested, and N_j is the number of laser scanned data points recorded for the j^{th} orientation.

The optimal θ determined were found to be dependent on ψ when only a single ψ was used for identification. Considering multiple ψ during identification resulted in consistent improvement in parameter precision for up to all 7 combinations of ψ tested. The aim of the current study was to ascertain whether specific ψ could be identified which maximise the precision to which θ could be determined.

2 Optimal design methodology

We developed an optimal design framework for determining ψ , for which ϕ , would be most sensitive to perturbations in the optimal stiffness parameters, $\hat{\theta}$. This would help improve the definition of the minima and therefore aid stiffness identification. As $\hat{\theta}$ is unknown, this was approximated from (Babarenda Gamage et al. 2011) ($2.12kPa$ for the first layer and $4.10kPa$ for the second layer) which was identified by combining 7 orientations as described in the previous section. $\hat{\theta}$ was used to construct synthetic experimental data from which the sensitivity of the model to perturbations to these parameters could be numerically calculated for any given ψ .

2.1 Quantifying model parameter identifiability

A local estimate of the sensitivity of ϕ to perturbations to $\hat{\theta}$ for a given ψ can be quantitatively determined by evaluating the Hessian matrix, \mathbf{H} , in the neighbourhood of $\hat{\theta}$. This region is termed the indifference region and takes the form of a hyperellipsoid when ϕ is plotted on the parameter space. Small indifference regions indicate low variance amongst all the parameters (Nathanson and Saidel 1985). The volume of this region is inversely proportional to $\det(\mathbf{H})$ (also known as the D-optimality criterion) and can be used to quantify the identifiability of $\hat{\theta}$ for a given ψ .

2.2 Improving model parameter identifiability

In order to improve identifiability of $\hat{\theta}$, we aimed to determine choices of ψ that minimised the indifference region at $\hat{\theta}$. This was achieved by maximising $\det(\mathbf{H}_{\hat{\theta}})$ using the following nonlinear optimization model.

$$\underset{\psi \in \mathbf{R}^{2n}}{\text{maximise}} \quad f(\det(\mathbf{H}_{\hat{\theta}}(\psi))) \quad (2)$$

where the parameter $\hat{\theta}_j$ represents the stiffness of the j^{th} layer of the two layer beam, and the decision variable ψ_{in} represents the 3D orientation of beam in spherical

coordinates, with $i = 1, 2$ indicating the azimuth and elevation angles respectively, for the n^{th} optimal orientation to identify. $\boldsymbol{\psi} = (0^\circ, 0^\circ)$ was taken to indicate a horizontal beam.

The nonlinear optimization was performed using the *fminunc* algorithm in Matlab 2012a (The MathWorks, Inc., USA: <http://www.mathworks.com>) which terminated once $\boldsymbol{\psi}$ was determined to the nearest 1° . No constraints were required on the orientations since the parametrisation of $\boldsymbol{\psi}$ in spherical coordinates was periodic with periods of $\psi_{1n} \in (0^\circ, 180^\circ)$ and $\psi_{2n} \in (0^\circ, 360^\circ)$ due to the symmetry of the layered cantilever arrangement.

3 Results and discussion

The optimal $\boldsymbol{\psi}$ identified are listed in table 1. Multiple initial $\boldsymbol{\psi}$ were used for optimization which lead to multiple local minima being found. The properties of these different local minima can be assessed from the determinant of the Hessian matrix with respect to the minima identified ($\det(\mathbf{H}_{\hat{\boldsymbol{\psi}}})$) in a similar manner as described in Section 2.1. Such properties can be exploited to steer the experimental design towards more practical scenarios, by favouring experiments where $\det(\mathbf{H}_{\hat{\boldsymbol{\theta}}})$ remains relatively large despite inaccuracies expected (for example in positioning the beam) during the practical application of the design.

Table 1: Optimal orientations identified for improving parameter precision

n	Optimal $\boldsymbol{\psi}$ ($^\circ$)	$\det(\mathbf{H}_{\hat{\boldsymbol{\theta}}})$
1	(35, 13)	4.6
2	(29, 31), (0, 15)	6.6

As a comparison, the maximum $\det(\mathbf{H}_{\hat{\boldsymbol{\theta}}})$ observed using synthetic data and for all combinations of up to 7 orientations considered in (Babarenda Gamage et al. 2011) was 2.3. This highlights a potentially significant improvement to parameter precision with minimal number of experiments needing to be performed.

4 Conclusion and future work

This study described a systematic approach to designing experiments that can be applied to determine the mechanical properties of soft tissues. Future work will look at performing experiments based on the results of this study to verify the approach and determine the extent to which the identified stiffness parameters vary within a specified measure of experimental error.

Acknowledgments

The authors are grateful for financial support from MSI NERF (UOAX0707), and thank Matthias Ehrgott for valuable discussions. P. M. F. Nielsen is supported by a RSNZ James Cook Fellowship.

References

- Babarenda Gamage, T. P., Rajagopal, V. , Ehrgott, M., Nash, M. P., and Nielsen P. M. F. 2011. “Identification of mechanical properties of heterogeneous soft bodies using gravity loading.” *International Journal for Numerical Methods in Biomedical Engineering* 27 (3): 391–407.
- Nathanson, M. H., and G. M. Saidel. 1985. “Multiple-objective criteria for optimal experimental design: application to ferrokinetics.” *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 248 (3): R378–386.

Demand Driven Throughput Assessment for Hunter Valley Coal Chain

Natashia Boland, Martin Savelsbergh, and Mohsen Reisi
School of Mathematical & Physical Sciences
University of Newcastle
Australia
mohsen.reisi@uon.edu.au

Extended Abstract

Hunter Valley Coal Chain (HVCC) is the largest coal export operation in the world. This operation includes transporting coal from over 40 mines located in the Hunter Valley area to the port of Newcastle in New South Wales, Australia. The port of Newcastle serves approximately 1200 vessels to export more than 100 million tonnes of coal per year. The system that supports this operation is a collection of mines, rail tracks, trains, and port terminals.

The coal is initially mined and stored at load point facility used by several mines and then transported, almost exclusively by train, to one of the terminals at the port of Newcastle. Afterwards, the coal is offloaded and stockpiled at the terminal before being transported to the vessels. This process is a collaboration between the rail operations, stockyards, and port operations, which are coordinated by the Hunter Valley Coal Chain Coordinator (HVCCC). The mission of the HVCCC is to plan a cooperative daily operation and long term infrastructure provision.

One of the most important and far-reaching decision problems faced by HVCCC is long-term capacity planning. The demand for coal continues to grow, and export demand through Newcastle is expected to increase in the near future. The optimal usage of the current facilities and machines will not necessarily lead to full accommodation of the future demand. This means that capital investment is required to upgrade the infrastructure and expand the capacities of the system. As upgrading infrastructure and expanding capacity are extremely expensive, system analysis plays a crucial role in ensuring the money is invested in the right section and at the right time.

HVCCC is trying to analyse and assess the throughput of the system to find out the bottleneck and investigate when and where they should invest to extend the current capacities. To achieve this they use a simulation model considering mines, load points, rail operation and port operation. The model considers a lot of details of the operation but is computationally intensive, therefore a few scenarios can only be analysed.

This research introduces an integer programming-based methodology instead of the simulation model to assess the throughput in a short amount of time. Two main aspects are considered in this work: solving the integer programming model efficiently and analysing the throughput.

First, in this work a pure integer programming model is introduced, that considers reasonable details of the problem in a daily level. This model finds an optimal way to provide

coal for a set of vessels without violating capacities of load points, junctions, wagons, dump stations, stacker, reclaimers, stockyards, and berths. Each vessel has a due date, and if its loading is finished after that, delay happens. The objective function tries to minimise the total delays of the vessels. As solving the model takes time, some of its variable that are similar in term of resource consumption are aggregated. Aggregating decision variables not only reduces size of model, but also breaks symmetry, which helps to reach more quickly the optimal solution. The aggregated problem can be solved faster than the original model, but it is still computationally intractable in some cases. Therefore, to tackle the issue, several strategies such as tightening constraints and per-processing are introduced in this work to reduce size of problem and provide stronger formulation.

The purpose of introducing and developing the integer programming model is to provide a tool that can assess the throughput of the system in an acceptable time. After working on the mathematical formulation and solution method, the next step is system analyses. First, different scenario needs to be created and tested to find out the reaction of system. The scenarios are varied by changing the tonnage of demands, increasing number of vessels (customers), and changing the arrival time of the vessels. Afterward, the solutions are analysed and compared using different metrics like number of delays, queue length, and stockyard utilization. The results are illustrated in some tables, and charts.

Stakeholder Engagement in Capital Budgeting at Counties Manukau District Health Board

Ali Broadbent
Catalyze Ltd.
New Zealand
alibroadbent@catalyze.co.nz

Abstract

How did Counties-Manukau District Health Board shift their capital budgeting process from one of competition for scarce resources between diverse stakeholders to a process with a high level of stakeholder alignment and ownership, while pressure on the capital budget continued to increase?

Catalyze worked with Counties-Manukau DHB to implement a process that prioritises capital investment using Decision Conferencing and Multi-Criteria Decision Analysis (MCDA). MCDA helps stakeholders to consider the cost of proposed investments along with both tangible and intangible benefits.

In a portfolio prioritisation Decision Conference, stakeholders score each proposed investment against criteria that articulate the organisations high level goals, and then weight the relative preference of the criteria. Scoring and weighting are interactive processes that result in shared understanding amongst stakeholders. Catalyze use MCDA modelling software to capture the scores and weights in real time, and an order of priority in terms of value for money for the capital requests is generated and displayed during the Decision Conference.

The new process at Counties-Manukau DHB has been through two iterations and has resulted in senior stakeholders enthusiastically participating in their capital budgeting process. Throughout the year there is an increased sense of confidence that they are spending their limited capital on the right things.

Scheduling Families of Jobs on Multiple Identical Machines to Minimize Total Tardiness

Simon Bull, Andrew Mason, Andrea Raith
Department of Engineering Science
University of Auckland
s.bull@auckland.ac.nz

Abstract

The scheduling of jobs to minimize tardiness is a common and difficult problem. Here we consider the scheduling of jobs in a particular factory to minimize total tardiness. Jobs have different processing times and due dates, and are pre-assigned to families according to the mould that each job requires. There are multiple identical machines but the number of moulds available for each job family is limited. Setups are not required between jobs of the same family on a machine but are required between jobs of different families to load and prepare the incoming mould, and such setups use the machine and incoming mould for the setup period. We consider a job-based IP model and a batch-based IP model that can each be solved to optimality for problems of the size required by the factory we are studying. Extending this, we look at the applicability of the models to random instances of similar problems and explore how the models can be used to find solutions to problem instances of much larger size.

1 Background and problem description

In a particular factory there are six identical machines. The factory receives orders for jobs, which each have a due date, a processing time and a required mould type. We consider the problem of scheduling the jobs on six identical machines, and at this factory the goal is to minimize the sum of tardinesses over all jobs. Jobs fall into families based on the mould that each job requires. There are sixteen mould families, and therefore sixteen job families. There are a limited number of identical moulds in each family, either one mould or two moulds. To process a job a machine must have the appropriate mould loaded, and must then process the job for exactly that job's processing time (i.e. there can be no idling or changing jobs part way through the processing of any job.) One mould and one machine are occupied at any time when a setup is underway or when a job is being processed. However, no mould or machine is required to unload a mould – once a job has been completed the mould is immediately available for loading onto a different machine. Processing

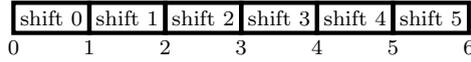


Figure 1: Time discretization, where shifts are periods and time index labels exist between shifts.

times are measured to the nearest shift, rounding up. If a job finishes part-way through a shift, the machine will idle until the next shift. All moulds take exactly one shift to to be set up on a machine.

Two previous students considered this problem. Helm (2011) describes a model that is based on job order but is not able to model the constraints on the number of moulds in each family. White (2012) gives a time-indexed and machine-indexed model that is able to solve the problem for very small instances. Unfortunately the model is very large and was not successfully solved for problems with more than ten jobs. Similar problems have come up in other work, including in Ibarra-Rojas et al. (2011) and Chen and Wu (2006). However, we have not identified anything that exactly solves the problem on identical machines with similar resource constraints (in our case moulds). The two features that distinguish our problem are the mould constraints and the family-based setups.

2 Definitions

In the factory we are concerned with it is natural to refer to time in terms eight-hour shifts, which we refer to as periods. Due dates, processing times and setup times are all measured in shifts. We consider shifts, and assign integer labels to the point in time between two adjacent periods. We assume all due dates may be interpreted as falling on one integer time label between two shifts. When referring to time intervals, we use the following conventions: if something begins at time t , it first occurs in the interval $[t, t + 1]$; if something finishes at time t , it last occurs in the interval $[t - 1, t]$; if something is due at time t , it can finish at time t and incur no tardiness. See Figure 1 for a graphical illustration of the discretization. The following notation will be used:

\mathcal{T}	Valid times, $\{0, 1, \dots, T_{\max}\}$
\mathcal{F}	Set of all families, $\{0, 1, \dots, n\}$
\mathcal{J}_f	Set of all jobs from family f
\mathcal{J}	The set of all jobs, $\mathcal{J}_0 \cup \mathcal{J}_1 \cup \dots \cup \mathcal{J}_n$
p_j	Processing time of job j
$c_{j,t}$	Tardiness of job j if started at time t , calculated using the due date d_j as $\max\{0, t + p_j - d_j\}$
m	The number of machines.
k_f	The number of moulds available for family f

All setups are taken to be 1 time period or shift.

We will not explicitly consider each machine individually. Instead we will consider the machines as a resource; at each time t there are m machines available; the sum of jobs being setup or processed at any time t must not exceed m .

3 Job-based model

We create a job-based time-indexed model. There are two classes of variables: binary variables $x_{j,t}$ that have value 1 if job j starts at time t and 0 otherwise, and integer $y_{f,t}$ variables that determine the number of machines reserved for a family f at time t .

The objective is to minimize the sum of job tardinesses:

$$\text{minimize } \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} c_{j,t} x_{j,t}$$

where the $c_{j,t}$ measures the tardiness of job j if started at time t . Given due date d_j and processing time p_j , $c_{j,t} = \max\{0, t + p_j - d_j\}$. However, by using a different definition for $c_{j,t}$ the problem can easily be extended to other objectives such as a weighted sum or tardinesses, a sum of tardy jobs or a sum of lateness.

To define the constraints in our model we introduce an auxiliary decision variable, $n_{f,t} \in \mathbb{N}$, that gives the number of machines running jobs from family f at time t . The variable $n_{f,t}$ is distinct from $y_{f,t}$ because $y_{f,t}$ is the number of reserved machines, i.e. those processing jobs or performing setups for family f , whereas $n_{f,t}$ measures the number of machines processing jobs but not performing setups. $n_{f,t}$ is only for convenience in defining constraints, and is defined as follows:

$$n_{f,t} = \sum_{j \in \mathcal{J}_f} \sum_{\tau=t-p_j}^t x_{j,\tau}$$

There are four classes of constraints; those that ensure each job is done exactly once (1), those that ensure machines are reserved for each family when jobs are being processed (2), those that ensure machines are reserved *before* jobs are started (3), and those that ensure the number of machines processing jobs or being setup for jobs never exceeds the number of physical machines (4). Constraint (5) ensures that no more machines are reserved for a family than there are moulds for that family. The constraints are:

$$\sum_{t \in \mathcal{T}} x_{j,t} = 1 \quad \forall j \in \mathcal{J} \quad (1)$$

$$n_{f,t} \leq y_{f,t} \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T}, t > 0 \quad (2)$$

$$n_{f,t} \leq y_{f,t-1} \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T}, t > 0 \quad (3)$$

$$\sum_{f \in \mathcal{F}} y_{f,t} \leq m \quad \forall t \in \mathcal{T} \quad (4)$$

$$y_{f,t} \leq k_f \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T} \quad (5)$$

$$x_{j,t} \in \{0, 1\} \quad \forall j \in \mathcal{J}, \forall t \in \mathcal{T}$$

$$y_{f,t} \in \mathbb{N} \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T}$$

The setups are dealt with in an unusual way that greatly reduces the complexity of this model compared with other models. Constraint (3) requires that if $n_{f,t}$ jobs from a family f are being processed at t then that many machines must be reserved for f in the previous time period, $t - 1$. Either the same number of jobs were begin processed in that previous time period, and so no setups are required, or fewer jobs were being processed and so a setup is required. This can be extended to longer

setup times or different setups for each family, but not to setup times that depend on both the mould being loaded and the mould being unloaded.

There is no cost associated with the variables $y_{f,t}$ and $y_{f,t}$ is not present in any equality constraints, and so the same solution has many equivalent representations with different $y_{f,t}$ values, and one has minimum $y_{f,t}$. Given the minimum $y_{f,t}$ solution we can readily determine when setups are performed: if constraint (3) is binding then no setup is performed at time t , whereas if constraint (3) is not binding then a setup is being performed. The number of setups performed at time t is $y_{f,t} - n_{f,t}$.

4 Full batch-based model

An alternative model considers batches of jobs rather than individual jobs. A batch is defined as a set of jobs $\mathcal{b} \subseteq \mathcal{J}_f$ from a single family. All jobs in \mathcal{b} are from the same family and so they may be processed sequentially on a single machine without idleness or mould swapping. In this model (\mathcal{b}, t) tuples are selected that represent setting up a single machine at time t and processing all jobs in \mathcal{b} from time $t + 1$ to time $t + 1 + \sum_{j \in \mathcal{b}} p_j$. The cost of selecting any such time and set of jobs is the sum of the tardinesses of each jobs in the set, *given that the jobs are processed in an order that minimizes their sum of tardiness*. For a given set of jobs the order that minimizes tardiness depends on the time at which the batch will be started.

This model is similar to the model in Section 3, with the exception that the decision variables are now related to batches and times rather than jobs and times. The following definitions are introduced:

\mathcal{B}_f	Set of all subsets of jobs (i.e. batches) from family f
\mathcal{B}	Set of all batches, $\mathcal{B}_0 \cup \mathcal{B}_1 \cup \dots \cup \mathcal{B}_n$
\mathcal{B}_j	The set of all batches that contain job j
$p_{\mathcal{b}}$	One greater than the total length of jobs in batch \mathcal{b} , or the time taken to perform a setup and process all jobs in \mathcal{b} sequentially
$c_{\mathcal{b},t}$	Sum of tardinesses of the jobs in batch \mathcal{b} if started at time t , (in the order that gives least value, and accounting for the setup from t to $t + 1$)

The decision variables are now $x_{\mathcal{b},t} \in \{0, 1\}$, which determine whether or not batch \mathcal{b} is started at time t . A value of 1 indicates that batch \mathcal{b} starts at time t , and 0 indicates that it does not. The decision variable $y_{f,t}$ now has a clearer definition: $y_{f,t}$ is the number of machines reserved to set up or process jobs in family f at time t , but as setups are now included with the decision variable $y_{f,t}$ has a simple explicit definition:

$$y_{f,t} = \sum_{\mathcal{b} \in \mathcal{B}_f} \sum_{\tau=t-p_{\mathcal{b}}}^t x_{\mathcal{b},\tau}$$

The constraints below are similar to those in the job-based model. Constraint (1) – that all jobs are completed exactly once – becomes (6). Constraints (2) and (3) are no longer required. Constraint (4) is exactly the same as constraint (7), and (5) is

exactly the same as constraint (8). The constraints are:

$$\sum_{b \in \mathcal{B}_j} \sum_{t \in \mathcal{T}} x_{b,t} = 1 \quad \forall j \in \mathcal{J} \quad (6)$$

$$\sum_{f \in \mathcal{F}} y_{f,t} \leq m \quad \forall t \in \mathcal{T} \quad (7)$$

$$y_{f,t} \leq k_f \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T} \quad (8)$$

$$x_{b,t} \in \{0, 1\} \quad \forall b \in \mathcal{B}, \forall t \in \mathcal{T}$$

4.1 Batch-based model column generation

Initial experiments with a full enumeration of batches suggest that when the integer variables are relaxed to allow non-integer values (as a linear program or LP), the optimal objective provides a much better lower bound on the true optimal objective than the bound obtained from the job-based model relaxed in a similar way. However, the full enumeration is very large as it includes every subset of jobs, with a minimum-tardiness ordering determined for every start time for every subset. Instead we generate variables as needed in a column generation framework.

We term the full batch-based model described at the beginning of Section 4 the master problem (MP), but create a restricted master problem (RMP) that only has a subset of the variables of the MP. Initially we add a trivial feasible solution where every job is a batch as the initial columns to the RMP. We create one generator or pricing problems (PP) for every family of jobs, and at each iteration one pricing problem is solved which returns new columns to be introduced into the RMP (Lübbecke and Desrosiers 2010). In our problem the PP return one or more (batch, time)-tuples at each iteration, and all are added to the RMP.

We formulate the generators as single machine scheduling problems that are very similar to the job-based model introduced earlier. Each generator considers a single family f and has binary decision variables $x_{j,t}$, whether job j begins at time t , and binary variable y_t that indicates whether this family should reserve a machine at time t or not.

As in the job-based model we use an auxiliary variable denoted $n_t \in \{0, 1\}$, that indicates whether or not the family is running a job at time t , which is defined as follows:

$$n_t = \sum_{j \in \mathcal{J}_f} \sum_{\tau=t-p_j}^t x_{j,\tau}$$

We introduce three new parameters, π_j , the dual variable of constraint (6), σ_t , the dual variable of constraint (7) and γ_t , the dual variable of constraint (8). These are interpreted respectively as a reward for processing job j , a cost for using a machine at time t and a cost for using a mould at time t .

The objective function of the generator is:

$$\text{Minimize } \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} (c_{j,t} - \pi_j) x_{j,t} - \sum_{t \in \mathcal{T}} (\sigma_t + \gamma_t) y_t$$

The constraints are very similar to those in previous models. Constraint (9) ensures that each job is started at most once (but does not require that each job is started).

Constraints (10) and (11) ensure that whenever a job is running a machine is reserved for that time period and the previous time period.

$$\sum_{t \in \mathcal{T}} x_{j,t} \leq 1 \quad \forall j \in \mathcal{J} \quad (9)$$

$$n_t \leq y_t \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T} \quad (10)$$

$$n_t \leq y_{t-1} \quad \forall f \in \mathcal{F}, \forall t \in \mathcal{T} \quad (11)$$

$$x_{j,t} \in \{0, 1\} \quad \forall f \in \mathcal{J}, \forall t \in \mathcal{T}$$

$$y_t \in \{0, 1\} \quad \forall t \in \mathcal{T}$$

A solution to this generator problem is a sequence of jobs that run at particular times, either with or without idle times between them. The generator always has a potential zero-cost solution, which is to select no jobs for processing. Any other solution can be interpreted as one or more batches. The jobs in the solution are ordered by start time, and we identify batches by considering adjacent jobs in this sequence. Two adjacent jobs, j_1 starting at time t_1 and j_2 starting at time t_2 , are in the same batch if $t_2 = t_1 + p_{j_1}$. Using that rule we iterate through the ordered list of jobs and start times building batches. At an optimal solution for the generator each batch has a negative (or zero) reduced cost for the master problem, and we return all batches to the master problem as potential entering columns.

5 Improving the Column Generation

5.1 Dual ordering

We make the following observation of optimal basis matrices: the dual variables corresponding to constraints (8) in the master problem, σ_t , tend to be ordered. We observe that the following is true, in most cases:

$$t_1 < t_2 \implies \sigma_{t_1} \leq \sigma_{t_2}$$

We tested this assumption with several problems, and while it is certainly not true that for all optimal basis matrices the inequality holds, it appears to be true that for many problems there is at least one optimal basis matrix for which the inequality holds for every pair of dual variables σ_{t_1} and σ_{t_2} . We added constraints to the dual problem of the master of the batch model that enforce the inequalities between subsequent dual variables. The constraints have the following form:

$$\sigma_t \leq \sigma_{t+1}, \forall t \in \mathcal{T}$$

These constraints have a profound effect on the number of iterations and time required to solve the problem to LP optimality, in most cases reducing the solve time by about factor of two.

In the master problem this introduces variables $s_t \in \mathbb{N}$, and changes machine constraint (7) to constraint (12):

$$\sum_{f \in \mathcal{F}} y_{f,t} - s_t + s_{t-1} \leq m \quad \forall t \in \mathcal{T} \quad (12)$$

Each variable s_t appears in the constraint related to machine utilization at time t and at time $t + 1$. A non-zero value for s_t moves machine availability from time t to

time $t + 1$; as an example, if all values of s_t are zero except at a particular t , then m machines are available at every time except for at t and $t + 1$. At t , $m - 1$ units of machine time are available, and at $t + 1$, $m + 1$ units are available. It may seem intuitive that such variables should never have non-zero value in an optimal solution but in some instances variables s_t do create opportunities for solutions that are not feasible for the original problem and have a better objective.

Consider a simple problem with three jobs from different families, each with processing time $p_j = 1$ and due date $d_j = 3$. If the earliest time interval begins at $t = 0$ then there is no feasible solution that results in zero tardiness – at least one job must be late. See Figure 2a for this example schedule. However, by shifting one unit of time from the interval beginning at $t = 0$ to the interval beginning at time $t = 1$, the schedule in Figure 2b is possible, which has zero tardiness.

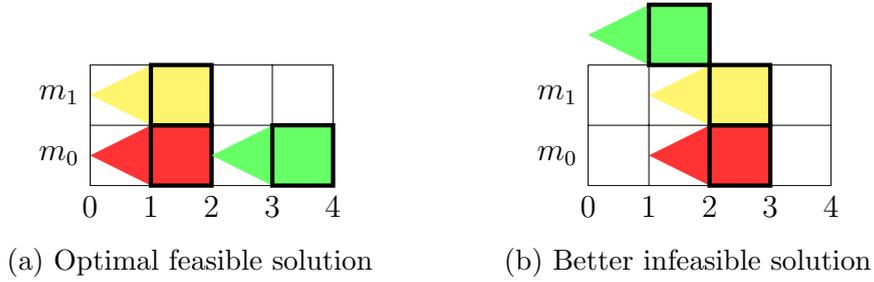


Figure 2: Two solutions to the problem described in Section 5.1. A square represents one unit of processing time, and a triangle represents a setup. The colour indicates job family. All jobs are identical in processing time and are due by time $t = 3$.

Despite the above described issue, we observed that in practice few problems have features that permit advantageously moving machine availability forward in time, and in cases where at optimality machine availability is being moved forward, we have been able to address the situation by adjusting the dual ordering constraints in the dual problem, introducing a constant c_t :

$$\sigma_t - \sigma_{t+1} < c, \forall t \in T$$

Equivalently, we introduce that constant c_t as a cost associated with the variables s_t in the primal problem. Our strategy has been to solve to LP optimality, determine whether or not any s_t variables have non-zero value and if they do assign a cost to them (at this stage by picking a large value) in the objective function and continue solving. The presence of the dual ordering constraints significantly improves the performance when solving to LP optimality, including the extra time required to remove non-zero s_t variables if they are present. In most cases it takes twice as long to find the LP optimal solution without the dual ordering constraints as it takes with them.

5.2 Déjà vu cuts

The column generators find a single machine schedule consisting of several batches, and at least one of those batches can be converted into a column with negative reduced cost that can be added to the master problem. At optimality there may be other columns in the schedule that have zero reduced cost, and these are likely to be batches that are in the master problem basis and have therefore already been

discovered by the generator. The constraint (13) forbids a batch of the jobs \mathcal{b} starting at time s and ending at time f from being feasible *in any job order*. To avoid violating the constraint, either a job must be processed during the period starting at time s , or a job must be being processed in the period starting at time e , or at least one of the jobs in \mathcal{b} must not begin between the times $s + 1$ and e .

$$\sum_{j \in \mathcal{b}} \left(1 - \sum_{t=s+1}^{e-p_j} x_{j,t}\right) + \sum_{j \in \mathcal{J}_f} \sum_{t=s-p_j+1}^s x_{j,t} + \sum_{j \in \mathcal{J}_f} \sum_{t=e-p_j+1}^e x_{j,t} \geq 0 \quad (13)$$

Our experiments with these déjà vu cuts indicate that they do not have a great affect on the performance; fewer iterations are required to reach optimality, but the addition of many cuts slows down the solve process.

6 Computing environment

We made use of the NeSI Pan Cluster at the University of Auckland to solve problem instances. We used the Gurobi Optimizer and its Python interface.

The time indexed model spends most of its solving time exploring a branch-and-bound search tree, which can readily be parallelized across threads on a single machine. The Gurobi Optimizer is able to take advantage of many threads in a parallel environment.

The batch-based was not (yet) solved to IP optimality but rather to LP optimality. The best IP solution was found using just those batches present in the problem at LP optimality. When solving, most time is spent generating new batches.

We implemented two parallel methods for solving the batch-based model. The first one cycles through all generators, at each iteration solving the master problem, providing dual variables to one generator at a time and receiving new batches from that generator, and like the job-based model the Gurobi Optimizer can parallelize the generators with multiple threads. Unfortunately the generators tend to be easy to solve but the column generation overhead is significant, and this parallelization is not as effective as with the job-based model.

The alternative implementation is a multi-process method where a single master process solves the master LP to find dual variables, and single processes for each job family which request dual variables from the master process and return entering batches. The master process initially finds dual variables and informs all family processes of these, and then waits for family processes to provide new batches and request dual variables.

For consistency with comparison, timings in Section 8 are from the former batch-based parallelization. When solving on the NeSI Pan Cluster we solved problems using 12 cores of a single node, 4 GB of memory and a wall time limit of one hour.

7 Test problems

In (Chen and Powell 2003) methods are considered for solving a similar multi-machine problem with job families, but with a different objective and without mould constraints. We generate example problems in a similar way to the method that they use, with an extension to generate the number of moulds for each family. This generation method is also similar to that used to generate instances in the CuSPLIB library of single machine scheduling problems (Yunes 2009).

For all test problems, we specify the number of machines, m , the number of families $|\mathcal{F}|$, the number of jobs, $|\mathcal{J}|$, the latest due date of any job, d_{\max} , and the longest processing time of any job, p_{\max} . Jobs are generated with uniformly distributed integer due dates in the range $[1, d_{\max}]$ and processing times in the range $[1, 6]$, and assigned to a family. The number of moulds are similarly generated – as low integers no greater than the number of machines. We generated many such test problems with different parameters and selected several that highlight the differences between the two models, and the problems and the performance of the models in solving them are summarized in Table 1.

8 Results

For the job-based model, we observe the objective value of the linear relaxation of the problem, the IP objective at optimality, and the *CPU* time required to find the optimal integer solution. Note that we allow one hour of *wall* time, but allow up to 12 cores to be used, so the *CPU* time can be much greater than 3600 s. For the batch model, we observe the objective value at LP optimality, the best integer solution found from the batches present at LP optimality, and the *CPU* time required to find the LP and IP solutions. Our key observation is that the batch-based model has a much better lower bound than the job-based model.

The performance for the two models on random problems is summarized in Table 1.

Table 1: Summary of sample problem parameters, and the performance of the two models on the problems. An asterisk indicates that the wall time limit was hit, and results are given for the best solutions at that time.

Name	N	$ \mathcal{F} $	$ \mathcal{J} $	d_{\max}	Job-based			Batch-based		
					LP	IP	CPU time	LP	IP	CPU time
Problem 1	1	2	10	5	65.0	75	0.43	75	75	11.60
Problem 2	1	2	10	5	108.2	118	5.33	118	118	12.07
Problem 3	1	2	20	60	0.0	0	0.23	0	0	68.03
Problem 4	1	2	40	40	966.8	1016	84.04	1016	1033	15317.70
Problem 5	8	20	100	100	7.0	7	1.76	7	7	6.27
Problem 6	4	20	100	100	75.7	124	270.70	124	124	72.05
Problem 7	7	6	30	25	677.0	796	6008.08	796	796	613.90
Problem 8	5	30	150	20	4687.1	*5765	*43124.10	5313	5313	1139.19

Note that the IP solution given for the batch-based model is not necessarily optimal; it is the best integer solution found from the columns present at LP optimality. However, in all of these examples it is also the optimal solution.

Note also that in problem 8 the job-based model exceeded the time limit. At that time the best integer solution it had found was 5765, and the LP lower bound had been improved from 4687.1 at the root node to 5017 – still a gap of over 13%.

The batch-based model performs very poorly for Problem 4; as there are only two families each generator has to consider many jobs and at each iteration the generators take a long time to find their optimal schedule. On the other hand, Problems 6, 7 and 8 show cases where the batch-based model performs significantly better than the job-based model. Each problem has on average just five jobs in each family and so the generators are very easy to solve at each iteration.

Although not shown here, the effect of the dual ordering variables is fairly consistent in reducing the solve time for the batch-based model. In Problem 8, for example,

without the ordering 1771 CPU seconds were required to solve the problem, rather than the 1139 CPU seconds required to solve the problem with the constraints and then fix the solution to remove any infeasibility.

9 Conclusions

We explored two time-indexed models that can exactly solve the problem at the particular factory. A job-based model can be solved as a single integer program, and can be solved for problems with up to one hundred jobs, if there is moderate lateness. A batch-based model with many more variables can be solved to its linear relaxation which gives a much better lower bound on the minimum sum of tardinesses, and in many cases the linear relaxation has the same objective as the optimal integer solution. The batch-based model out-performs the job-based model for some classes of problem, generally when there are many families but few jobs in each family. However in smaller problems with very few families, such that the families contain many jobs, the job-based model performs much better than the batch-based model.

We added a dual-ordering constraint to the RMP in the column generation model that relaxes making finding optimal solutions easier, and once an optimal solution is found we are able to remove the relaxation and find a correct solution. The dual-ordering constraints result in a significant decrease in the time required to solve the problem.

References

- Chen, Jeng-Fung, and Tai-Hsi Wu. 2006. "Total tardiness minimization on unrelated parallel machine scheduling with auxiliary equipment constraints." *Omega* 34 (1): 81 – 89.
- Chen, Zhi-Long, and Warren B. Powell. 2003. "Exact algorithms for scheduling multiple families of jobs on parallel machines." *Naval Research Logistics (NRL)* 50 (7): 823–840.
- Helm, Denis. 2011. "Production Scheduling with set up times." University of Auckland.
- Ibarra-Rojas, Omar J., Roger Z. Ros-Mercado, Yasmin A. Rios-Solis, and Mario A. Saucedo-Espinosa. 2011. "A decomposition approach for the piecemoldmachine manufacturing problem." *International Journal of Production Economics* 134 (1): 255 – 261. Enterprise risk management in operations.
- Lübbecke, M.E., and J. Desrosiers. 2010. "Column Generation." *Wiley Encyclopedia of Operations Research and Management Science, John Wiley and Sons, Chichester, UK.*
- White, Amelia. 2012. "Scheduling the production of surge arrester at Siemens." University of Auckland.
- Yunes, T. 2009. CuSPLIB 1.0: A Library of Single-Machine Cumulative Scheduling Problems. <http://moya.bus.miami.edu/~tallys/cusplib/>.

Mathematical Programming and Metaheuristic Approaches Applied to Biological-Based Fluence Map Optimization in Radiotherapy

Guillermo Cabrera^{a,b}, Manuel Chica^c, Matthias Ehrgott^a, Andrew Mason^a

^aDepartment of Engineering Science, University of Auckland (New Zealand)

^bPontificia Universidad Católica de Valparaíso (Chile)

^c European Centre for Soft Computing (Spain)

{gcab623, m.ehrgott, a.mason}@auckland.ac.nz

manuel.chica@softcomputing.es

Abstract

Intensity modulated radiation therapy (IMRT) is one of the most effective techniques in cancer treatment. Its main goal is to eradicate all clonogenic cells from the tumour without compromise of surrounding normal tissues. One specific problem in IMRT is fluence map optimization (FMO). The main goal of FMO is to find the optimal set of beamlet intensities given some clinical criteria. Both physical and biological criteria have been developed to tackle the FMO problem. Although most physical models are mathematically tractable and can be solved to optimality they do not include some important clinical considerations as, for example, radio-biological tissue response. On the other hand, biological models although more meaningful usually do not have desirable mathematical features. In this paper we carry out a comparison between metaheuristics and exact methods applied to an unconstrained non-linear biological model based on the well-known generalized equivalent uniform dose. We compare the results in terms of objective function value, time and the obtained dose volume histogram. We apply our algorithms to a non-clinical prostate case.

Key words: radiation therapy, fluence map optimization, generalized equivalent uniform dose, mathematical optimization, metaheuristics

1 Introduction

The main goal in radiation therapy is to eradicate all clonogenic cells from the tumour without compromise of surrounding normal tissues. However, because of the physics of radiation delivery, there is a trade-off between tumour control and normal tissue damage. The most common form of radiation treatment is intensity modulated radiation therapy (IMRT). IMRT allows the radiotherapist to modulate radiation across a beam, which is particularly important for volumes with non-convex shapes

in difficult anatomical situations (Ehrgott et al. 2008). Radiation modulation is possible thanks to a specific physical device called a multi-leaf collimator (MLC) which is able to block the radiation from specific beams. Therefore, IMRT can generate convenient dose distributions to deliver more radiation to the target while sparing surrounding organs at risk (OARs).

IMRT treatment planning is usually divided into three sequential sub-problems: beam angle optimization (BAO), fluence map optimization (FMO) and sequencing of the MLC. In BAO the main goal is to define an optimum number of delivery angles and their orientations (Ehrgott and Johnston 2003). FMO aims to find an optimal set of beamlet intensities given some clinical criteria. Usually hundreds of beamlets are involved in the FMO problem. Finally the sequencing problem aims to find the optimal delivery sequence of the previously defined fluence intensities seeking mainly to reduce the time of patient exposure to radiation.

In this paper we solve the FMO problem. FMO has been tackled from both physical and biological points of view. The former approach, also known as dose-volume approach, links the delivered dose to both tumour and normal tissues with tumour control and complications for the OARs, respectively. Dose-volume models usually maximize the dose delivered to the target and minimize the dose to the OARs subject to both bound constraints and dose-volume constraints (DVC). Most of the physical models are linear, mixed-integer, or quadratic models (Ehrgott et al. 2008). This allows scholars to find clinically acceptable treatment plans using well known exact techniques such as linear and quadratic programming. This approach, although it is very well-known, presents several weaknesses. From a mathematical point of view one can argue that some of its parameters (e.g. weights in quadratic models) have neither clinical nor physical meaning (Ehrgott et al. 2008). Moreover, measures such as mean delivered dose do not take into account the effect of hot/cold spots that can have huge consequences on the tumour control and radiation-related complications after the treatment (Thomas et al. 2005). From a clinical point of view dose-volume models do not consider some important aspects of radio-biological response e.g. cell fraction survival, oxygenation, repopulation, and radio sensitivity.

Biological models, also known as dose-response, relate the delivered dose to the biological response of the irradiated structures. Their main goal is to maximize the tumour control probability (TCP) while minimize the normal tissues complication probability (NTCP) of OARs. Some authors have pointed out the advantages of biological models over physical ones (Thomas et al. 2005; Wu et al. 2002).

In this paper we solve an unconstrained FMO model based on the generalized equivalent uniform dose (gEUD) (Wu et al. 2002). Although the gEUD function is convex (Choi and Deasy 2002) we cannot state the same for our unconstrained model (Olafsson, Jeraj, and Wright 2005). Therefore, we propose the use of metaheuristics as an alternative to exact methods. Concretely, two evolutionary algorithms are implemented, the CHC (Eshelman 1991) and differential evolution (DE) (Storn and Price 1997) algorithms. In addition, the conjugate-based method L-BFGS-B (Zhu et al. 1997) is deployed and compared with the latter algorithms in terms of objective function value, time and the obtained dose-volume histogram (DVH).

This paper is organized as follows. In Section 2 the main concepts related to both IMRT and FMO problems are introduced. In Section 3, we present the biological model used for our study. In Section 4, the optimization methods are explained. Finally, we discuss the results in Section 5 and we highlight some conclusions and

future work in Section 6.

2 Preliminaries

In IMRT region R (target or OAR) is divided in $|R|$ small volumes called *voxels*. The source of radiation is a set of fixed beam angles K which are divided in J beamlets or *bixels*. Information about the effect produced by one unit of intensity from *bixel* j on *voxel* i in region R is defined by the dose deposition matrix A^R . In this paper we assume that A^R is given. Below we present the formula to calculate the dose deposited at each *voxel* i for some vector w of beamlet intensities in Equation 1:

$$D_i^R = \sum_{j=1}^J A_{ij}^R w_j \quad \forall i = 1, 2, \dots, |R| \quad (1)$$

where D_i^R is the total dose deposited at *voxel* i in region R by the intensity vector w . Therefore, to solve the FMO problem we need to find a set of beamlet intensities w_j that produce a dose vector D^R that meets given clinical criteria. For a more comprehensive explanation of IMRT concepts see Ehr Gott et al. (2008).

Some exact methods have been developed to solve the FMO problem. Linear programming, mixed integer programming, quadratic programming, non-linear programming (NLP) and multi-objective optimization have been proposed before (Ehr Gott et al. 2008). Particularly most of the strategies to solve biological FMO models are based on conjugate gradient methods. Those approaches, although fast, do not ensure optimality (Aleman et al. 2010).

Metaheuristics have been mainly used to solve BAO (Bertsimas et al. 2012). There are also many multi-objective metaheuristic proposals which integrate the BAO and FMO problem but considering dosimetric deviation functions as the optimization objectives (Schreibmann et al. 2004). Regarding the use of metaheuristics to solve biological FMO models, only the work of Harmann and Bogner was found (Hartmann and Bogner 2008).

3 A biological FMO model based on gEUD

The biological NLP model used in this paper is based on the concept of gEUD. gEUD can be defined as the biologically equivalent dose that, if delivered uniformly, leads to the same response as the actual non uniform dose distribution (Niemierko 1997). One advantage of gEUD is the penalization of hot/cold spots in OAR/target regions without the need of several parameters as in other biological models.

Different models have been proposed using gEUD concept. Wu et al. (Wu et al. 2003) presented a model which combines a dose-response formulation with DVC. Some authors have suggested the superiority of gEUD-based models over dose-volume optimization in terms of OAR sparing with equal or even better target coverage (Thomas et al. 2005; Choi and Deasy 2002). For a complete analysis of these functions see Romeijn, Dempsey, and Li (2004) and Choi and Deasy (2002).

In this paper we focus on an unconstrained gEUD-based model proposed originally in Wu et al. (2002). Using this model, solutions that improve the sparing of critical structures while maintaining the target dose can be obtained (Wu et al. 2002; Olafsson, Jeraj, and Wright 2005). Furthermore gEUD based models provide

a large search space making it easier for the optimization system to balance competing requirements in search of a better solution (Choi and Deasy 2002; Wu et al. 2002). The mathematical expression of gEUD is shown in Equation 2:

$$gEUD(w; R, a) = \frac{1}{|R|} \left(\sum_{i=1}^{|R|} D_i^a \right)^{\frac{1}{a}} \quad (2)$$

where $|R|$ is the number of voxels of the structure R , a is the structure-dependent parameter and D_i corresponds to the i -th element of intensity vector D^R . For target structures a should be negative, whereas for normal structures a should be greater than 1. As $abs(a)$ increases the function becomes more sensitive to cold/hot spots in targets/OARs, respectively. Therefore, for those normal tissues that allow certain levels of radiation without a functional compromise (also called parallel structures), parameter a should be set close to 1. For serial structures (those that must be irradiated as little as possible) values for parameter a should be set greater than 10 (Thomas et al. 2005; Wu et al. 2002). Then, gEUD can be seen as a link between physical and biological models because it behaves as biological ones do but it depends highly on the intensities and also on the structure dependent parameter, a .

The NLP unconstrained model used in this paper is based on the gEUD concept presented above. Equations 3,4, and 5 show this model:

$$\min_{w \geq 0} f(w) = -\ln L(w; T, a_T, \nu_T, eud_0^T) - \ln U(w; N, a_N, \nu_N, eud_0^N) \quad (3)$$

where:

$$L(w; T, a_T, \nu_T, eud_0^T) = \left(\left(1 + \frac{gEUD(w; T, a)}{eud_0^T} \right)^{\nu_T} \right)^{-1} \quad (4)$$

$$U(w; N, a_N, \nu_N, eud_0^N) = \left(\left(1 + \frac{eud_0^N}{gEUD(w; N, a)} \right)^{\nu_N} \right)^{-1} \quad (5)$$

Parameters eud_0^T and eud_0^N correspond to the prescribed EUD values for target and normal tissues respectively. Parameters T and N denote the target and normal tissue, respectively. Then $|T|$ and $|N|$ correspond to the number of voxels of structures T and N , respectively. Finally, $\nu > 0$ is a user-defined parameter that indicates the importance of each structure.

Contour plots of the objective function f are presented in Figure 1a. These contour plots are in the EUD space generated by prostate and rectum structures only. Based on contours in Figure 1a we can easily realize that the objective function is convex in EUD space. However, nothing can be said about its convexity in the intensity space (Choi and Deasy 2002). In Figure 1b we show the same function in the 3D space.

4 Optimization strategies

4.1 L-BFGS-B algorithm

L-BFGS-B (Zhu et al. 1997) is a limited-memory quasi Newton optimization algorithm for solving large non-linear unconstrained optimization problems. The L-BFGS-B algorithm allows us to include bounds on decision variables which meets

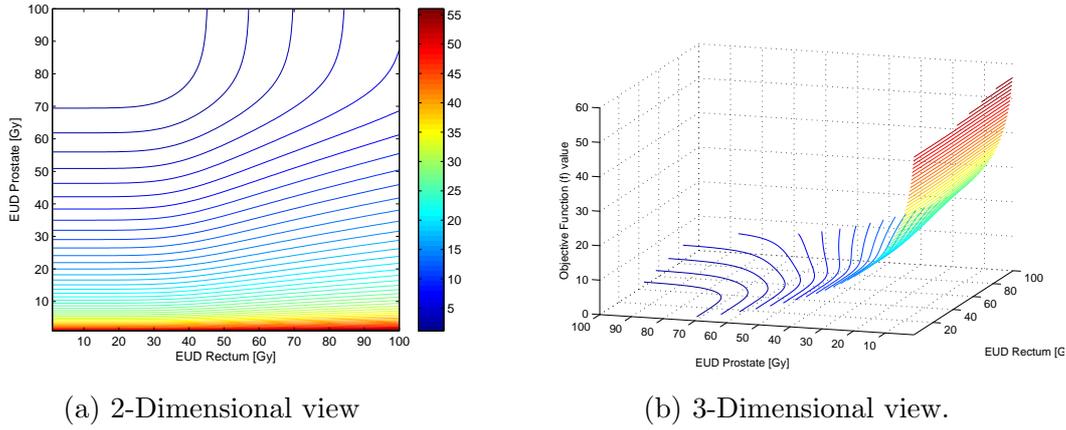


Figure 1: Objective function $f(w)$ contours in the EUD space.

our requirement of $w_j \geq \theta$. Basically, L-BFGS-B algorithm builds and iteratively refines a quadratic model of the function being optimized. Using the information from gradients calculated at previous iterations the algorithm calculates positive definite Hessian approximations. This approximate Hessian matrix is then used to make quasi-Newton step. For a more detailed explanation of the algorithm see (Zhu et al. 1997).

The algorithm requires computation of the first derivative of the objective function. Gradient vectors can be calculated as follows (Olafsson, Jeraj, and Wright 2005).

$$\frac{\partial (-\ln L(w; T, a_T, v_T, eud_0^T))}{\partial w_j} = -\frac{v \cdot (1 - L)}{|T| \cdot gEUD(w; T, a)^a} \sum_{i=1}^{|T|} (D_i^{a-1} A_{ij}) \quad (6)$$

$$\frac{\partial (-\ln U(w; N, a_N, v_N, eud_0^N))}{\partial w_j} = \frac{v \cdot (1 - U)}{|N| \cdot gEUD(w; N, a)^a} \sum_{i=1}^{|N|} (D_i^{a-1} A_{ij}) \quad (7)$$

where $gEUD$, L and U were defined in Equations 2, 4 and 5 respectively.

4.2 Metaheuristics

Differential Evolution (DE)

DE (Storn and Price 1997) is a parallel direct search method based on evolutionary algorithms (EAs). DE combines simple arithmetic operators with the classical crossover, mutation, and selection operators within an easy to implement scheme and with few control parameters. The fundamental idea of DE is a scheme for generating trial solutions by adding the weighted difference vector between two population members to a third one. The DE algorithm is summarized in the following steps:

Population initialization: Initialize each solution k of the first generation of the population ($t = 1$), $w_j^k(t)$, according to a uniform probability distribution.

Mutation or differential operation: Then, the algorithm generates a differential vector z_j^k for each $w_j^k(t)$ solution of the population at generation t according to Equation (8):

$$z_j^k = w_j^{r_1}(t) + F \cdot [w_j^{r_2}(t) - w_j^{r_3}(t)], \quad (8)$$

where k is the solution's population index at generation t ; r_1, r_2, r_3 are three randomly generated integers (for the k^{th} solution) with uniform distribution and their values are lower than or equal to the population size, and mutually different. F is the *mutation factor* ($F > 0$) which controls the amplification of the difference between two individuals and which is normally fixed for the run of the algorithm.

Recombination operation: In order to increase the diversity of the new trial solution $w_j^k(t+1)$, a recombination operator is applied by replacing certain intensity values of solution $w_j^k(t)$ by the values of the previously generated differential vector z_j^k . The values to be replaced are randomly selected with a uniform distribution according to the *recombination rate* $CR \in [0, 1]$. The replacement is done as follows:

$\forall j$ intensity value of $w_j^k(t)$:

If $\text{Rand}(j) \leq CR$ then $w_j^i(t+1) = z_j^k$

Otherwise, $w_j^k(t+1) = w_j^k(t)$

Selection operation: If the new trial solution $w_j^k(t+1)$ is better than $w_j^k(t)$, then the latter is replaced by the new trial solution.

After preliminary experimentation, the DE variant which uses a binomial discrete recombination (DE/Random/1/bin) showed better performance than the one using an exponential recombination (DE/Random/1/exp). For more details about these DE variants, see Das and Suganthan (2011). Therefore, we will use the DE/Random/1/bin algorithm in the experimentation with parameters CR and F set to 0.9 and 0.5, respectively.

Real-coded CHC

Originally, CHC (Eshelman 1991) was proposed as a binary-coded EA combining a selection strategy with high selective pressure and several components inducing a strong diversity. However, as the FMO problem is a real-valued parameter problem, we have extended the above CHC scheme to deal with real-coded solutions. The main components of the real-coded CHC are:

An elitist selection: The solutions of the population are merged with the new population and the best individuals are selected to compose the new population.

A highly disruptive crossover: The blend crossover (BLX- α) crossover operator guaranteeing that the two trial solutions are always at the maximum binary-converted Hamming distance from their two parents, thus proposing the introduction of a high diversity in the new population and reducing the tendency of premature convergence. For more details about the BLX- α crossover operator see (Eshelman and Schaffer 1993).

An incest prevention mechanism: Before mating, the Hamming distance between the potential parents is calculated and if half this distance does not exceed a fixed difference threshold, they are not allowed to mate and no trial solution coming from them is included in the population. Therefore, only the most diverse potential parents are mated. However, the required diversity automatically decreases as the population naturally converges.

Besides, CHC is also characterized by a restart mechanism to encourage the achievement of a suitable and fast rate of convergence. For our experimentation we set the population size to 50 individuals and $\alpha = 0.5$ for the BLX- α operator.

5 Computational experiments

In the section the prostate case generated using CERR (Deasy, Blanco, and Clark 2003) is presented. Obtained results are also analysed in this section. The target corresponds to the prostate plus some margin and OARs are bladder and rectum. Four a priori fixed coplanar beam angles were considered with a total of 116 beamlets whereas the total number of voxels is more than 22,000. Equation (3) is determined by parameters a , ν and eud_0 . For the target, a , ν and eud_0 are -8, 12 and 74, respectively. The values for the bladder are 2, 5 and 30; and the values for rectum are 8, 6 and 40, respectively (Thomas et al. 2005; Peng et al. 2012; Wu et al. 2002).

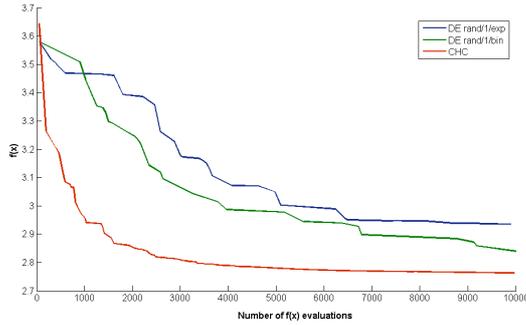
In order to enrich the experimentation with a low quality baseline for the approaches presented in the paper, a random search algorithm is implemented. The random search algorithm randomly generates solutions to the FMO problem until a stopping criteria is achieved (number of evaluations of the objective function). Each solution w_j^k is generated by randomly setting each intensity with a value within a range $[LB_{w_j}, UB_{w_j}]$. The best solution is always maintained, being the output of the algorithm.

In Table 1 the final objective values reached by the algorithms as well as the EUD values for each organ are shown. Each algorithm was run 10 times to study the stability of its results. Then, we show the mean and standard deviation of the results in all the runs. In addition, the time spent by the algorithms are listed in the last column of the table. The time of the metaheuristic algorithms is fixed by the number of evaluations which is their stopping criterion.

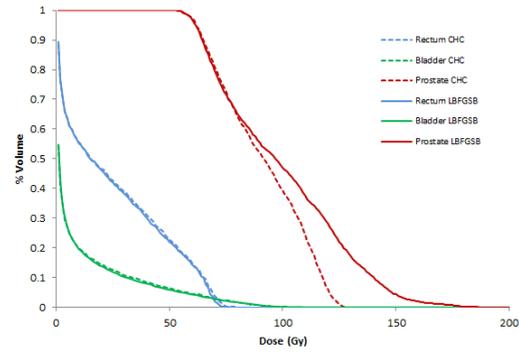
Table 1: Mean and standard deviation of the objective function, EUD values, and elapsed time in the 10 runs of the metaheuristic and mathematical methods.

Method		$f(x)$	EUD(Prostate)	EUD(Bladder)	EUD(Rectum)	Time (s)
Random search	\bar{x}	3.697	73.993	22.370	62.887	213.736
	σ	0.105	1.492	2.039	1.887	0.036
DE	\bar{x}	2.952	74.068	21.034	55.551	215.834
	σ	0.040	1.285	0.724	1.001	0.043
CHC	\bar{x}	2.778	74.258	20.620	54.094	214.025
	σ	0.006	0.229	0.129	0.188	0.055
L-BFGS-B	\bar{x}	2.706	74.490	20.488	53.577	53.300
	σ	0.001	0.044	0.034	0.031	3.622

As can be observed, the best results are obtained by the L-BFGS-B method. The value of the objective function after 53 seconds is 2.706. CHC is the best metaheuristic algorithm obtaining better results than DE. Clearly, the random search



(a) Convergence of the DE and CHC algorithms.



(b) DVH plots for the solutions obtained by L-BFGS-B and CHC.

baseline is easily outperformed by all the methods, showing that this FMO problem cannot be solved by means of a simple random-based algorithm.

The objective function and EUD values achieved by the L-BFGS-B method and the CHC algorithm are very similar. For instance, the difference in the objective function is just 0.072 and the EUD values are also close. However, the L-BFGS-B method is faster than all the metaheuristic algorithms. It only spends 53 s to obtain these results when CHC needs almost four times longer, that is 214 s.

The evolution of the objective function values in the metaheuristic algorithms can be analysed from Figure 2a. The descent of the objective value is plotted through the function evaluations of the DE and CHC algorithms. The maximum number of evaluations was 10,000. The better performance of the CHC algorithm with respect to the DE algorithm is clear. But what is more important is the rapid descent of the objective function value in the first 2,000-3,000 evaluations. Although the metaheuristic algorithms are still improving the objective value during the last 8,000 evaluations, these improvements are small, needing more time than the L-BFGS-B method in achieving the same results.

Doses obtained by L-BFGS-B and CHC are presented in a DVH plot in Figure 2b. This tool is widely used in radiotherapy to evaluate the performance of treatment plans. We can see how the irradiated volumes of the organs are very close in the cases of CHC and L-BFGS-B. The most important difference is that the solution provided by CHC presents a more homogeneous irradiation to the prostate than those one provided by L-BFGS-B.

6 Conclusions and future work

In this work we have compared two different strategies to solve the a gEUD based model for the FMO problem. The well known L-BFGS-B algorithm was compared with two metaheuristics, the CHC and DE evolutionary algorithms. The algorithms were compared according to the objective function value, computational time, and the DVH generated by their solutions.

We found that in terms of objective function and the EUD values, the L-BFGS-B method was just slightly better than the metaheuristics. CHC showed better results than DE and although the CHC algorithm is slower than the L-BFGS-B method, it is quite competitive for the problem. Moreover, in terms of DVH we can conclude that the CHC algorithm tends to generate more uniform doses for the target than

the L-BFGS-B method.

Despite these promising results, more experimentation and optimization analysis must be done. Other metaheuristics exploiting some of the features we drafted in this paper could be a fruitful research area. Furthermore, solving BAO and biological-based FMO models by combining mathematical and metaheuristics could be another very interesting research field.

Acknowledgments

This work has been supported by BECASCHILE and the Spanish Ministerio de Ciencia e Innovación under project TIN2009-07727 including EDRF funding.

References

- Aleman, D.M., D. Glaser, H.E. Romeijn, and J.F. Dempsey. 2010. “Interior point algorithms: guaranteed optimality for fluence map optimization in IMRT.” *Physics in Medicine and Biology* 55 (18): 5467.
- Bertsimas, D., V. Cacchiani, D. Craft, and O. Nohadani. 2012. “A hybrid approach to beam angle optimization in intensity-modulated radiation therapy (in press).” *Computers & Operations Research*.
- Choi, B., and J.O. Deasy. 2002. “The generalized equivalent uniform dose function as a basis for intensity-modulated treatment planning.” *Physics in Medicine and Biology* 47 (20): 3579.
- Das, S., and P.N. Suganthan. 2011. “Differential evolution: A survey of the state-of-the-art.” *IEEE Transactions on Evolutionary Computation* 15 (1): 4–31.
- Deasy, J.O., A.I. Blanco, and V.H. Clark. 2003. “CERR: A computational environment for radiotherapy research.” *Medical Physics* 30 (5): 979–985.
- Ehrgott, M., C. Gler, H.W. Hamacher, and L. Shao. 2008. “Mathematical optimization in intensity modulated radiation therapy.” *JOR* 6:199–262.
- Ehrgott, M., and R. Johnston. 2003. “Optimisation of beam directions in intensity modulated radiation therapy planning.” *OR Spectrum* 25:251–264.
- Eshelman, L.J. 1991. “The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination.” Edited by G.J.E. Rawlins, *Foundations of Genetic Algorithms*. Morgan Kaufmann, 265–283.
- Eshelman, L.J., and J.D. Schaffer. 1993. “Real-coded genetic algorithms and interval-schemata.” Edited by L.D. Whitley, *Foundations of Genetic Algorithms*, Volume 2. Morgan Kaufmann, 187–202.
- Hartmann, M., and L. Bogner. 2008. “Investigation of intensity-modulated radiotherapy optimization with gEUD-based objectives by means of simulated annealing.” *Medical Physics* 35:2041.
- Niemierko, A. 1997. “Reporting and analyzing dose distributions: A concept of equivalent uniform dose.” *Medical Physics* 24 (1): 103–110.

- Olafsson, A., R. Jeraj, and S.J. Wright. 2005. "Optimization of intensity-modulated radiation therapy with biological objectives." *Physics in Medicine and Biology* 50 (22): 5357.
- Peng, F., X. Jia, X. Gu, M.A. Epelman, H.E. Romeijn, and S.B Jiang. 2012. "A new column-generation-based algorithm for VMAT treatment plan optimization." *Physics in Medicine and Biology* 57 (14): 4569.
- Romeijn, H.E., J.F. Dempsey, and J.G. Li. 2004. "A unifying framework for multi-criteria fluence map optimization models." *Physics in Medicine and Biology* 49 (10): 1991.
- Schreibmann, E., M. Lahanas, L. Xing, and D. Baltas. 2004. "Multiobjective evolutionary optimization of the number of beams, their orientations and weights for intensity-modulated radiation therapy." *Physics in Medicine and Biology* 49 (5): 747.
- Storn, R., and K. Price. 1997. "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces." *Journal of Global Optimization* 11 (4): 341–359.
- Thomas, E., O. Chapet, M.L. Kessler, T.S. Lawrence, and R.K. Ten Haken. 2005. "Benefit of using biologic parameters (EUD and NTCP) in IMRT optimization for treatment of intrahepatic tumors." *International Journal of Radiation Oncology*Biophysics* 62 (2): 571 – 578.
- Wu, Q., D. Djajaputra, Y. Wu, J. Zhou, H.H. Liu, and R. Mohan. 2003. "Intensity-modulated radiotherapy optimization with gEUD-guided dose - volume objectives." *Physics in Medicine and Biology* 48 (3): 279.
- Wu, Q., R. Mohan, A. Niemierko, and R. Schmidt-Ullrich. 2002. "Optimization of intensity-modulated radiotherapy plans based on the equivalent uniform dose." *International Journal of Radiation Oncology*Biophysics* 52 (1): 224 – 235.
- Zhu, C., R.H. Byrd, P. Lu, and J. Nocedal. 1997. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization." *ACM Transaction on Mathematical Software* 23 (4): 550–560.

Structural Equation Modelling of Undergraduate Management Students' Perceptions of Feedback in a New Zealand University

Robert Y. Cavana, Kala S. Retna & Arthur Ahimbisibwe
Victoria Business School, Victoria University of Wellington, New Zealand
bob.cavana@vuw.ac.nz

Abstract

A previous paper presented the results of a survey based on a large sample of undergraduate students on management courses in a New Zealand university, exploring students' perceptions of feedback given to them on their formative assessments. A factor analysis and simplified regression analysis were undertaken with the collected data.

In this paper, we discuss a more advanced statistical analysis of this data involving Structural Equation Modelling. Data were subjected to exploratory factor analysis to identify the underlying relationships between measured variables and to identify a set of latent constructs underlying a sequence of measured variables. Using SPSS v19, a factor analysis was conducted by the Principal Components Analysis approach with Varimax rotation to confirm the suitability of the construct indicators. Six meaningful factors explaining 60% of the common item variance emerged.

Confirmatory Factor Analysis was then conducted for examining construct validity by assessing how well individual items represent the construct. The results from the final measurement model were used to create the structural model that tested the strength and significance of the theorized relationships. The final Structural Equation Model with path coefficients is provided.

The results and insights from this statistical analysis will be discussed in this paper.

Key words: feedback, formative assessment, higher education, student learning, survey analysis, structural equation modelling.

1. Introduction

Reforms in universities are appearing in various forms that aim to contribute to the quality of learning and teaching. In the pursuit of efforts to enhance student learning, paying attention to feedback is one aspect that plays a central role in understanding the relationship between student progress and achievement (Bandura, 1991; Espasa & Meneses, 2010; Fedor, 1991; Weaver, 2006;). Feedback on learning from students and teachers is also one of the key areas of concern for New Zealand (NZ) universities as reported in recent research (VUW, 2009). Students are paying customers of tertiary institutions and part of their demands for quality education is receiving feedback for assessments and coursework. Also, in recent times, much emphasis has been made to shift from teacher to student-centred learning (Rust, 2002). Emphasis on student-centred

learning is part of the global quality movement that seeks to address accountability in all aspects of higher learning (Leckey & Neill, 2001). One aspect of the quality accountability by universities is the quality of feedback that is given through formative assessments. In a recent empirical study, Retna, Chong and Cavana, (2009) also emphasised the importance of feedback to student satisfaction and learning in tutorials.

A previous paper by Retna & Cavana (2010, 2013) briefly reviews the literature on feedback and its importance in relation to student learning. That paper also outlined an empirical survey of undergraduate management students at a NZ university, and provided a quantitative and qualitative analysis of the students' perceptions on feedback. The quantitative analysis involved a factor analysis and simplified regression analysis. The results supported the initial hypotheses that improvement of performance (both work quality and results) and the quality of feedback, lead to higher levels of overall student satisfaction with feedback provided on management courses (by tutors and lecturers).

In this paper, we discuss a more advanced statistical analysis of this data involving Structural Equation Modelling. Data are subjected to exploratory factor analysis to identify the underlying relationships between measured variables and to identify a set of latent constructs underlying a sequence of measured variables. Confirmatory Factor Analysis is then conducted for examining construct validity by assessing how well individual items represent the construct. The results from the final measurement model are used to create the structural model that tests the strength and significance of the theorized relationships. The final Structural Equation Model with path coefficients is provided and briefly discussed.

2. Method

A survey questionnaire was used in this research as the primary tool for collecting data. In an educational setting, the use of a questionnaire is a useful approach in terms of factors such as time and efficiency. The anonymity of a questionnaire allows students to respond with ease and comfort without the perceived fear of being penalised in their assessments. In order to identify some attributes experienced by students on receiving feedback on their assessment, a small scale pilot study (85 students) was conducted with a third year management course at a New Zealand university. Using the literature on student feedback and also from the analysis of the pilot study, a questionnaire was derived focusing on three main dimensions: improvement of performance, the need for feedback and quality of feedback.

One faculty administrator and one academic, who were not involved in teaching, administered the questionnaires with 828 students on undergraduate management courses in the Commerce Faculty at a New Zealand university. To avoid the presence of academics and tutors who had been involved in the programme, the survey was conducted during the last lesson of the trimester. Prior permission was sought from lecturers involved in the programme to leave their classroom before the survey was conducted. All participants of this research were third year undergraduates and were selected for three reasons: accessibility, large sample and their rich experiences of receiving feedback for their assessments. Though, 828 questionnaires were administered, only 613 were returned, a response rate of 74 per cent.

The questionnaire consisted of three parts and served to fulfil the quantitative, qualitative and demographic profiles for analysis. The first part had 20 questions and related to the quality of feedback, improvement of performance, and need for feedback

by students, with one key question on the overall satisfaction of feedback given on management courses. A 5-point itemised Likert rating of Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree was used for data collection, with Strongly Agree coded as 5; to Strongly Disagree as 1.

The second part had two questions that required students to suggest specific things that feedback had helped in their learning and also to list two to three types of their preferences for feedback. The final part of the questionnaire gathered demographic information such as age, nationality/ethnicity, and gender. Demographic details of the respondents to the research are summarised in Table 1.

Table 1. Demographic details

Demographic information regarding the respondents		
Age Group (years)		
17 or less	6	1%
18 - 20	383	63%
20 - 25	182	30%
25 - 30	21	3%
30 + above	15	2%
	607	100%
Nationality / Ethnicity *		
NZ Maori	54	9%
Pacific Islander	20	3%
NZ European/Pakeha	365	57%
Chinese	63	10%
Other	133	21%
	635	100%
Gender		
Male	277	46%
Female	331	54%
	608	100%

* includes some double selections

3. Data analysis and results

In this study, data analysis involved two steps: the factor analysis and the structural relationship analysis using the Structural Equation Modelling (SEM) method. The Exploratory factor analysis aimed to identify the underlying relationships between measured variables and to identify a set of latent constructs underlying a sequence of measured variables. The structural relationship analysis was used to examine the simultaneous relationship between all the constructs.

3.1 Exploratory Factor Analysis

Data were first subjected to exploratory factor analysis to identify the underlying relationships between measured variables and to identify a set of latent constructs underlying a sequence of measured variables. Using IBM SPSS Statistics 19, a factor analysis was conducted by the Principal Components Analysis (PCA) approach with Varimax rotation to confirm the suitability of the construct indicators. Varimax rotation generally yields more stable results and is easier to interpret. PCA approach was chosen because it provides a linear summarization of the data into simpler components and produces exact scores rather than estimates. PCA is also the simplest of the true eigenvector-based multivariate analyses that often reveals the internal structure of the

data in a way that best explains the variance in the data by providing the user with a lower-dimensional picture when viewed from its most informative viewpoint. Six meaningful factors explaining 60.1% of the common item variance emerged and are reported in Table 2. Additionally, all items loaded cleanly on the hypothesized constructs exceeding .50 (Hair et al., 2009). Although one item's (I use feedback to improve my results) factor loading was below .5 and had cross loadings was dropped and not considered in subsequent analyses because this demonstrated lack of construct validity. The six hypothesised constructs were interpreted as follows: factor 1 was termed '*Need for feedback*' while factor 2 was termed '*How to improve*'. Factor 3 was termed '*Poor quality feedback*', factor 4 was termed '*Support for feedback*', factor 5 was termed '*Importance placed on Feedback*' and factor 6 was termed '*Improvement of work quality*': and is the endogenous (dependent) variable in this study. Table 2 shows the rotated component matrix for factor analysis.

Table 2. Rotated component matrix^a

	Component					
	Factor					
	1	2	3	4	5	6
NFB1	.555	.184	.139	.338	-.359	.040
NFB2	.747	.102	-.020	-.079	.067	.135
NFB3	.769	.016	-.002	.014	-.193	.084
NFB4	.759	.167	.004	-.037	.104	.014
NFB5	.584	-.040	.041	.297	.012	-.081
HTI1	.130	.776	-.019	.131	-.055	.154
HTI2	.100	.762	.008	-.063	.230	.194
PQF1	.014	.066	.764	.017	.035	-.111
PQF2	.057	-.068	.707	-.160	.248	-.040
SFF1	.151	.326	-.072	.597	-.041	.348
SFF2	.037	-.139	-.075	.626	.318	.286
IPF1	-.028	.036	.434	.110	.533	.091
IPF2	-.026	.104	.186	.094	.694	.059
IWQ1	.115	.398	.176	.074	-.148	.551
IWQ2	.024	.152	-.100	.192	.082	.700
IWQ3	.057	.050	-.160	.084	.247	.747
IWQ4	.043	.033	.087	-.004	.018	.835
IWQ5	.044	.193	-.049	.210	-.170	.691
IWQ6	-.032	.084	-.288	.447	.203	.563

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.
 a. Rotation converged in 12 iterations.

The survey responses were anchored between *strongly disagree* (1) and *strongly agree* (5). Hence all responses of 3.0 indicate neutrality and responses greater than 4.0 indicate strong agreement. Mean responses for all the 19 items were in the range of 2.28-4.51. All the mean and standard deviation values are reported in Table 3.

3.2 Confirmatory factor analysis and Structural Equation Modelling

In tracing structural relationships between these constructs, SEM in AMOS Version 19.0 was employed for data analysis. SEM involves estimation of parameters by

minimizing the discrepancy between the model-implied covariance matrix and the observed covariance matrix (Joreskog & Sorbom, 1989). SEM was chosen because it is a confirmatory approach that provides explicit test statistics for establishing convergent and discriminant validity important to management research, tests an overall model rather than individual coefficients, allows for error terms and reduces measurement error through the use of multiple indicators (Kearns & Sabherwal, 2007). The approach chosen was to separate analysis of the measurement and structural models in a two-step process recommended by Anderson and Gerbing (1988). This allows refinement of measures before testing of the structural model and is consistent with previous studies (Byrne, 2001; Kearns & Sabherwal, 2007). In the first phase, a measurement model was used to measure the fit between the theorized model and observed variables and to establish reliability and validity (eg see Cavana, Delahaye & Sekaran, 2001, pp209-215). In the second phase, results of the measurement model were used to create a structural model in order to measure the strength of the theorized relationships.

3.3 The measurement model

Using the raw SPSS data file as input, Confirmatory Factor Analysis (CFA) was conducted for examining construct validity by assessing how well individual items represent the construct. Based on the larger sample size (N=613), the Maximum Likelihood (ML) was selected since it provides more reliable results. The robustness of SEM using ML estimation has been demonstrated in previous studies especially in dealing with non-experimental data and in reporting traditional as well as robust goodness of fit measures; that are accurate when even the data violates the normalcy rule of ML estimation (Kearns & Sabherwal, 2007; Byrne, 2001). Construct validity and discriminant validity were tested first to demonstrate the dimensionality of the constructs in the measurement model and nomological validity was also tested for the robustness of the structural model.

3.3.1 Construct validity

Construct validity was examined in two ways: (1) High factor loadings with significant t-values are mostly good indicators of the construct validity and (2) A high squared correlation value for a construct also indicates good construct validity (Anderson & Gerbing, 1988; Joreskog & Sorbom, 1989). All six constructs demonstrated good model fit when subjected to Hu and Bentler's (1999) criteria and Rigdon's (1996) criteria. All the factor loadings of all items in all six constructs were high ($\lambda > 0.4$) and significant ($p < 0.001$), far above the usual statistical significance cut off criterion ($p < .05$) and hence acceptable. Similarly, the corresponding t-values ($t \geq 9.08$) indicated that all the factor loadings were significant as shown in Table 2. This provides evidence that the measurement items are significantly related to their construct measures. Following guidelines by Fornell and Larcker (1981), squared correlation values were then calculated for each construct. The Average Variance Extracted (AVE) for each of the six constructs exceeded the suggested threshold of 0.50, indicating that the variance captured by a construct was larger than the variance due to measurement errors. Hence, the construct validity of the six constructs was established.

3.3.2 Content validity

Content validity was achieved by selection of survey items from the existing management theory and used tested items from previous research (Retna et al., 2009). The research instrument was first reviewed by three professors from the Victoria School of Management. A pilot testing by undergraduate management students for instrument

refinement was also conducted. Based on these responses and comments, item scales that were unclear and ambiguous were either improved or deleted. Following the guidelines set forth by Dillman (1991) questions were brief and to the point, addressing only a single issue at a time and avoided phrases that could elicit a socially acceptable response.

3.3.3 Construct reliability

Although, traditionally Cronbach alpha Coefficient is used to assess construct reliability, it suffers from restrictive assumption of equal importance of all indicators. For this matter, the standard errors based on maximum likelihood were computed in SEM to provide an indication of reliability with confidence intervals based on the t-distribution. The underlying logic here is that as the standard error approaches zero, the test statistic for its related parameter cannot be estimated. Similarly, standard errors that are extremely large indicate parameters that cannot be determined (Byrne, 2001, p75). Since standard errors are influenced by the units of measurement in observed or latent variables, as well as the magnitude of the parameter itself; no definitive criterion of small or large has been confirmed (Joreskog & Sorbom, 1989). The final results of the measurement model for the six constructs and their relationships with respective items are all reported in Table 2.

Table 2. Survey item loadings, construct validity and reliability from final measurement

N=613							
Construct	Survey Question	Mean	Stand. Dev.	Stand. loading	SE	t-value	R-Square
Need for feedback	Feedback is important to me (NFB1)	4.37	.765	.622	.032	14.85	.386
	I deserve feedback when I put so much effort in (NFB2)	4.33	.792	.631	.033	15.18	.398
	I always read the feedback on my assignments (NFB3)	4.51	.722	.714	.029	17.58	.510
	It is more important for me to see the reason why I received a particular grade (NFB4)	4.36	.746	.638	.031	15.34	.407
	I always collect my assignments (NFB5)	4.31	.859	.473	.037	10.93	.224
How to improve	Feedback tells me what I need to do to improve my performance (HT11)	4.25	.710	.666	.037	12.61	.444
	Feedback tells me what the expectations of the tutors are (HT12)	3.92	.798	.602	.040	11.86	.362
Poor quality feedback	Feedback was inconsistent or contradictory (PQF1)	3.08	1.536	.507	.080	9.68	.257
	Gave feedback that I couldn't understand (PQF2)	2.82	.995	.644	.058	11.00	.414
Support for feedback	I use feedback to improve my results (SFF1)	3.66	.866	.663	.046	12.60	.440
	Marker offered opportunities to clarify their feedback (SFF2)	3.19	1.030	.437	.047	9.47	.191
Importance placed on Feedback	Feedback is only useful when it is positive (IPF1)	2.28	1.163	.547	.067	9.54	.299
	The grade is more important to my learning than feedback (IPF2)	3.32	1.076	.497	.059	9.08	.247
Improvement of work quality'	Feedback made me think further about the topics (IWQ1)	3.62	.889	.532	.036	13.09	.283
	Feedback was provided that I could use in future assignments/courses (IWQ2)	3.58	.994	.707	.037	18.764	.500
	Critical feedback was given on the quality of the work (IWQ3)	3.45	.901	.704	.034	18.62	.496
	Feedback showed me how to critically assess my work (IWQ4)	3.35	.887	.677	.034	17.60	.458
	Feedback helped me focus on areas I could improve (IWQ5)	3.89	.870	.675	.033	17.58	.456
	Overall, I was satisfied with feedback given in my management courses (IWQ6)	3.40	.992	.684	.038	17.84	.467

3.3.4 Discriminant validity

The discriminant validity test was performed in order to establish the distinction among all the constructs used in this study (Anderson & Gerbing, 1988). Chi-square (χ^2) difference tests were run for all possible construct pairs. For each pair, a comparison was made between the χ^2 values for the constrained model and the unconstrained model. The constrained model represents a case in which the variances and covariance for the construct pairs were constrained to unity. The χ^2 differences were significantly less for the unconstrained models compared to the constrained models, suggesting that the better model was the one in which the factors are separate but correlated (Anderson & Gerbing, 1988). Discriminant validity was also established for any construct pair, when the AVE of each construct exceeded the square of the construct correlations shown in the Table 3. Similarly, no correlation exceeded the prescribed limit of 0.90 which was a good indicator that there was no item redundancy and that multicollinearity was also absent in this study. Also, using bootstrapping a confidence interval of (+/-) 4 standard errors was calculated for each of the construct correlations to determine if the interval contained the value 1. None of the confidence intervals contained the value 1 indicating that the correlations between these constructs differ significantly from unity and that the probability of perfect correlation was extremely low (Kearns & Sabherwal, 2007).

Table 3. Correlations for study constructs (N=613)

	F1	F2	F3	F4	F5	F6
F1	1.00	.063	.002	.047	.144	.027
F2	.251**	1.00	.0019	.068	.0087	.162
F3	.048	.044	1.00	.011	.638	.025
F4	.216**	.260**	-.105**	1.00	.000	.308
F5	.012	.093*	.799**	.001	1.00	.058
F6	.164**	.402**	-.158**	.555**	-.028	1.00

Peason correlations appear below the diagonal. Squared correlations appear above the diagonal. **. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

3.3.5 Social desirability, self-reporting and common method variance

Common method variance is a potential problem in research when all measurements are provided by a single respondent. Because self-reporting, consistency motif, acquiescence, social desirability, affectivity and transient mood state lead to common method variance; it is of concern in survey research when sampling perceptual data (Podsakoff, Mackenzie, Lee & Podsakoff, 2003). Common methods bias leads to type 1 & 2 errors where the researchers may accept or reject the null hypothesis when they should not have done so. Common methods bias was addressed in three ways; firstly using the strategies to ameliorate the problems of self-report data by designing a questionnaire to avoid implying that one response is better than the other, avoided socially accepted responses, decomposed questions relating to more than one possibility and avoided complicated syntax (Kearns & Sabherwal, 2007). Common method variance was further assessed using Harman's one factor test (Podsakoff et al, 2003). The underlying logic for this test is that if common method bias accounts for the relations among variables, then a factor analysis should yield a single factor when all the items are analysed together. No single factor emerged or one general factor accounted for most of the variance implying that no substantial common method variance was present. The single factor accounted for 23.94% of the total variance. Finally, a confirmatory factor analysis approach was used to test a model positing that a single latent factor underlies the study variables, by linking all the items to a single

factor for common method variance (Kearns & Sabherwal, 2007). The unstandardized regression weights for the common latent factor were all found to be 0.4, and when squared was 0.16 confirming that common method variance was not a problem in this study.

3.3.6 Non-response bias

Non response bias was established in t-tests by comparing the average values for each of the constructs for the first quartile completed questionnaires received versus the last quartile completed questionnaires allowing the late questionnaires to proxy the perceptions of non-respondents. Mean differences for each of the constructs did not reveal any significant difference between the early and late questionnaires (2-tailed t-tests, $p < 0.05$). This comparative test depicted the absence of non-response bias in this study.

3.4 The structural model

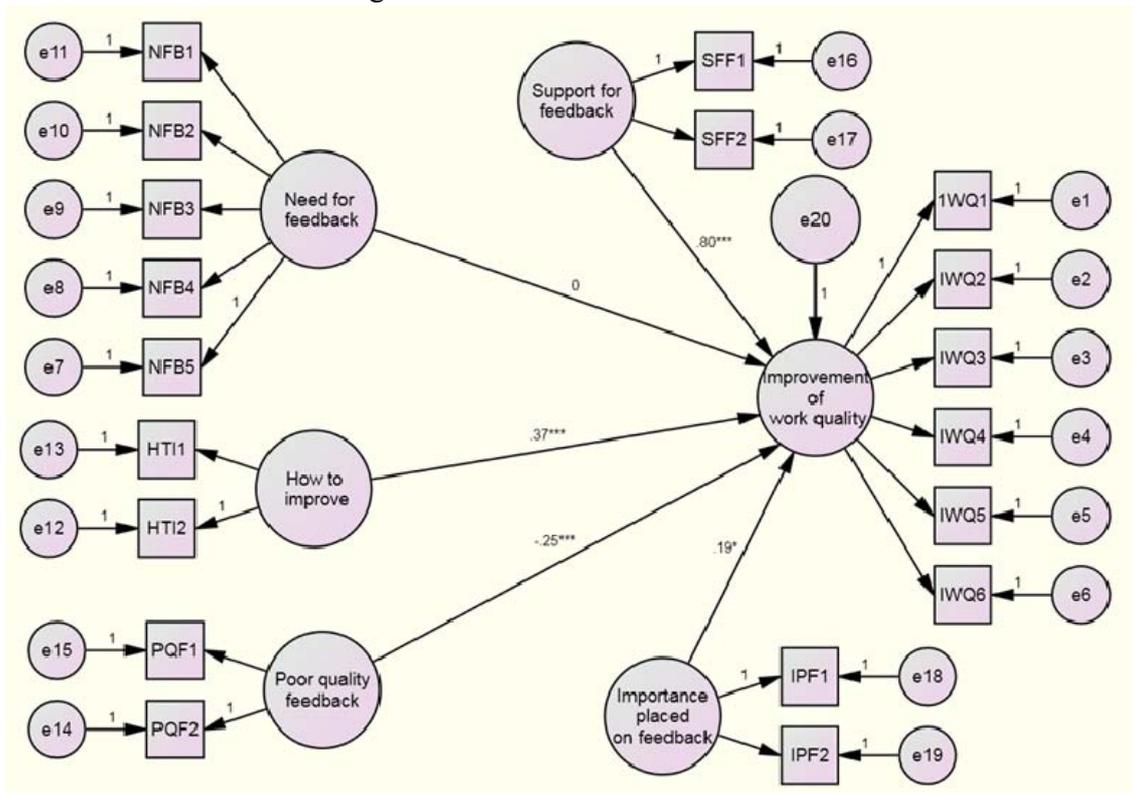
Results from the final measurement model were used to create the structural model that tested the strength and significance of the theorized relationships. The final SEM model with path coefficients is shown in Figure 1. This final structural model accounts for 88.3% of the variation in '*improvement in work quality*'. Thus, the model is very successful in accounting for a substantial portion of the variability in '*improvement in work quality*'. Surprisingly, one path coefficient did not demonstrate the expected results; the path coefficient between '*Need for feedback*' and '*improvement in work quality*' was initially hypothesized to be positive and significant but instead it was found to be non-significant.

3.4.1 Goodness-of-fit

Goodness-of-fit was used to describe how well the statistical model fits the sample data and to determine whether the data supports a hypothesized model in SEM. It was established by multiple indices to negate bias associated with the use of a single index. The measures that were used included among others χ^2/df , TLI, CFI, GFI, AGFI and RMSEA (Joreskog & Sorbom, 1989). Although χ^2 which is a function of the differences between the observed covariances and the covariances implied by the model is recognized as a measure of fit, it is frequently affected by the size of correlations within the model and can produce inaccurate probability values hence it was replaced with the χ^2/df (Kearns & Sabherwal, 2007). The TLI compares the lack of fit of a target model to the lack of fit of a baseline model, usually the independence model and is one of the indexes affected least by sample size. The CFI also has the advantage of reflecting fit at all sample sizes and measures the comparative reduction in noncentrality. Values above 0.85 are desirable for both the TLI and CFI. The RMSEA computes average lack of fit per degree of freedom and is less affected by sample size. For both, values below 0.08 are recognized as adequate (Joreskog & Sorbom, 1989).

The SEM results in Figure 1 suggest that the model adequately fits the data with the following absolute fit indices: $\chi^2=672.899$, $df=147$, $\chi^2/df= 4.578$, probability level=0.00. Other fit indices also suggested a good fit to the model, the CFI=0.813, TLI=0.782, and RMSEA=0.077 and RMR was .080. The goodness of fit indices were also appropriate as follows: GFI=0.894, AGFI=0.863. Based on these values, the final structural model was deemed acceptable since the hypothesised model adequately fits the sample data (Byrne 2001; Joreskog & Sorbom, 1989).

Figure 1. The final structural model



Note: All model path coefficients are standardised. *, **, *** represent *p* (significance) levels of 0.05, 0.01 and 0.001 respectively.

3.4.2 Results from Hypotheses testing

Using 613 observations provided by undergraduate management students, survey data supported 4 of the study's 5 hypotheses. The results reveal a positive and significant association between 'How to improve' and 'Improvement of work quality'. ($\beta=0.37$, $p<0.001$) hence supporting H2: 'How to improve' positively influences 'Improvement of work quality'. Surprisingly, the results showed no significant association between 'Need for feedback' and 'Improvement of work quality' ($\beta=0.00$, $p>.05$) thus rejecting H1: 'Need for feedback' positively influences 'Improvement of work quality'. The results also point out that 'Poor quality feedback' has a negative and significant effect on 'Improvement of work quality' ($\beta=-0.25$, $p<0.001$) hence supporting H3: 'Poor quality feedback' negatively influences 'Improvement of work quality'. The results reveal that 'Support for feedback' has the strongest positive and significant effect on 'Improvement of work quality' ($\beta=0.80$, $p<0.001$) hence supporting H4: 'Support for feedback' positively influences 'Improvement of work quality'. There was also a positive and significant association between 'Importance placed on Feedback' and 'Improvement of work quality'. ($\beta=0.19$, $p<0.05$) hence supporting H5: 'Importance placed on Feedback' positively influences 'Improvement of work quality'.

4. Concluding comments

This paper has examined various aspects of feedback provided to students on their formative assessments, and the satisfaction from this feedback, as expressed by a large sample of undergraduate management students at a New Zealand university. Structural Equation Modelling was undertaken with the data collected from a short questionnaire

administered in the last class of the trimester. The results supported 4 of the 5 hypotheses related to testing the relationships between the independent variables and the dependent variable ie 'Improvement of work quality'.

It must be emphasised that this research is exploratory. Nevertheless, a future paper by the authors will compare the quantitative and qualitative results from the previous study (Retna & Cavana, 2010 & 2013) with the results of the structural equation modelling undertaken in this paper. It is hoped that the further comparative study will reveal some interesting insights and guidelines for future research in this important area, involving providing feedback to tertiary education students to help with their learning and education while they are studying at university.

5. References

- Bandura, A. 1991. "Social theory of self-regulation." *Organisational Behaviour and Human Decision Processes* **50**: 248-287.
- Byrne, B.N. 2001. *Structural Equation Modeling with AMOS: Basic Concepts, Applications and Programming*. Rahwah, NJ Lawrence Erlbaum.
- Cavana, R. Y., B.L. Delahaye, and U. Sekaran. 2001. *Applied Business Research: Qualitative and Quantitative Methods*. Brisbane, Wiley.
- Espasa, A., J. Meneses. 2009. "Analysing feedback processes in an online teaching and learning environment: an exploratory study." *Higher Education* **59**: 277-292.
- Fedor, D. B. 1991. "Recipient responses to performance feedback: A proposed model and its implications." *Research in Personnel and Human Resources Management* **9**: 73-120.
- Fornell C., D.F. Larcker. 1981. "Evaluating structural equation models with unobservable variables and measurement error." *Journal of Marketing Research* **18**: 39-50.
- Hair, J.F., W.C. Black, B.J. Babin, and R.E. Anderson. 2009. *Multivariate Data Analysis* (7th Ed.). Englewood Cliffs, NJ, Prentice Hall.
- Hu, L., P.M. Bentler. 1999. "Cut off criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives." *Structural Equation Modelling* **6**: 1-55.
- Joreskog, K. G., D. Sorbom. 1989. *LISREL 7: A Guide to the Program and Applications*, 2nd ed. Chicago, IL: SPSS.
- Kearns, G.S., R. Sabherwal. 2007. "Antecedents and consequences of information systems planning intergration." *IEEE Transactions on Engineering Management* **54**: 628-643.
- Leckey, J., N. Neill. 2001. "Quantifying quality: the importance of student feedback." *Quality in Higher Education* **7**: 19-32.
- Podsakoff, P. M., S.B. Mackenzie, J. Lee, and N.P. Podsakoff. 2003. "Common methods biases in behavioral research: a critical review of the literature and recommended remedies." *Journal of applied Psychology* **88**: 879-903.
- Retna K.S., R.Y. Cavana. 2010. "An exploratory analysis of undergraduate management students' perceptions of feedback in a New Zealand University", *Proceedings of the 24th ANZAM conference, Adelaide* (Adelaide, Australia, University of South Australia).
- Retna K.S., R.Y. Cavana. 2013. "Undergraduate management students' perceptions of feedback in a New Zealand University." *Journal of Management and Organization* [Forthcoming]
- Retna K. S., E. Chong, and R.Y. Cavana. 2009. "Tutors and tutorials: students' perceptions in a New Zealand University." *Journal of Higher Education Policy and Management* **31I**: 251-260.
- Rigdon, E.E. (1996). "CFI versus RMSEA: a comparison of two fit indices for structural equation modelling." *Structural Equation Modelling* **3**: 369-79.
- Rust, C. (2002). "The impact of assessment on student learning." *Active Learning in Higher Education* **3**: 145-148.
- VUW 2009. *Symposium on Tertiary Assessment and Higher Education Student Outcomes*, held at Victoria University of Wellington, New Zealand, in November 2008.

Review of Modelling for LMPs in Electricity Markets

Bhujanga B Chakrabarti
System Operations, Transpower
bhujanga.chakrabarti@transpower.co.nz

Ramesh Rayudu
Victoria University, Wellington
Ramesh.Rayudu@vuw.ac.nz

Abstract

This article reviews the transmission branch power flow and the loss modelling in the market clearing model of New Zealand Electricity Market (NZEM) and the Independent System Operators (ISOs) in the US. A case study is conducted with a loss less 5-bus network using the two methods of modelling of transmission flow. It is shown that the two methods are technically equivalent.

Next, it describes some outstanding anomaly in the determination of Locational Marginal Price (LMP) of electricity. Unique LMP with constant and stepped demand curves is discussed, and non-unique LMPs, particularly multiple dual solutions are also discussed. It also looks at the solutions prescribed in the literatures, so far.

Key Words: Branch flow model, Loss model, LMP market, Multiple dual solutions, PTDF

1 Introduction on LMP Markets

In the deregulated operation of the New Zealand Electricity Market (NZEM), Transpower, the System Operator (SO) uses a security constrained DC OPF based market clearing engine to determine the dispatch and the nodal prices. This spot pricing market to dispatch electricity has been in operation since 1 October, 1996. The clearing engine is called the Scheduling, Pricing and Dispatch (SPD) model. The model uses the network configuration at the time of scheduling, load bids or load forecast, offers from the market participants for different products (eg, energy, 6 second reserves, 60 second reserves) to compute a price based dispatch presented by Alvey, Goodwin, Ma, Streiffert, & Sun(1998). The model also uses a linearised loss model for the transmission circuits. The SPD model co-optimises energy and the different reserves. ISOs in the US are also LMP markets introduced in late Nineties or early 2000s. It models some components differently than ours in their market clearing engine, but the end outcome is similar to our LMP. This article describes only branch flow model and loss models for both markets.

A number of problems are emerging as the market is maturing. We find a number anomalies such as non-existent of price, primal multiple dispatch solutions, dual multiple price solutions. These are briefly reviewed and presented solutions prescribed in the literature, so far.

2 Comparison of Branch flow and Loss Modelling

There are a number of differences in modelling of components in market clearing engines between NZEM and the ISOs in the US. We only describe the flow and loss modelling.

2.1 Flow and Loss models in the NZEM

2.1.1 Flow model

Power flows in the line between i and j is given by (2.1). We introduce transmission loss (P_{ij}^L) in the line flow equation. A factor of $\frac{1}{2}$ is used to discount double counting of transmission loss twice for each direction for the same transmission line. This is further discussed in the next section.

$$(2.1) \quad P_{ij} = B_{ij}(\theta_i - \theta_j) + \frac{1}{2} P_{ij}^L : \tau_{ij}; (\forall (i, j) \in L_{ij})$$

Here $j \in N_i$ signifies set of all lines connected at i between i and j 's.

$(\forall (i, j) \in L_{ij})$ defines the set of all lines connected between (i, j) at both i and j

$B_{i,j}$ = Line susceptance between i and j , in pu

θ_i, θ_j = Bus voltage angles at bus i and j , in radian.

The power balance at each bus must be preserved. Given the conventions defined above, these are given by (2.2).

$$(2.2) \quad P_{gi} - P_{di} = \sum_{j \in N_i} P_{ij} : \lambda_i; \forall i$$

$j \in N_i$: All nodes connected to bus i , as defined earlier.

$P_{d,i}$ Demand at bus I , $P_{g,i}$ generation at bus i .

Notice that the branch losses are now coupled with power balance constraint (2.2) at each bus through branch flows. λ_i is the dual variable associated with the power balance equality constraint. It represents the locational marginal energy price at the bus i . At the solution, λ_i captures the marginal generation cost, marginal cost of loss and the marginal cost of network congestion and the shadow prices of other binding constraints.

2.1.2 Linear loss model

There are different ways of defining a simplified transmission loss formula. The loss formula and sensitivities depend on the selection of slack bus (single, distributed slacks). We describe the piece-wise linear loss model that is used in NZEM. The transmission loss along a line can be represented, using usual notation, as in (2.3) and a paper presented by Chakrabarti, Edwards, Callaghan, & Ranatunga (2011).

$$(2.3) \quad \begin{aligned} P_{ij}^L &= g_{ij}(V_i^2 + V_j^2 - 2V_i V_j \cos \theta_{ij}) ; \forall (i, j) \in L_{ij} \\ P_{ij}^L &= 2g_{ij}(1 - \cos \theta_{ij}) ; \forall (i, j) \in L_{ij} \\ P_{ij}^L &= g_{ij}\theta_{ij}^2 ; \forall (i, j) \in L_{ij} \end{aligned}$$

It is seen that under the simplest form the transmission losses relate quadratically to the angle difference. In per unit system, it can be expressed as (2.4).

$$(2.4) \quad P_{ij}^L = R_{ij} P_{ij}^2 ; \forall (i, j) \in L_{ij} \text{ Where, } R_{ij} \text{ is the branch resistance, in pu.}$$

Therefore using linear approximation for losses, it is inappropriate to evaluate the loss due to the whole flow in the line using the loss sensitivity of only one segment. However when the range of power flow, and hence the angle is known, the linear approximation can be used. Fig.1 shows the four linear segments which fit best the

quadratic loss curve. The x-axis represents the MW flow and the y-axis represents the loss in different segments. The angle between each segment with the horizontal line gives the loss sensitivity for the segment. That means loss sensitivity remains constant for that whole segment.

The SPD model in NZEM uses 1 loss segment for transformers, 3 loss segments for AC lines and 6 loss segments for HVDC lines. For example, AC lines use variable segment length method and set the break-points on the power flow axis at 0, 0.310134, 0.690025 and 1pu of line rating.

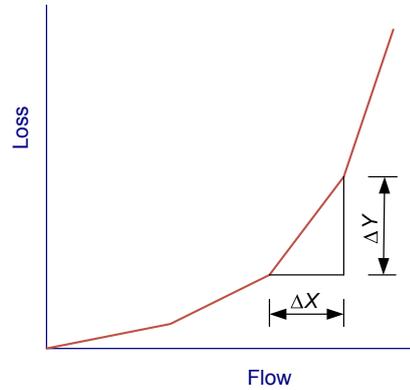


Figure-1. Linear Loss segments

It can be noted that the break points are linear function of the line ratings and the loss sensitivities are the linear function of the average gradient. Therefore these can be calculated off-line for future uses. Static loss of each branch is modeled as load equally at each end. Dynamic loss for each branch is modeled at the receiving end only.

The SPD determines the flow while solving the optimisation problem. While it determines flow in a line it also takes care the various constraints related to loss calculation.

For example, for 3segment loss calculation in AC line, using loss coefficient: A = 0.310134

Loss segment 1

- a. Seg 1 folw, $F1 = (\text{coff } A) * \text{Branch MW capacity}$.
- b. Loss factor of seg 1, $m1 = Rpu * \text{Branch MW capacity} * 0.75 * (\text{coeff } A)$
- c. Loss in segment 1, $L1 = F1 * m1$

Loss segment 2

- d. Seg 2 folw, $F2 = (1 - \text{coff } A) * \text{Branch MW capacity}$.
- e. Loss factor of seg 2, $m2 = Rpu * \text{Branch MW capacity}$
- f. Loss in segment 2, $L2 = F2 * m2$

Loss segment 3

- g. Seg 3 folw, $F3 = 10000$
- h. Loss factor of seg 3, $m3 = Rpu * \text{Branch MW capacity} * (2 - 0.75 * (A))$
- i. Loss in segment 3, $L3 = F3 * m3$.

2.2 Flow and Loss models in the ISOs in the US

2.2.1 Flow model

The transmission branch line flow using PTDFs is given by (2.5)

$$(2.5) \quad T_l = \left[\sum_j (PTDF_{l,j}) . PG_j - \sum_k (PTDF_{l,k}) . PD_k \right]; l = 1, 2, 3 \dots L$$

$j \in J$ Set of all Generators

$k \in K$ Set of all Loads

$PTDF_{i,j}$ shows the sensitivity of line flow over injection(j)/withdrawal(k)

In this model, flow is determined using power transfer distribution factor (PTDF).

2.2.2 Loss Model

Market Wide Energy Balance Equation is shown in (2.6), as presented in the California ISO Market Optimisation Details, (2009) and Chakrabarti, Edwards, Callaghan & Ranatunga (2011).

$$(2.6) \quad \sum_{unit \in G} En_{unit}^t - \sum_{load \in L} En_{load}^t = En_{req}^t + En_{loss}^t; t \in T$$

$$(2.7) \quad En_{loss}^t = En_{loss}^{base,t} + \Delta En_{loss}^t; t \in T$$

$$(2.8) \quad \Delta En_{loss}^t = \sum_{unit \in G} \alpha_{node}^t \cdot (En_{unit}^t - En_{unit}^{base,t}) - \sum_{load \in L} \alpha_{node}^t \cdot (En_{load}^t - En_{load}^{base,t})$$

Single market wide energy balance is shown in (2.6).

En_{loss}^t Total system loss at time t

$En_{loss}^{base,t}$ System Loss in the base case at time t

ΔEn_{loss}^t Incremental system loss at time t

α_{node}^t Incremental loss sensitivity

2.2.3 Incremental loss sensitivity

Following DC method and using matrix form, we get

$$[\theta] = [B^{-1}] \cdot [P] = [X] \cdot [P]$$

where

$$(2.9) \quad [X] = [B^{-1}]$$

Thus the θ vector can be obtained by multiplying X-bus matrix with the power injection vector. The B matrix should be tailored suitably considering slack bus, before the B-matrix is inverted or factored into LDU matrices. This gives the X matrix.

$$(2.10) \quad \theta_{ij} = \theta_i - \theta_j = \sum_{l \neq slack} X_{i,l} \cdot P_l - \sum_{l \neq slack} X_{j,l} \cdot P_l$$

Thus total transmission loss becomes,

$$(2.11) \quad P^L = \frac{1}{2} \sum_{(i,j)} g_{ij} \left[\sum_{l \neq slack} X_{i,l} \cdot P_l - \sum_{l \neq slack} X_{j,l} \cdot P_l \right]^2$$

$$(2.12) \quad P^L = \frac{1}{2} \sum_{(i,j)} g_{ij} \left[\sum_{l \neq slack} X_{i,j,l} \cdot P_l \right]^2$$

where,

$$X_{i,j,l} = X_{i,l} - X_{j,l}$$

Incremental loss at node n,

$$(2.13) \quad \alpha_n = \frac{\partial P^L}{\partial P_n} = \sum_{(i,j \in L_{i,j})} g_{i,j} \theta_{ij} X_{i,j,n}$$

This is incremental system loss due to injection at node i . This says how much total system loss will change for an incremental change in injection at bus n , holding all other

injections and withdrawals constant, except the reference bus and bus n .

2.2.4 Branch Flows: A numerical case study

A 5-bus system is considered and flows in different branches are calculated using PTDFs and also by DC method.

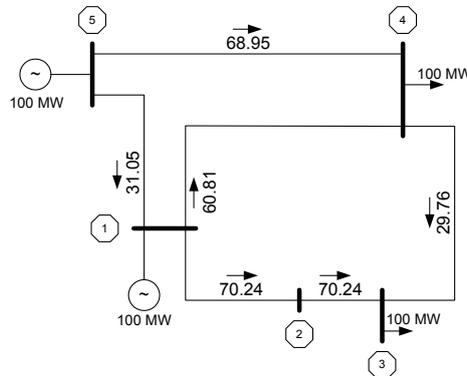


Fig. 2: Flows in the Studied network by LF solution

YBUS Matrix

	1	2	3	4	5
1	224.72	-35.58	0	-32.89	-156.25
2	-35.58	128.18	-92.6	0	0
3	0	-92.6	126.27	-33.67	0
4	-32.89	0	-33.67	100.23	-33.67
5	-156.25	0	0	-33.67	189.92

Table 1: YBUS matrix of the studied 5-bus system

Z bus matrix (Bus 1 is the reference bus)

0.0188	0.0153	0.0055	0.0010
0.0153	0.0211	0.0075	0.0013
0.0055	0.0075	0.0133	0.0024
0.0010	0.0013	0.0024	0.0057

2.2.5 Power Transfer Distribution Factor (PTDFs) or Shift Factors (SFs)

1. Get Y-bus matrix by inspection or using incidence Matrix(A); $YBUS = A^T yA$
2. Modify Y-bus matrix depending on slack bus. Call it B matrix
3. Get Z-bus from step 2: $Z = B^{-1}$
4. Define a vector of load and generation with +ve sign for generation and -ve sign for load. Call it "gsh" vector
5. Bus angles (radians) can be calculated as: $k = Z * gsh$
6. Flow can be obtained by: $F(i,j) = -B(i,j) * (k(i) - k(j))$.
7. Define perturbation at buses in a vector form, gsh (injection at one bus and withdrawals in other bus by small amount)
8. Recalculate the k- vector ($k = Z * gsh$)
9. Calculate the participation factor $PF(ij) = -B(i,j) * [k(i) - k(j)]$

PTDFs gives a measure of how much branch flow changes in a branch due to injection(S) of 1 MW and withdrawal(s) 1 MW at given buses. Net flow in a line can be calculated from the algebraic sum of appropriate (PTDFs*injection/withdrawals), as used in linear system. Column 2 of Table-2 shows the PTDFs of different lines for injection at bus 1 and withdrawal at 4 with bus 1 as the reference bus and so on. An

alternative algorithm is shown in the appendix.

2.2.6 DC Power Flow

In a loss less system, power flows in different lines are given by:

$$(2.14) [B] * [\theta] = [P]$$

$$[\theta] = [B]^{-1} * [P]$$

$$P_{i,j} = -B_{i,j} * (\theta_i - \theta_j)$$

In our problem, inserting the values B - matrix after removing 1st row and 1st column and $P = [0 \ -100 \ -100 \ 100]^T$, we get the flows which are shown in figure 2.

Line	Ref=1 G1=1MW, L4=1 MW	Ref=1 G1=1MW, L3=1 MW	Ref=1 G5=1MW, L4=1 MW	Ref=1 G5=1MW, L3=1 MW	G1=1, G5=1 L4=1 L3=1 MW	Col2+col5)*100= col6*100. Check
1-4	0.4376	0.2481	0.36	0.1706	0.6082	60.82
1-5	0.3685	0.209	-0.5195	-0.6791	-0.3106	-31.06
1-2	0.1939	0.1595	0.1595	0.5085	0.7024	70.24
5-4	0.3685	0.209	0.4805	0.3209	0.6894	68.94
3-4	0.1939	-0.4571	0.1595	-0.4915	-0.2976	29.76
2-3	0.1939	0.5429	0.1595	0.5085	0.7024	70.24

Table- 2: PTDFs for different lines

Flows in lines can be found out by algebraic sum of flows by different injection and withdrawal patterns. The 6th column flows are checked by introducing the injections and withdrawals in the “gsh”-vector and the result matched with the sum of column 1 and column 5 results. This in turn checked with the DC load flow equation and the result matched exactly with 6th column for 100 MW injection at bus 5 and withdrawal at buses 3 and 4 by 100 MW each; with bus 1 is slack. The detail is available in Chakrabarti, Edwards, Callaghan, & Ranatunga (2011).

3 Unique and non-unique LMPs

3.1 Definition of LMP

LMP is a cost of supplying an increment of load at a particular location. We can think of the LMP as a change of the total production cost to deliver additional increment of load to the location. The components of LMP are: Energy Component–Marginal generation price. Loss Component - is the marginal cost of additional losses caused by supplying an increment of load at the location. Congestion Component - equal zero for all locations if there are no binding constraints.

1. Unique LMP with fixed and stepped demand curves.
2. Supply (S) and Demand (D) curves do not intersect, D is above S: Existence of clearing price depends on market rule.

3.2 Unique LMP with fixed and stepped demand curves

Interpretation of cleared MW and price is straight forward for case 3a and 3b. Clearing price in the case shown in 3c in not very clear. It will depend on the market rules.

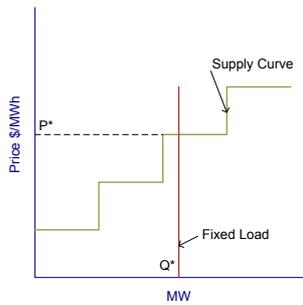


Fig 3a: Constant Demand

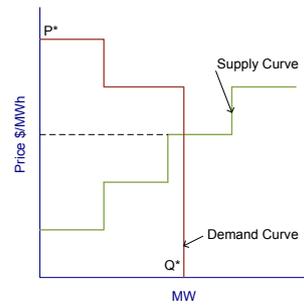


Fig 3b: Stepped Demand

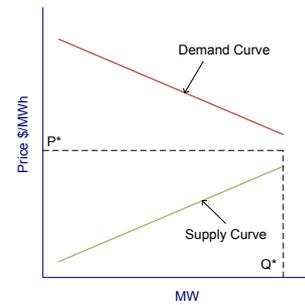


Fig 3c: S and D does not intersect

3.3 Primal and Dual Multiple Solutions

1. Non-unique LMPs where supply and demand curves intersect at multiple locations (horizontally). This is called primal multiple solutions.
2. Non-unique LMPs where supply and demand curves intersect at multiple locations (vertically). This is called multiple dual solutions.

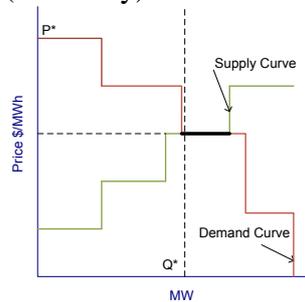


Fig 4: Multiple primal solutions

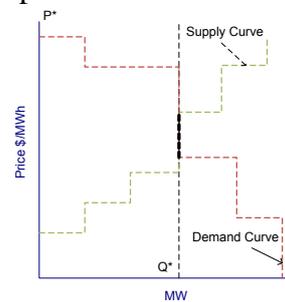


Fig 5: Multiple dual solutions

Fig. 4 shows multiple primal solutions and more than one participant tends to set the price. Fig. 5 shows the multiple dual solutions and different prices could be viewed, for example highest price can be seen when we look from the top of the curve and the lowest price when we look from the bottom of the curve. In this case dispatch is unique but the clearing price is not unique. Based on market rules the clearing price could be the highest accepted generation offer or the mid-point price.

3.4 Numerical Example illustrating multiple dual solutions

A three bus system is considered in Fig-6 to illustrate the multiple dual solutions (MDS). The objective is to minimise the total cost of dispatch (maximise the total surplus) with subject to flow constraints and generator and load limit constraints. The PTDFs are used to express the flows in different lines.

This section is a review based on a paper written by Feng, Xu, Zhong, & Ostergaard (2012). Thus major contribution rests on the authors of the paper. The problem in the above paper is rerun and examined and explained the findings of the authors.

The PTDFs (also called $T_{lG(i)}$ or $T_{lD(i)}$), with bus 3 as the reference bus, for load at bus 1 alone, and generation at bus 2 alone are calculated independently. These are shown below.

	PTDF 1-2	PTDF 1-3	PTDF 2-3
Dem=1MW at bus1	-0.333	-0.667	-0.333
Gen=1 MW at bus2	-0.333	+0.333	+0.667

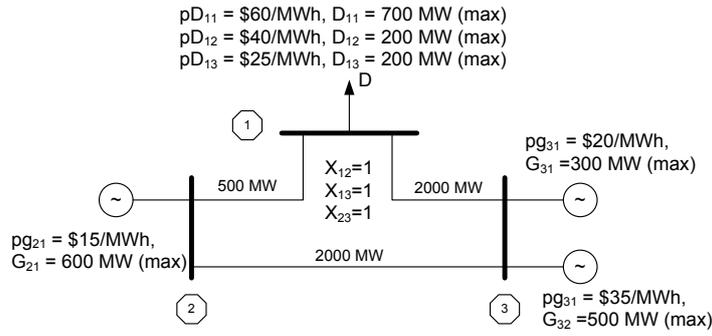


Figure-6: 3-bus network illustrating MDS

3.5 Optimisation problem (ISO Type)

The objective	Min. $Z = (15P_{g21} + 20P_{g31} + 35P_{g32}) - (60PD11 + 40PD12 + 25PD13)$
Total Power balance	$(P_{g21} + P_{g31} + P_{g32}) - (P_{d11} + P_{d12} + P_{d13}) = 0 \quad : \lambda$
Power Flow Constraints	$L2-1: 1/3(P_{d11} + P_{d12} + P_{d13}) + 1/3P_{g21} \leq 500; \mu_1$ $L3-1: 2/3(P_{d11} + P_{d12} + P_{d13}) - 1/3P_{g21} \leq 2000; \mu_2$ $L2-3: -1/3(P_{d11} + P_{d12} + P_{d13}) + 2/3P_{g21} \leq 2000; \mu_3$
Generation capacity constraints:	$P_{g21} \leq 600, \tau_1^+; P_{g21} \geq 0, \tau_1^-; P_{g31} \leq 300, \tau_2^+; P_{g31} \geq 0, \tau_2^-;$ $P_{g32} \leq 500, \tau_3^+; P_{g32} \geq 0, \tau_3^-;$
Demand specified at a bus:	$P_{d11} \leq 700, \beta_1^+; P_{d11} \geq 0, \beta_1^-; P_{d12} \leq 200, \beta_2^+; P_{d12} \geq 0, \beta_2^-;$ $P_{d13} \leq 200, \beta_3^+; P_{d13} \geq 0, \beta_3^-;$
Ref bus angle	$a_3 = 0 \quad : \pi_3;$

The KKT condition of optimality for generators and demands are shown in (3.1 and 3.2)

For generator i.

$$(3.1) \quad p_{gi} + \lambda + \sum_{l=1}^L \mu_l T_{lf(i)} + \tau_i^+ - \tau_i^- = 0; i \in \text{generators}; p_{gi} = \text{Gen}(i) \text{ offer price}$$

For demand j

$$(3.2) \quad -p_{dj} - \lambda - \sum_{l=1}^L \mu_l T_{lf(i)} + \beta_i^+ - \beta_i^- = 0; j \in \text{Loads}; p_{dj} = \text{Dem}(j) \text{ offer price}$$

The problem is solved by Linear programming “cplex” solver called from the GAMS program. Optimal Solution is shown in (3.3).

$$(3.3) \quad \lambda = 25$$

$$P_{d11} = 700MW, P_{d12} = 200MW, P_{d13} = 0; \beta_1^+ = -\$35, \beta_2^+ = -\$15, \beta_3^+ = \beta_3^- = 0, PG_{21} = 600MW$$

$$PG_{31} = 200MW, PG_{32} = 0MW; (G21)\tau_1^+ = -\$10, (G31)\tau_2^+ = \$5, (G32)\tau_3^+ = \tau_3^- = 0$$

It can be seen that the values of some multipliers in the solution are not correct, for example the value of μ_1 is shown zero by the solver even though the line 1-2 is at limit

of 500 MW ($1/3*900+1/3*600$). Optimal dispatch solution is correct. But the prices are difficult to calculate from the values of the multipliers. Also it can be noticed that all generators and loads are at limits, there is no marginal generator in this case. The following equations are developed using the non-zero multipliers as shown in (3.4).

$$\begin{aligned}
 &G21: 15 + \lambda + \frac{1}{3}\mu_1 = 0; G31: 20 + \lambda + \tau_2^+ = 0; G32: 35 + \lambda - \tau_3^- = 0 \\
 (3.4) \quad &d11: -60 - \lambda + \frac{1}{3}\mu_1 + \beta_1^+ = 0; d12: -40 - \lambda + \frac{1}{3}\mu_1 + \beta_2^+ = 0; \\
 &d13: -25 - \lambda + \frac{1}{3}\mu_1 - \beta_3^- = 0
 \end{aligned}$$

There are 7 variables to be found out but there are only 6 equations. We need one additional equation to solve the problem of finding the values of multipliers. The following boundaries as shown in (3.5- 3.8) are evaluated.

$$(3.5) \text{ Gen 32 is bounded at the LL, } S_h = \min\{p_{gi} + \sum_{l=1}^L \mu_l T_{f_{g(i)}}, i \in LL \text{ Gen}\} = 35$$

$$(3.6) \text{ Gen 21 is bounded at the UL, } S_l = \max\{p_{gi} + \sum_{l=1}^L \mu_l T_{f_{g(i)}}, i \in UL \text{ Gen}\} = 15 + \frac{1}{3}\mu_1$$

$$(3.7) \text{ Load 11 is bounded at the UL, } L_h = \min\{p_{dj} + \sum_{l=1}^L \mu_l T_{f_{D(j)}}, j \in UL \text{ Load}\} = 40 - \frac{1}{3}\mu_1$$

$$(3.8) \text{ Load 13 is bounded at the LL, } L_l = \max\{p_{dj} + \sum_{l=1}^L \mu_l T_{f_{D(j)}}, j \in LL \text{ Load}\} = 25 - \frac{1}{3}\mu_1$$

The area of surplus bounded by (a+b+c).Q* is the allocatable surplus where the following (3.9) definitions apply, referring the fig. 7,

$$(3.9) \quad a = L_h - \min(L_h, S_h); b = \max(L_l, S_l) - S_l; c = \min(L_h, S_h) - \max(L_l, S_l); d = S_l \geq 0$$

In our example,

$$(3.10) \quad a = 5 - \frac{1}{3}\mu_1; b = 20 - \frac{1}{3}\mu_1; c = 0; d = 15 + \frac{1}{3}\mu_1$$

We now write additional equation (3.11) for the reference bus.

$$(3.11) \quad \rho = -\lambda = d + \frac{b+c}{a+2b+c} \cdot (a+b+c)$$

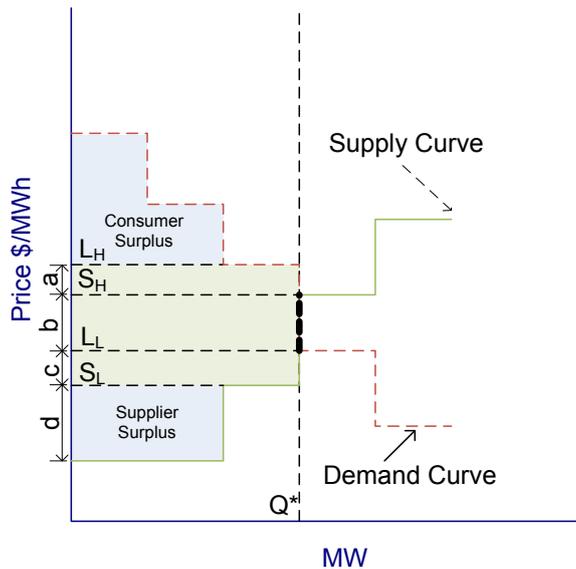


Fig 7: Dual multiple solutions: proposed Solution

Using (3.10 and 3.11), we get our 7th equation (3.12) as follows.

$$(3.12) \quad -\lambda = 15 + \frac{1}{3} \mu_1 + \frac{1}{9} (60 - \mu_1) \cdot \frac{75 - 2\mu_1}{45 - \mu_1}$$

Solving, we get

$$(3.13) \quad [\lambda, \tau_2^+, \tau_3^-, \beta_1^+, \beta_2^+, \beta_1^-, \beta_2^-, \mu_1] = [-27.5, 7.5, 7.5, 20, 0, 15, 37.5]$$

Nodal market clearing price at any bus $f(i)$ is uniquely determined by:

$$(3.14) \quad \rho_{f(i)} = -\lambda - \sum_{l=1}^L \mu_l T_{lf(i)}$$

Using these values, we get the bus prices using (3.14) and these are:

Bus1=\$40/MWh, Bus 2=\$15/MWh and bus 3 = \$27.5/MWh.

4 Summary

Branch power flows and loss models used in the market clearing engines in the NZEM and the ISOs in the US are reviewed. It is seen that both loss models and flow models are equivalent though the processing is different. LMP in both markets captures its three components. Attempt is made to explain the unique and non-unique prices that arise from time to time in different markets. Non-uniqueness in price resulted from multiple dual solution and its resolution is explained in detail. More research is required for further understanding on this issue.

5 Disclaimer

The views expressed in this paper do not reflect the views of the System Operator at Transpower or Victoria University, Wellington.

6 References

- Alvey, T., D. Goodwin, X. Ma, D. Streiffert, & D. Sun. 1998 "A security-constrained bid-clearing system for the New Zealand wholesale electricity market." *IEEE Transaction on Power System*, vol. 13, no. 2:340-346.
- Chakrabarti, B. B., C. Edwards, C. Callaghan, and S. Ranatunga. 2011 "Alternative Loss Model for the New Zealand Electricity Market using SFT". *proc. IEEE General Meeting, Detroit*.
- Wood, A. J., B. F. Wollenberg. 1996. *Power Generation, Operation and Control*, 2nd edition, John Wiley & Sons, Inc., New York.
- D.Feng, Z. Xu, Jin Zhong, Jacob Ostergaard. 2012. "Spot Pricing When Lagrange Multipliers are Not Unique". *IEEE Transaction on Power System*, vol 27, no1 :314-322.
- California ISO Market Optimisation Details. 2009. Technical Bulletin 2009-06-05. www.caiso.com

Acknowledgements

We sincerely thank Mr Kieran Devine, and Mr Doug Goodwin of Transpower NZ Ltd for their interest in our research and providing with engineering and technical support. We also thank Mr Tony Neighbours of Transpower for extending editorial support.

7 Appendix: Alternative Method of PTDF Calculation

Branches, m=6; Buses, n=5, Reference bus, R=1

% AA is a full Incidence Matrix [Element X bus]6X5

% Primitive Admittance matrix Ypre is given [1/x(ii)], 6X6

YBUS= AA'* ypre*AA, 5X5

% Bus 1 is reference, turn this row as unit vector. YBUS is now 5X5

Z=inv(YBUS); 5X5; SF=ypre*AA*Z, 6X5

Limits to Collective Action – Development of an Evolutionary Game Model

John R. Cody, Robert Y. Cavana & David G. Pearson
School of Social & Cultural Studies and Victoria Business School,
Victoria University of Wellington, New Zealand
jrcody@xtra.co.nz

Abstract

The presentation outlines the development of a theoretical model that compares the relative fitness of strategies of contribution to, and defection from contributing to, a collective good.

Three stages of development are described: Hirshleifer and Martinez Coll's (1988) hypothesis related to limits of reciprocity; Heckathorn's (1996; 1998) radical extension of the model to include sanctioning and ideology; and, Ziegler's (1997) approach to simulating stratified social systems.

The current project uses a VENSIM (Ventana Systems 2010) version of the model. VENSIM facilitates the use of two features of System Dynamics modelling: experiments with feedback loops and formal validation of the model.

The conceptual core of the model is Heckathorn's proposition that there are five logical combinations of payoff in the core two-by-two matrix of an evolutionary game. In addition to the much analysed Prisoner's Dilemma ($T > R > P > S$) changes in relative costs and constraints can create Assurance, Chicken, Privileged and Altruist's Dilemma games. Each game can be rationalised using a distinctive ideology to explain the balance between individual and collective interests. The sequence of games in a scenario is determined by the shape of a production function and the relative value of the collective good.

The model is being applied to problems related to persistent population health disparities.

Key words: system dynamics, evolutionary games, collective action.

References

- Heckathorn, D.D. 1996. "The dynamics and dilemmas of collective action." *American Sociological Review* **61**: 250-277.
- Heckathorn, D.D. 1998. "Collective action, social dilemmas and ideology." *Rationality and Society* **10**: 451-479.
- Hirshleifer, J., J.C. Martinez Coll. 1988. "What strategies can support the evolutionary emergence of cooperation?" *Journal of Conflict Resolution* **32**: 367-98.
- Ventana Systems. 2010. The Ventana Simulation Environment: Vensim Professional for Windows.
- Ziegler, R. 1997. "The normative structure of solidarity and inequality." *Rationality and Society* **9**: 449-467.

Districting for the New Zealand Census: MIP-Heuristic Approaches

Matthew Crowder
Department of Engineering Science
University of Auckland
New Zealand
mcro114@aucklanduni.ac.nz

Andrew Mason
Department of Engineering Science
University of Auckland
New Zealand
a.mason@auckland.ac.nz

Abstract

The census districting problem is one of several closely related problems where an organisation wishes to partition a region into contiguous and compact districts with roughly equal populations. This problem is extremely difficult to solve if formulated as an integer programming problem but heuristic algorithms are able to find good but not optimal solutions in a reasonable amount of time. MIP (Mixed Integer Programming) solvers have advanced in recent years to a point where given an appropriate model, they can be used to explore a large neighbourhood of potential solutions very quickly and can thus be used in heuristic algorithms or MIP-Heuristics. We successfully developed a new method for finding good solutions to this problem using dynamic programming and by devising a large neighbourhood around an incumbent solution.

Key words: partitioning, districting, census, MIP-Heuristics

1 Introduction

The 2013 New Zealand census will require everybody in New Zealand, an estimated 4.6 million people, to fill out census forms. Statistics New Zealand recruits a temporary workforce of approximately 7,500 people to deliver these forms.

The census districting problem is one of several closely related problems where an organisation wishes to partition an area or region into contiguous and compact districts with roughly equal populations. Statistics New Zealand therefore has the problem of devising some method to geographically apportion New Zealand among the census collectors. The workload assigned to each census collector must be practicable for the time available. In addition, Statistics New Zealand requires the area assigned to each worker to be compact and contiguous. These areas are called sub-districts. New Zealand is split into around 45,000 discrete geographic areas called meshblocks, which are combined to form the required sub-districts.

The districting problem we wish to consider can be modelled as a graph partitioning problem (Mehrotra, Johnson, and Nemhauser 2012) by associating a node with meshblocks contained in a region and connecting nodes by an edge whenever

two meshblocks border each other. The weight on a node is equal to the workload of the corresponding meshblock. The arc lengths represent the distance between two meshblocks. The safety of census collectors, very important to Statistics New Zealand, is improved by removing any arc that crosses a dangerous boundary like rivers and major highways. A feasible solution to the above problem is a partitioning of the nodes in the graph such that the nodes in any set of the partition induce a connected subgraph (to ensure geographic contiguity of the sub-districts) and the sum of the node weights lies within a specified range (to satisfy the workload requirements).

2 Formulation

2.1 Graph Representation

Let $G(V, E)$ be a graph where the set of all nodes V represent all meshblocks contained within the geographic border of the region we are partitioning and the edges E are the pairs of all meshblocks that share a common border. Let the cost of the arcs $c_{ij} \in \mathbb{R}$ be the distance between two meshblocks $(i, j) \in E$ and the weight of each node $w_i \in \mathbb{Z}$ the workload of the meshblock $i \in V$.

A subdistrict of G can be any set of nodes $S \subseteq V$. A subdistrict is only feasible if it is contiguous. Stated mathematically, the sub-graph $G(S, \{(i, j) \in E \mid i \in S, j \in S\})$ must be connected. Furthermore, S must be weight feasible, that is $W_{min} \leq \sum_{i \in S} w_i \leq W_{max}$ for the workload limits (W_{min}, W_{max}) defined by Statistics New Zealand.

The cost C_S of a sub district S if used in a partitioning of G , is defined as the sum of the arcs lengths in the minimum spanning tree (denoted $T(S, E')$) formed over the nodes S and arcs E' . Equivalently $C_S = \sum_{(i,j) \in E'} c_{ij}$

2.2 Set Partitioning Problem

A sub-district has two representations, as a tree that is a subgraph of G and as a column of the set partitioning model below. The optimal partitioning of G into a non-overlapping set of trees (or sub-districts) from a set of trees \mathcal{T} can be formulated as the following integer programme:

Let a_{ij} equal 1 if a node $i \in V$ is in the tree $j \in \mathcal{T}$, and 0 otherwise and let x_j equal 1 when tree $j \in \mathcal{T}$ with cost C_j is in a partitioning of G and 0 otherwise. The census districting problem can then be stated as the following set partitioning problem, where every tree $j \in \mathcal{T}$ is weight feasible

$$\begin{aligned} \min \quad & \sum_{j \in \mathcal{T}} C_j x_j \\ \text{s.t} \quad & \sum_{j \in \mathcal{T}} a_{ij} x_j = 1 \quad \forall i \in V \\ & x_j \in \{0, 1\} \quad \forall j \in \mathcal{T} \end{aligned}$$

A feasible solution to the above constraints will be a set of trees $\mathcal{P} = \{j \mid x_j = 1, j \in \mathcal{T}\}$ that partition G i.e. $\cup_{p \in \mathcal{P}} S_p = V$, $\cap_{p \in \mathcal{P}} S_p = \emptyset$ and the optimal partition \mathcal{P}^* partitions G into trees that form a minimum cost spanning forest.

3 MIP-heuristic Algorithms

The census districting problem is intractable for problems with a large number of nodes. However, heuristic algorithms can be used to find good, but not necessarily optimal solutions. The power of heuristic algorithms lies in the speed at which they can explore potential solutions. MIP solvers have advanced in recent years to a point where given an appropriate model, they can be used to explore a large neighbourhood of solutions very quickly. Hence they can be used as part of a heuristic algorithm or MIP-Heuristic.

Raffensperger (2008) found good solutions to the census districting problem by solving a set partitioning problem to explore possible partitions of a graph. The columns in his set partitioning model were generated using a variant of Kruskal’s algorithm and subgradient optimization. It may be possible to find solutions of even better quality by designing a MIP-Heuristic algorithm that repeatedly solves the set partitioning problem for the purpose of generating new columns similar to the columns in an incumbent solution.

Figure 1 shows the general set of steps within our MIP-heuristic approach. First, some method is used to generate a large number of weight-feasible trees. Second, these trees are added as columns to the set partitioning problem. Third, a MIP solver finds the optimal solution to the set partitioning problem. Finally, if the MIP solver finds an improved solution, some method is used to generate trees in the neighbourhood of this solution otherwise we generate more trees as in the initial step. This process is then repeated until some stopping criterion is met.

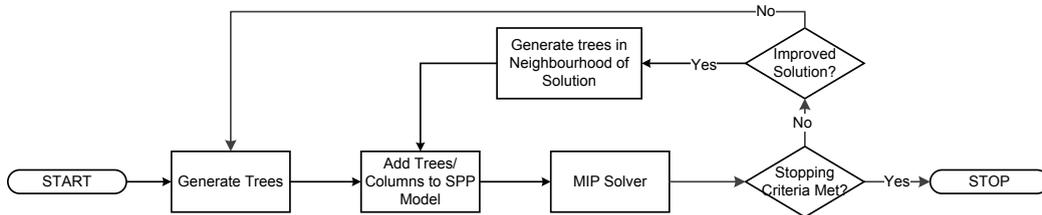


Figure 1: A generalised MIP-Heuristic algorithm

4 Tree Generation using Dynamic Programming

Dynamic programming is an algorithmic scheme for optimization that solves a complex problem by breaking it up into subproblems that are easier to solve. Dynamic programming algorithms are very fast because each subproblem is solved only once and it is stored by the algorithm so that it can be used when solving a subsequent subproblem. Dynamic programming cannot be easily used to find the optimal partitioning of an arbitrary graph because there is no way to break up the problem so that the optimal partition of one set of nodes can be combined with another and still be optimal. However, this is not true in the case of a tree. Furthermore, a feasible partition of a spanning tree for the full graph is also feasible for the full graph itself. We can therefore generate many initial columns for the set partitioning problem by finding the optimal partition to a number of different spanning trees of the full graph.

Figure 2 shows an example spanning tree rooted at node v with two child nodes a, b . Alongside each node i is its respective weight w_i . We can see that for $W_{max} = 9$ and $W_{min} = 6$ there is only one feasible way to partition these nodes. However for

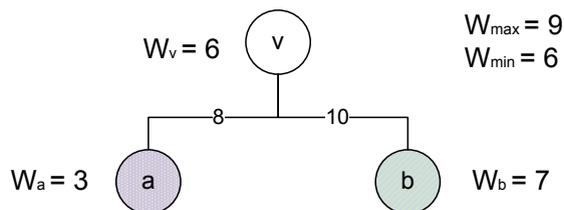


Figure 2: Example tree

larger problems finding the least cost way of partitioning a tree into weight feasible sub-trees is much harder. A dynamic programming algorithm solves this by breaking up the problem into smaller sub-problems that are easier to solve.

4.1 Constructing Random Spanning Trees

Our aim in partitioning spanning trees is not to find the best partition but to produce as many good initial columns for the set partitioning problem as practicable for a given problem. To achieve this the construction of spanning trees for input to the dynamic programming algorithm is randomised. Moreover, to avoid spanning trees that give poor solutions we limit random trees to those having a total length close to that of the minimum spanning tree. To generate these random trees we used a randomised version of Kruskal’s Algorithm (Kruskal 1956) that chooses arcs using a GRASP (Greedy Randomised Adaptive Search Procedure) algorithm (Feo and Resende 1995). In this algorithm the step in Kruskal’s Algorithm of choosing the shortest arc that does not form a loop is replaced by instead choosing randomly from a restricted candidate list of arcs (the parameter RCL gives the size of this list). The restricted candidate list is comprised of the RCL shortest arcs not yet in the tree. Any arcs that form a loop are removed from this list and replaced by the next shortest arc.

5 Columnwise Neighbourhood Search

The number of potential sub-districts that exist for most census districting problems is massive. Consequently, brute force exhaustive enumeration of all potential columns is not possible. However, one might have more success in obtaining good districting solutions, by being strategic with what columns are added to the set partitioning model. One way of doing this is to define a neighbourhood of potential columns (or trees) around the current optimal solution and adding all these columns to the set partitioning model which can be then resolved. Doing this repeatedly can thus be classified as a steepest descent heuristic.

A tree has two representations firstly as a column in the SPP model $t \in \{0, 1\}^n$ and secondly as a subset of the nodes of G . To limit the number of columns added to the SPP model a reasonable sized neighbourhood containing good quality trees is desired. The classical way of defining a neighbourhood around a solution (or tree in our case) is to limit our attention to a fixed sphere around the solution. The distance between t and the modified tree t' is defined to be $|t, t'| = \sum_{j=1}^n |t_j - t'_j|$. Thus we can define a neighbourhood $\mathcal{N}^k(t)$ to be the search space around the tree t where $|t, t'| \leq k$, which is equivalent to the set of all trees reachable from t in one “k change” move. As the elements of t are binary a move from t to t' can

be thought of as flipping one or more elements of t to get t' . A k change move $t' = k_change(t, \{i_1, i_2, \dots, i_k\})$ defines a move from one value of t to another t' , where i_1, i_2, \dots, i_k are the elements of t to flip to obtain t' .

To make implementation easier we define three moves that can be used to explore the neighbourhood of a tree. The “add” move involves adding a nodes that border the tree. Conversely, the “remove” move removes nodes from the tree. A “swap” move performs an add move and a remove move sequentially (add k nodes and remove k nodes), this move was found to generate much more trees than the former two moves.

6 Implementation of Algorithms

The Python programming language was used in this project as it allows fast prototyping of algorithms; C programs of similar complexity require significantly more development time. The set partitioning problem is modelled and solved in Gurobi, Gurobi provides an API that is used to add and remove columns from the set partitioning model and return solutions (see Figure 1). All runs were performed on machines with an Intel Xeon (4 x 2.6 GHz) processor and 6GB of RAM. The Python module NetworkX’s graph implementation was used to create and manipulate graphs (nodes and arcs). The data structure of the graphs are implemented using Python dictionary data structures. This “dictionary-of-dictionary” structure allows for the fast manipulate of nodes and neighbours in large graphs (Hagberg, Schult, and Swart 2012).

7 Results

We wished to compare the quality of the solutions produced by our DP algorithm to that of the approach taken by Raffensperger (2008). We were provided with a list of trees generated using his algorithm. To provide a direct comparison we generated the exact same number of trees using the DP-MIP algorithm. The trees from the different algorithms were added to separate set partitioning models and solved using Gurobi. Table 1 lists the IP/LP objectives found by solving the set partitioning model and the number of columns in each set partitioning problem. Using the DP to generate the trees resulted in an objective function value mostly on par for the three larger problems and gave a better objective for the smaller problems. The Python code of the DP algorithm took a very long time to run; whereas in his paper Raffensperger reports a time to solve similarly sized problems as between 2 and 10 minutes. The IP solve times for each of these problems were insignificant in comparison to the time needed to generate the trees (less than one percent of the total running time).

To ensure that the computational effort required in partitioning trees using the DP algorithm is well spent, we compared it to simply generating many random trees. Random trees are generated by growing a tree from a randomly selected node and picking a random arc connected to any currently selected node. We then solved a minimal spanning tree over the resulting random nodes to calculate its cost. We generated columns for the SPP model in batches of 10,000 trees, solving the set partitioning problem after each batch was finished. Figure 3 shows the typical performance (in this case for the 1000 node problem) of both the DP-MIP algorithm

Problem size	# of Trees Generated	Time (mins) to generate trees	DP-MIP Hybrid Algorithm		Raffensperger's Method	
			IP	LP	IP	LP
30	2,556	0.16	7.400×10^2	7.400×10^2	8.955×10^2	8.955×10^2
100	7,619	0.43	5.841×10^3	5.820×10^3	5.851×10^3	5.836×10^3
1000	76,378	31.00	1.695×10^6	1.687×10^6	1.696×10^6	1.689×10^6
5000	385,782	321.46	1.881×10^7	1.875×10^7	1.877×10^7	1.872×10^7

Table 1: Solution quality of SPP when given columns generated by the DP and John Raffensperger

and the purely random tree generation algorithm. It can be observed that the DP-MIP algorithm performs significantly better. It was very difficult for Gurobi to find a feasible solution when the set partitioning model contained a small number of columns/trees (10,000) that were constructed randomly; it could not find a feasible solution even after running for 12 hours. However, when we generated 1 million random trees the MIP solver was able to find a solution to the problem in around 30 minutes. Contrasting this, after using the DP Algorithm to generate 10,000 columns Gurobi was able to solve the problem in 0.78 seconds.

Figure 4 shows the performance of the CNS(Columnwise Neighbourhood Search) algorithm with different neighbourhood sizes for the 1000 node problem. We can see that the smaller neighbourhoods ($k=1$, $k=2$) perform better than the larger neighbourhoods. It was found that the CNS approach did not improve the quality of the solutions found using the DP-MIP algorithm. Moreover, the CNS approach generated less unique trees in the same time as the DP-MIP algorithm.

8 Conclusions

These results give good evidence that we have been successful in implementing a dynamic programming algorithm that partitions spanning trees to use as good initial trees for other MIP-Heuristic algorithms. Furthermore, the hybrid DP-MIP algorithm was the most effective method found as it reached the best solution overall in the fastest time. The GRASP approach for constructing spanning trees was a very important part of our DP-MIP algorithm.

Disappointingly, the Columnwise Neighbourhood Search algorithm did not find higher quality solutions than the solutions found using the DP-MIP algorithm. Additionally, it was found that the large neighbourhoods of trees surrounding the current best solutions were worse than smaller ones. One could conclude that expanding the neighbourhood too much is not a good strategy. We could consider more judicious selection criteria when adding columns such as the reduced cost of the variables in the LP.

Acknowledgments

I would like to acknowledge and extend my gratitude to my supervisor Dr Andrew Mason for his advice and help on this project and to Dr John F. Raffensperger for generously sharing his ideas, results and test data.

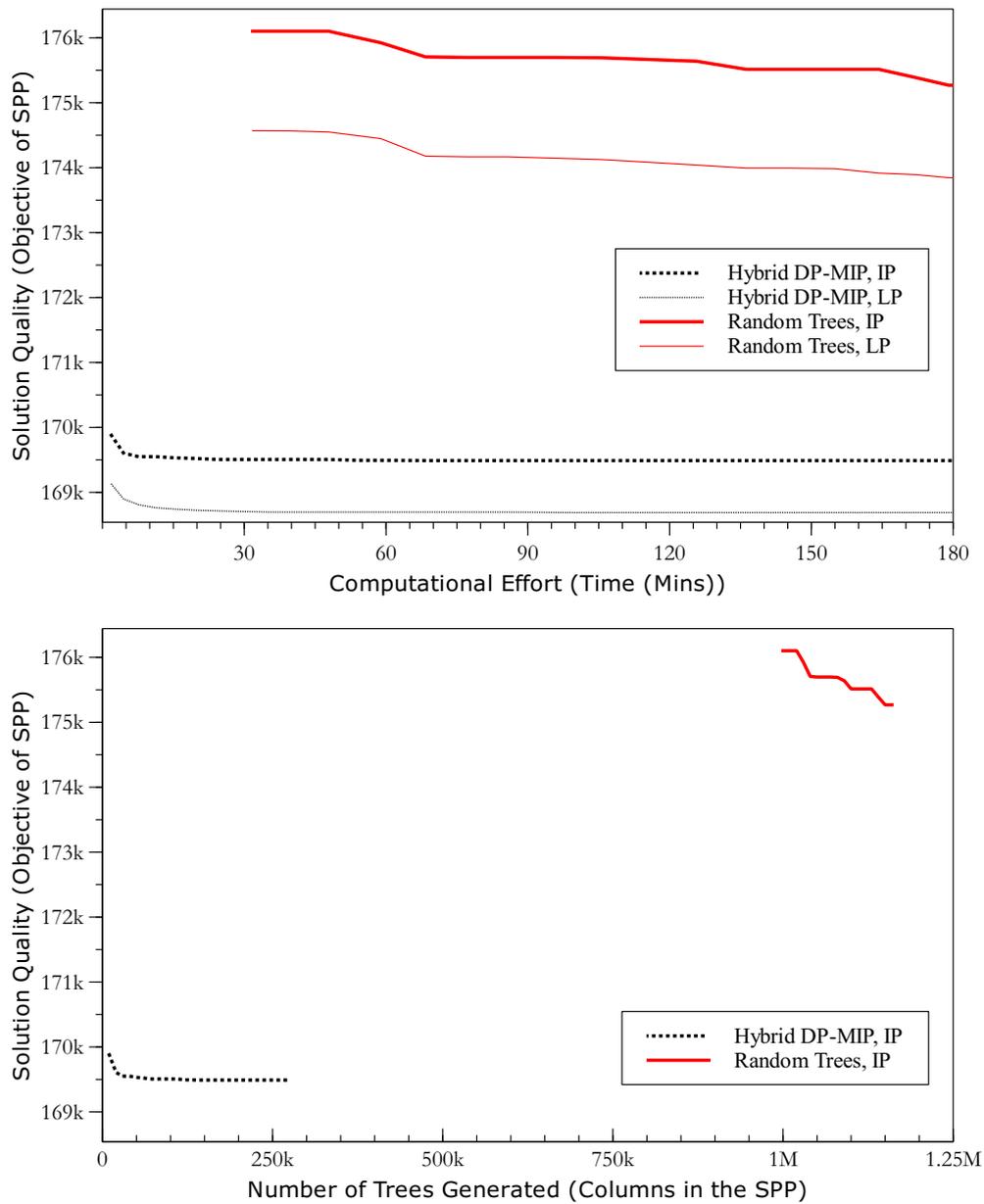


Figure 3: Plots of solution quality vs. computational effort in terms of the number of trees and time for the DP-MIP Algorithm and purely random tree generation for the 1000 node problem

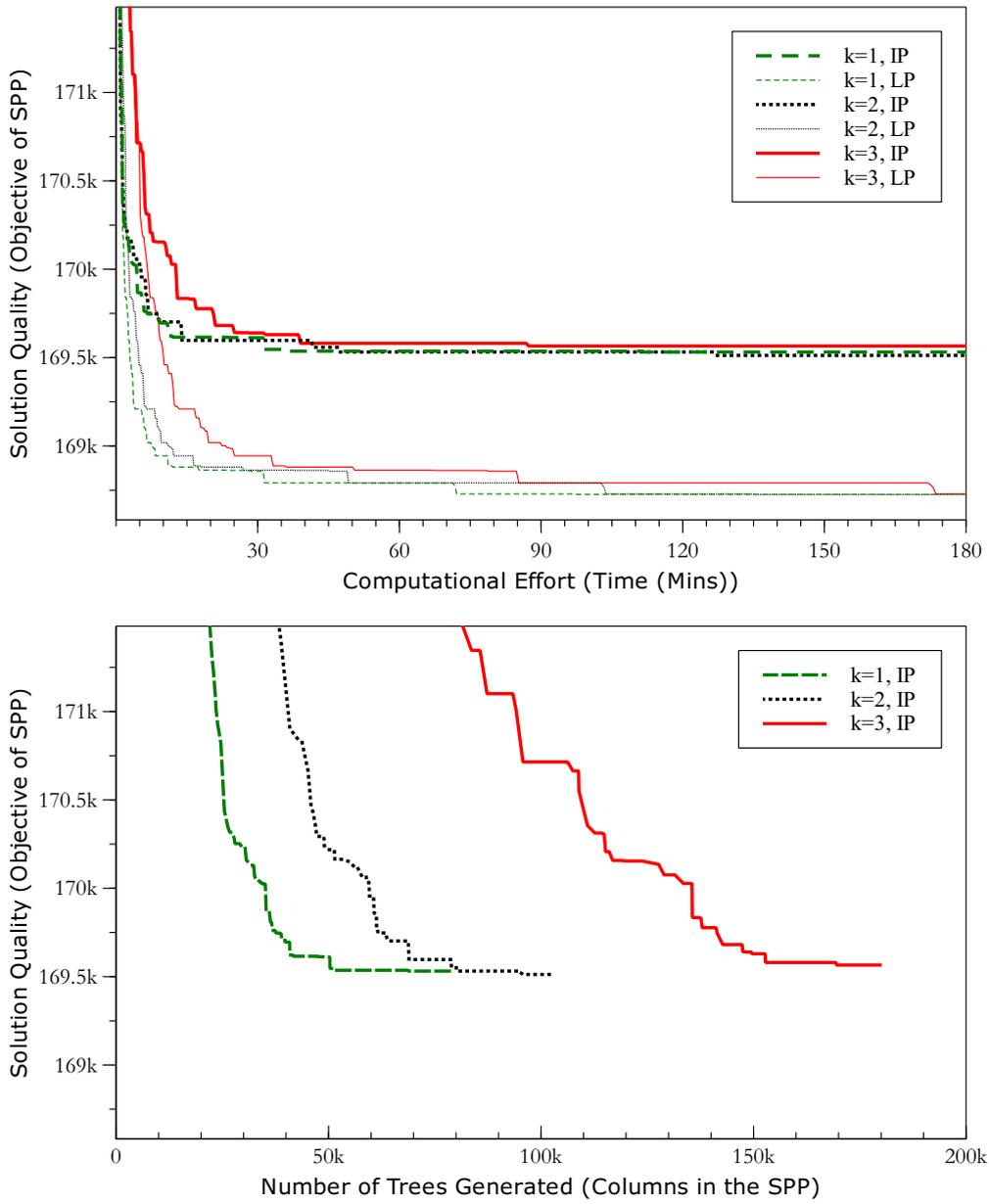


Figure 4: Performance of CNS algorithm with different neighbourhood sizes $k = 1, 2, 3$ for the 1000 node problem

References

- Feo, T.A., and M.G.C. Resende. 1995. “Greedy randomized adaptive search procedures.” *Journal of Global Optimization* 6 (2): 109–133.
- Hagberg, Aric, Dan Schult, and Pieter Swart. 2012. NetworkX Reference. http://www.networkx.lanl.gov/_downloads/networkx_reference.pdf.
- Kruskal, JB. 1956. “On the shortest spanning subtree of a graph and the traveling salesman problem.” *Proceedings of the American Mathematical Society* 7 (1): 48–50.
- Mehrotra, Anuj, Ellis L Johnson, and George L Nemhauser. 2012. “An optimization based heuristic for political districting.” *Management Science* 44 (8): 1100–1114.
- Raffensperger, John F. 2008. “A New Algorithm for the Collection Area Problem.” *Proceedings of the 43rd Annual Conference of the Operational Research Society of New Zealand*. 187–189.

Maximizing the Size of a Diamond, Cut from a Given Rough Stone

Anthony Downward, Yeong Fatt Thai, Golbon Zakeri
Department of Engineering Science, University of Auckland.

Abstract

In this work we consider the problem of fitting a single convex polyhedron inside another convex polyhedron, so as to maximize the utilized space. Specifically, we are thinking about this in the context of diamond cutting where you wish to maximize the value of a cut diamond by finding the optimal orientation and translation within a rough stone. This rough stone can be thought of as a convex polyhedron defining the feasible region for the cut diamond: all of the diamond vertices must lie inside this region. This problem is non-linear and non-convex, making it difficult to solve.

For the two-dimensional case, where we are dealing with polygons, there are a number of ways to approach this problem, for example enumeration techniques. However, to avoid the need to enumerate and evaluate all possible solutions, here we will consider a column generation method for this problem; this method uses a variant of the entering variable criterion from the revised simplex method to identify a set of diamond orientations, of which, one will be the global optimal solution.

We also consider a nonlinear programming formulation, for which we can find the global optimum solution by using piecewise approximations of a non-convex quadratic objective function.

In order to better understand this problem in three-dimensions we have developed a tool to visualize the three-dimensional fitting of a diamond within the rough stone. This software, written in C++, using OpenGL allows the user to understand the constraints imposed by the rough-stone on the cut diamond. The user is able to interact with the orientation of the diamond, and the software continuously resolves for the optimal magnification as the diamond is rotated. This is done, using hot starts, and a variant of the dual simplex method; however, since the \mathbf{A} matrix is changed due to rotation, we lose both primal and dual feasibility. Using this software one can conjecture heuristic techniques to improve the magnification of the diamond, which then can be implemented mathematically.

The Department of Engineering Science has been using this software to advertise the degree to prospective students on University open days. This is especially useful due to the accessibility of the software and the ability to relate the concepts visualised to standard linear programming theory. At the end of this talk we will demonstrate this software and present a heuristic technique that we have developed to find locally optimal solutions in this case.

Key words: non-linear, non-convex, optimization, diamond, shape fitting.

Use of Hydro Resources for Irrigation and Electricity Production

A. Downward, G. Zakeri, Z. Farishta, F. Wahid
Department of Engineering Science, University of Auckland.
g.zakeri@auckland.ac.nz

Abstract

This paper is primarily concerned with proper valuation of water. Water is utilized in multiple purposes such as electricity generation and farming. Using a central planning model of the New Zealand electricity market (NZEM), we can estimate the cost to the electricity sector associated with various levels of irrigation. We discuss such a model and how realistic it may be.

The recent transfer of Tekapo A and B from Meridian to Genesis has raised further questions surrounding hydro management and contracting. For example, when two separate firms operate on the same river chain, how does this affect the efficiency of the water usage? To reproduce the efficiency of the single-ownership structure, we investigate the use of water transfer prices.

1 Introduction

Fresh water is an increasingly scarce resource. Supply of water is uncertain and unevenly distributed through time and space. Thus, using water today will affect the amount of water available in the future. Typically, this is modelled by way of an opportunity cost which aims to reflect the marginal value of conserving water. Due to various sources of uncertainty, especially hydrological inflows, this opportunity cost can be difficult to estimate.

Water in New Zealand is used for irrigation purposes as well as hydro generation. Irrigation is classified as a consumptive use for water (i.e. fresh water is lost through irrigation), however hydro generation is not classified as a consumptive activity as the water is released to a down stream reservoir. New Zealand ranks fourth in the OECD 30 for the size of its renewable freshwater resources on a per capita basis (Ministry of Forestry and Environment 2010). However, shortages (or threats of future shortages) still occur at certain times of the year (especially during dry years), reflecting the high level of inflow variability. In absence of a water market where participants can signal the value (to themselves) of having water, it is imperative to compute the value of water to allocate it for the most efficient use. There are well developed models (e.g. SDDP and DOASA models) that compute the value of water, in the context of electricity generation, by efficiently scheduling hydro generation in the face of inflow uncertainty. SDDP and DOASA use decomposition and sampling

methods applied to very large scale stochastic programs. Such methods can be used as preliminary tools to assign a value to water. We will present a framework for this method and discuss some drawbacks associated to such a mechanism for determining water values in NZ.

On a related matter, one recent change in the electricity market structure which has the potential to reduce the efficient use of water, is the disaggregation of the Waitaki river-chain, where Tekapo A and B were sold to Genesis. Now there are two separate companies (Genesis and Meridian) both operating on one system. Since Tekapo A and B are the upstream reservoirs, a large part of Meridian's inflows are now controlled by Genesis. Later in this paper, we will demonstrate, by way of example that this disaggregated system will be less efficient than an otherwise identical integrated system. We finally discuss the use of water transfer prices to eliminate the efficiency loss.

2 Management of hydro resources

To determine the most efficient use of water during a dry year, it is worthwhile to estimate the cost (to the electricity sector) of allocating water for irrigation purposes. Using SDDP or similar models (e.g. DOASA, developed within EPOC (Philpott et al. 2010)) we are able to compute water value surfaces. An example for a single reservoir is shown in Figure 1 in the next section. This surface gives the value of water as a function of the volume of water stored in the reservoir. The gradient of the function is the marginal value of additional water, and this is what determines how a competitive generator ought to offer into the market. These water values however only pertain to central plans or a perfectly competitive market.

Below, we present an example illuminating the differences between how a central planner values water as compared to a profit maximizing firm with market power. The time horizon is over 5 periods, with the value of water in the final period set to 0. Each period will last one hour. In this example, we consider two entities: a hydro generator with a reservoir, and the *rest of the market* which we will approximate by linear offer (or cost) curve of slope 1. First, we will consider the most simple case where there are no inflows, with a fixed demand in each period of 200MW.

2.1 Perfect competition

Suppose W_t is the amount of water available (in MWh) for hydro generation at time t . Now in any time period you want to optimize your use of water so as to maximize the savings (in terms of fuel costs) in the current period plus any future savings.

In this example, the savings from the using water in period t are given by:

$$S_t(x_t) = \frac{1}{2} (200^2 - (200 - x_t)^2),$$

where x_t is the amount of hydro generation at time t and thus $200 - x_t$ is the amount of generation from the rest of the market (offering with a slope of 1). Therefore using a dynamic programming recursion we can formulate the central planner's problem as follows:

$$\mathcal{C}_t(W_t) := \max_{0 \leq x_t \leq W_t} \{S_t(x_t) + \mathcal{C}_{t+1}(W_t - x_t)\}, \quad t = \{1, \dots, T - 1\}.$$

In the final stage, T , the optimal decision is to use all the water in the reservoir up to the demand, thus

$$x_T^C(W_T) = \min \{200, W_T\},$$

giving

$$\mathcal{C}_T(W_T) = S_T(x_T^C(W_T)).$$

2.2 Monopolist behaviour

With a monopolist running the hydro generator, the incentives are very different. The monopolist wishes to maximize its total profit (price \times quantity), where the clearing price is determined by the marginal price of the most expensive dispatched generator). For any given period, t , the monopolist's profit, given a hydro dispatch of x_t can be found to be:

$$P_t(x_t) = x_t(200 - x_t).$$

The monopolist wishes to maximize these profits over time, so once again we use a dynamic programming recursion, this time to formulate the monopolist's problem:

$$\mathcal{M}_t(W_t) := \max_{0 \leq x_t \leq W_t} \{P_t(x_t) + \mathcal{M}_{t+1}(W_t - x_t)\}, \quad t = \{1, \dots, T-1\}.$$

In the final stage, T , the optimal decision is to use all the water in the reservoir up to the monopolist strategy (100MW), therefore:

$$x_T^M(W_T) = \min \{100, W_T\},$$

giving

$$\mathcal{M}_T(W_T) = P_T(x_T^M(W_T)).$$

2.3 Comparison

At the optimal solutions for the two scenarios, we find the following water value curves. Note that the water values are lower, under the monopolistic scenario,

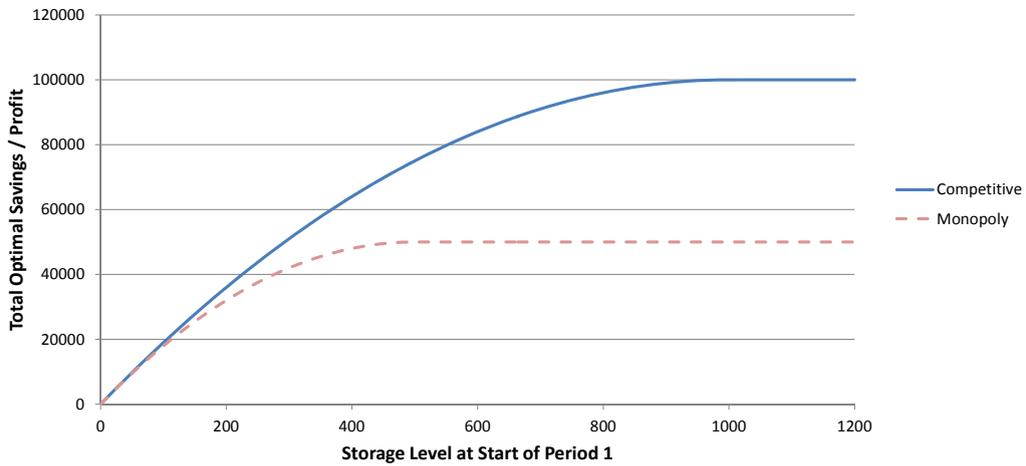


Figure 1: Comparison of water values.

however, the prices are higher. This is because the monopolist focuses only on its profits rather than the total welfare of the market.

Table 1: Comparison of savings and revenue.

	Competitive	Monopoly
Savings	\$97,750	\$75,000
Revenue	\$25,500	\$50,000

From these curves we can find the optimal hydro release plan in the central plan is to use 170MW in each period, whereas the monopolist uses 100MW. From this release plan, we can see that even though the monopolist values water less, it also uses less water, thereby maintaining higher prices. As shown in Table 1 below, we can see that the monopolist increases its revenue by a factor of two over the competitive plan, but in doing so increases the amount of thermal generation required (reducing savings by almost 25%). These differences show that one ought to be aware, when modelling the New Zealand electricity market, that assuming that the hydro resources are operated to maximize overall welfare may not reflect true behaviour in the market. However, computing the optimal hydro release policies for markets with imperfect competition is complicated by a lack of convexity which can lead to situations where no equilibrium exists. Thus we will assume a perfectly competitive environment for the remainder of this paper.

3 The cost of irrigation

In order to assess the impact of irrigation of the power generation sector, we assume that if irrigation allocations increase, hydro generators will anticipate this change and adapt their hydro generation policy to account for the expected decline in inflows. Each irrigation level will thus have a different policy associated with it.

For each discrete level of irrigation, historical tributary inflows along the Waikato River chain are modified to incorporate the impact of irrigation. These modified inflow sequences are then used within DOASA to compute an optimal water release policy for hydro generators.

3.1 Cost to a competitive electricity system

In this section, the DOASA National Model is used to explore the impact of irrigation in a hypothetical scenario of centrally planned electricity generation in New Zealand. The DOASA National Model is a multi-reservoir model that attempts to overcome the curse of dimensionality by making use of sampling to estimate future cost functions. The models objective is to minimize national cost of electricity generation.

The DOASA National Model formulates a simplified model of the New Zealand electricity generation system. The DOASA National model includes: 6 storage reservoirs with finite storage and generation capacity; 33 hydro stations; 12 thermal generation units with finite generation capacity and variable fuel cost; 3 demand nodes; 1 week stages (52 in a year); 3 demand load blocks per week (peak, shoulder and off-peak). For a full list of the assumptions and simplifications see (Farishta 2011).

In this paper we will present the results from the model under the assumption of perfect hindsight. Specifically we will simulate the optimal hydro management for 2008 and 2009, with known demand and inflow. We will test four discrete irrigation levels between 0 and 8 cumecs and estimate the costs to New Zealand under these assumptions. For further information about modelling this with uncertain inflows see (Farishta 2011).

For each year and irrigation level, we simulate the water release policy to estimate the thermal plant utilisation, and hence the overall costs; these are shown in figure 2. Thus by comparing different levels of irrigation, we can estimate the marginal costs associated with additional irrigation.

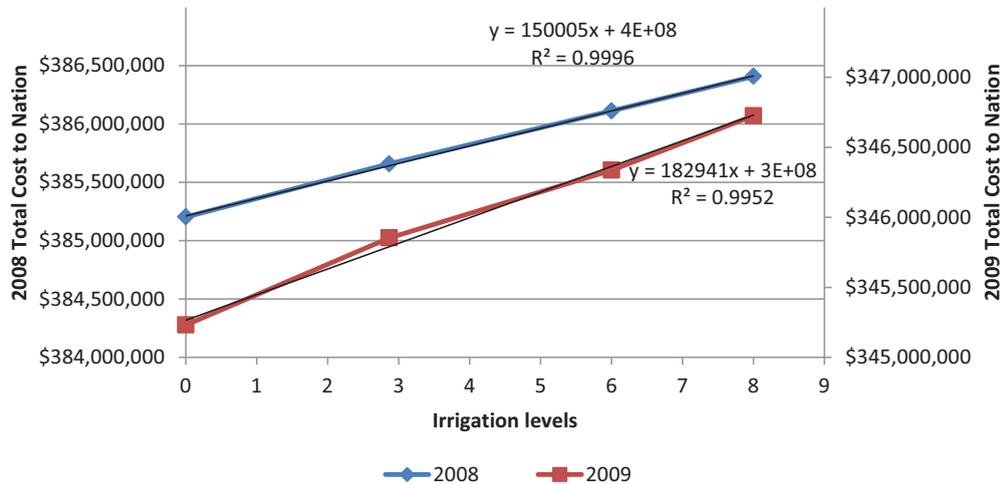


Figure 2: DOASA – Perfect hindsight model.

Under perfect hindsight, we estimate the each addition cumec irrigation will increase the national cost of electricity generation in 2008 by \$150,000, while for the year 2009 we estimate one cumec irrigation will increase national cost of electricity generation by around \$183,000.

4 Disaggregation of a river-chain

In this section we will examine the efficiency losses from a lack of coordination in a river-chain. This work has been inspired by the recent sale of Tekapo A and B from Meridian to Genesis, which prompted the question as to whether the upstream stations being owned by a separate firm could lead to issues with water supply for the downstream reservoirs. Wahid explored this in detail in (Wahid 2011), below we will present some of those results.

We will compare the efficiency of a simple two reservoir model under two different ownership structures (*integrated* and *disaggregated* ownership). Here two hydro generation plants connected via a single river (flow arc): in the integrated setting both stations are operated by a single firm; whereas for the disaggregated model we operate these stations separately. The key assumptions that we make in this work is that the generators are price takers (that is, their actions do not affect the price) and that prices are deterministic. We model the operation of the plants over a 48 period day.

The objective of each company is to maximize their individual revenues by generating electricity by releasing water from their reservoirs. As a result they primarily will aim to generate at full capacity during high prices. When the river chain is under integrated ownership the upstream generator may wish to release water earlier (when prices are not at their highest) in order to fill the downstream reservoir before for generation during high prices. This is to ensure that during these high price periods both reservoirs have sufficient water reserves in order to power their respective generators at full capacity. Hence the sole owner of the river chain is aiming to capture the maximum value of the water for high price periods, across the complete river chain. However, under the disaggregated river chain, there is a lack of coordination in the timing of the releases between the upstream and downstream generators. The owner of the upstream generator is trying to maximize the value of the water they have, while the owner of the downstream generator capitalizes on both the water stored inside its own reservoir and the water received from the releases of the upstream generator.

The operation of this simple river-chain can be formulated as a linear programming model. It determines the optimal actions of the generators a single trading day. The operation of the river-chain is optimized under the two market structures. In order to understand how the parameters of the model influence the inefficiency, the model is solved for various combinations of different downstream generation capacities and starting reservoir levels.

4.1 Formulation for integrated model

Table 2: Parameters for simple river chain model.

Variable/parameters	Description
$x_{N(t)}, x_{S(t)}$	Water level of the reservoirs North and South reservoirs respectively indexed over each time period t
$g_{A(t)}, g_{B(t)}$	Amount of electricity (MW) produced by hydro generators A and B respectively over the time period t
c_A, c_B	Generation capacity of generator A and B respectively
c_N, c_S	Reservoir capacity of usable water
p_t	Price of energy at time t

Using the parameters given in Table 2, the primal formulation of the Integrated model for the Simple River Chain over 48 time periods is given by:

$$\mathcal{I} := \text{maximize } z_1 = \sum_{t=1}^{48} p_t (g_{A(t)} + g_{B(t)}) \quad (1)$$

$$\text{subject to } x_{N(t+1)} - x_{N(t)} + g_{A(t)} = 0, \quad [\pi_{1(t)}] \quad \forall t, \quad (2)$$

$$x_{S(t+1)} - x_{S(t)} + g_{B(t)} - f_{F1(t)} = 0, \quad [\pi_{2(t)}] \quad \forall t, \quad (3)$$

$$g_{A(t)} - f_{F1(t)} = 0, \quad [\pi_{3(t)}] \quad \forall t, \quad (4)$$

$$g_{A(t)} \in [0, c_A], g_{B(t)} \in [0, c_B], \quad \forall t, \quad (5)$$

$$x_{N(t)} \in [0, c_N], x_{S(t)} \in [0, c_S], \quad \forall t, \quad (6)$$

The constraints of the integrated river-chain optimization problem \mathcal{I} are described in Table 3. Associated with each set of constraints, i , we define dual variables, $\pi_{i(t)}$. Each dual variable gives the marginal value of the resource associated with

that constraint. For example, given that $\pi_{1(t)}$ is the dual variable corresponding to constraint (2), $\pi_{1(t)}$ represents the value of an extra unit of water available to the north reservoir at time t .

Table 3: Explanation of constraints.

Equation	Description
(2), (3)	Water balance constraints of the North and South reservoirs respectively
(4)	The volume of water arriving at reservoir B in time period t must equal the volume discharged by A in time period t . Note, this assumes water discharged by A instantaneously arrives at reservoir B . However, this could be easily modified to include a delay of d periods by replacing (4) with $g_{A(t)} - f_{F1(t+dt)} = 0$
(5)	Generation capacities of plants A and B respectively
(6)	Reservoir capacities of North and South reservoirs respectively

Table 4: Explanation of dual variables for integrated model.

Dual	Description
$\pi_{1(t)}$	The value of an extra unit of water available to reservoir A at time t
$\pi_{2(t)}$	The value of an extra unit of water available to reservoir B at time t
$\pi_{3(t)}$	Change in objective if one unit of water is removed from the flow between reservoirs A and B at time t . These is only an interval of allowable change if $g_{A(t)} > 0$ otherwise the set of basic variables changes.

4.2 Formulation for disaggregated model

The disaggregated model gives rise to separate formulations for each reservoir: \mathcal{D}_N and \mathcal{D}_S , respectively. To solve the disaggregated model without water transfer pricing, Plant A's optimal generation plan is found by solving \mathcal{D}_N . Plant B's schedule is then found by solving \mathcal{D}_S with the optimal flows, $f_{F1(t)}^*$, from \mathcal{D}_N , fixed.

$$\mathcal{D}_N := \text{maximize } z_2 = \sum_{t=1}^{48} p_t g_{A(t)} \quad (7)$$

$$\text{subject to } x_{N(t+1)} - x_{N(t)} + g_{A(t)} = 0, \quad [\gamma_{1(t)}] \quad \forall t, \quad (8)$$

$$g_{A(t)} - f_{F1(t)} = 0, \quad [\gamma_{2(t)}] \quad \forall t, \quad (9)$$

$$g_{A(t)} \in [0, c_A], x_{N(t)} \in [0, c_N], \quad \forall t, \quad (10)$$

$$f_{F1(t)} \geq 0, \quad \forall t. \quad (11)$$

$$\mathcal{D}_S := \text{maximize } z_3 = \sum_{t=1}^{48} p_t g_{B(t)} \quad (12)$$

$$\text{subject to } x_{S(t+1)} - x_{S(t)} + g_{B(t)} = f_{F1(t)}^*, \quad [\mu_{1(t)}] \quad \forall t, \quad (13)$$

$$g_{B(t)} \in [0, c_B], x_{S(t)} \in [0, c_S], \quad \forall t, \quad (14)$$

Table 5: Explanation of dual variables for disaggregated model.

Dual	Description
$\gamma_{1(t)}$	The value of an extra unit of water available to reservoir A at time t .
$\gamma_{2(t)}$	Change in objective if Plant A increases its generation by one unit in period t . If Plant A is generating during period t ($g_{A(t)} > 0$) then $\gamma_{2(t)} = 0$. However, if $g_{A(t)} = 0$ then $\gamma_{2(t)}$ is equal to the difference between p_t and the lowest price that A is generating at.
$\mu_{1(t)}$	The value of an extra unit of water available to reservoir B at time t .

In the disaggregated model, $\gamma_{2(t)}$ gives the change in objective if plant A were to use a unit of water to generate a unit of power in period t . If A is already generating in period t then there is no change in the objective. However, if $g_{A(t)} = 0$ then $f_{F1(t)} = 0$, so any increase in the right-hand side must increase the generation at time t . At the optimal solution generation will reduce during the lowest priced period currently being generated during, leading to a marginal change in profit of $p_{(t)} - \gamma_{1(t)}$. This is the amount that firm A, would need to be paid in order to shift generation into time period t . On the other hand, $\mu_{1(t)}$ is the most that firm B would be willing to pay to receive additional water at time period t . Finally, we compare the outcomes from the different ownership structures to find the level of inefficiency created from splitting up the chain. We define this inefficiency measure to be the percentage drop in revenue from the integrated model: $\eta = \frac{100}{z_1} (z_1 - (z_2 + z_3))$.

4.2.1 Example

In order to understand when inefficiencies are likely to occur, we solved both models for a variety of downstream station capacities and starting reservoir levels, while keeping the capacity of the upstream station fixed. We used a price sequence from 30 June 2008, where there were relatively flat prices during the day, but they peaked at night (periods 38-45). As shown in Figure 3, inefficiencies occur when there is little water in the downstream reservoir; this is because the downstream reservoir is reliant on inflows from upstream in this situation. Inefficiencies also increase with the capacity of the downstream generator. Maximum inefficiencies occur when the downstream reservoir has initially no water for generation and the downstream generator's capacity is ten times that of the upstream generator. In this scenario there is an approximately 3.6% of revenue loss under the disaggregated ownership compared to the integrated ownership. In Table 6 we show the generation and profits made by the stations under the different ownership structures for the situation where generators A and B have capacities of 100MW, and 1000MW, respectively, and the South reservoir is initially empty.

4.3 Explanation of water transfer prices

This section focuses on incorporation of *water transfer prices* into a river chain where there are two firms operating stations in a single river-chain. In the case of the integrated model where there is a single firm controlling all the reservoirs, the upstream reservoir may release water early so that the downstream reservoir has water available in the high price periods. This behaviour ensures system efficiency. However, we have seen that once the ownership is disaggregated, this no early re-

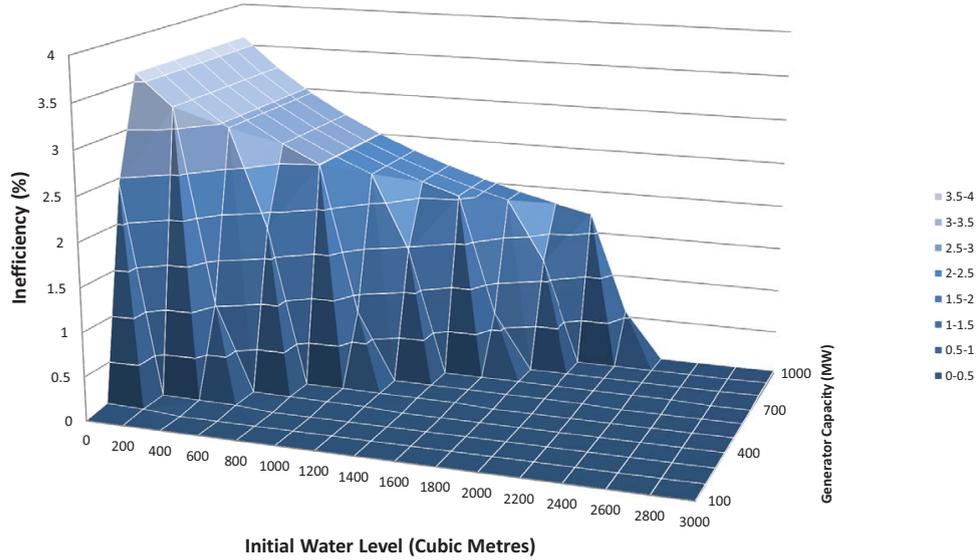


Figure 3: Inefficiency as a function of downstream capacity and initial water level.

lease no longer happens, leading to an overall loss in system efficiency. However, this system efficiency can be recovered by introducing water transfer prices. Water transfer prices allows the downstream reservoir to pay the upstream reservoir to release water early so that water is available downstream when it is needed.

Both (Lino et al. 2003) and (Wahid 2011) give the water transfer price as the dual variable of the water balance constraint of the south reservoir. This corresponds to the value of an extra unit of water at the south reservoir ($\pi_{2(t)}$). When water transfers are priced based on this dual variable it can be shown that we recover the integrated solution, thereby restoring the overall efficiency of the system. Moreover, this water transfer price is guaranteed to be between the cost incurred by firm A, and the value received by firm B, meaning that both firms would be willing to negotiate such a transfer price. With the inclusion of these water transfer prices the objective function for the \mathcal{D}_N given in (7) becomes:

$$\text{maximize } z_4 = \sum_{t=1}^{48} (p_t g_{A(t)} + \pi_{2(t)} \Delta g_{A(t)}) \quad (15)$$

where $\Delta g_{A(t)}$ gives firm A's deviation from their optimal generation plan from \mathcal{I} . On the other hand, the objective function of \mathcal{D}_S (12) becomes:

$$\text{maximize } z_5 = \sum_{t=1}^{48} (p_t g_{B(t)} - \pi_{2(t)} \Delta g_{A(t)}), \quad (16)$$

where firm B is incurring a cost due to firm A shifting generation between time periods.

When there is a time delay on the flow between A and B the water transfer prices are shifted so that B pays the water transfer price for period t for A to release water in period $t - d$ where d is the delay. This means the water transfer prices for periods $T - d + 1, \dots, T$ is zero as any water A releases in these periods will not reach B before the end of the time horizon.

Table 6: Comparison of the revenue earned between the integrated and single ownership for the scenario of 1000 MW generator A’s capacity with no water reserves for the South reservoir

		Integrated		Disaggregated	
Period	Price (\$/MW)	Dispatch A (MW)	Dispatch B (MW)	Dispatch A (MW)	Dispatch B (MW)
41	\$498.23	100	900	100	0
40	\$498.23	100	0	100	600
42	\$498.20	100	100	100	100
36	\$389.87	100	0	100	0
39	\$389.48	100	0	100	0
43	\$384.64	0	0	100	100
44	\$384.64	0	0	100	100
45	\$384.64	0	0	100	100
35	\$383.47	100	0	100	0
38	\$383.47	100	0	100	0
37	\$382.90	100	0	0	0
27	\$380.27	100	0	0	0
22	\$379.35	100	0	0	0
Generator Earnings		\$418,347.00	\$498,227.00	\$419,487.00	\$464,150.00
Total Earnings		\$916,574.00		\$883,637.00	

To solve the disaggregated model with the inclusion of water transfer prices we: solve the integrated model and extract the water transfer price, $\pi_{2(t)}$; solve the disaggregated model and extract the optimal generation plan of firm A, $g_{A(t)}^*$; solve the disaggregated model with inclusion of the benefits and costs of the water transfer price, $\pi_{2(t)}$, and firm A’s disaggregated generation plan, $g_{A(t)}^*$.

Solving the model in this way in a deterministic setting recovers the solution to the integrated model, ensuring there is no inefficiency loss. However, due to the water transfer price, there may be a wealth transfer from firm B to firm A.

References

- Farishta, Zabin. 2011. “Should Farmers get the water rights or Energy Companies.” Technical Report, The University of Auckland. preprint.
- Lino, P., L.A.N. Barroso, M.V.F. Pereira, R. Kelman, and M.H.C. Fampa. 2003. “Bid-based dispatch of hydrothermal systems in competitive markets.” *Annals of Operations Research* 120 (1): 81–97.
- Ministry of Forestry and Environment. 2010, December. “Freshwater demand (allocation).” Technical Report.
- Philpott, Andy, Ziming Guan, Javad Khazaei, and Golbon Zakeri. 2010. “Production inefficiency of electricity markets with hydro generation.” *Utilities Policy* 18 (4): 174–185.
- Wahid, Faisal. 2011. “What might happen in the Tekapo A and B transfer.” Technical Report, The University of Auckland. preprint.

Optimal Delivery of Arc Modulated Radiation Therapy in Cancer Treatment

J. Du, M. Ehrgott, and A. Raith
Department of Engineering Science
The University of Auckland
New Zealand
jdu018@aucklanduni.ac.nz

Abstract

Arc Modulated Radiation Therapy (AMRT) is a recently developed radiation therapy technique used to treat patients with cancer. AMRT delivers radiation in one continuous gantry rotation and finds optimal beam intensities at equally spaced angles to maximise tumour control and minimise normal tissue complication probability. In order to achieve an optimal delivery of a treatment plan, it is necessary to modulate the radiation intensity through a device called the multi-leaf collimator (MLC). By controlling the movement of the collimator leaves, we wish to deliver the planned intensities as closely as possible.

In this paper, we compare two different methods proposed in literature for optimal AMRT leaf sequencing. These methods sequence each leaf pair independently, and we present the findings of our comparisons in terms of computation time, delivery error, and beam-on time. We also discuss the impact of technological restrictions on the MLC leaves that may introduce multiple leaf pair dependencies and violate the independence assumption.

Key words: Arc modulated radiation therapy (AMRT), multi-leaf collimator (MLC), leaf sequencing.

1 Introduction

Arc Modulated Radiation Therapy (AMRT) was introduced in 1995 as a novel approach to radiation therapy treatment. Radiation is delivered by a linear accelerator (*linac*), and consists of a single continuous gantry rotation (*arc*). This is different to Intensity Modulated Radiation Therapy (IMRT) where the linac turns off when moving between delivery positions, and is more efficient than Intensity Modulated Arc Therapy (IMAT) which uses multiple delivery arcs (Zhu et al. 2010).

The radiation is modulated by a device inside the linac called the *multi-leaf collimator* (MLC) which consists of 20-100 movable metal leaf pairs. The leaves move to block radiation, forming several *apertures* (opening shapes see middle of Figure 1) over different positions to achieve planned *intensity profiles* (Kamath et al. 2009). Naturally, this involves controlling the MLC leaves so that enough radiation



Figure 1: Rotating linac (left), a typical MLC (middle), an intensity matrix (right). Source: OncoSantana; Varian medical systems; Chen, Luan, and Wang (2011).

is delivered to the tumour while minimising the radiation exposure to healthy tissue.

We can represent the MLC as a grid with $M \times N$ *bixels* (elements), where M is the number of leaf pairs and N is the width of the MLC (*leaf range*). This directly maps to a set of intensity matrices, one for each delivery angle, where the elements represent the planned intensity to be delivered to the patient.

The difficulty with AMRT leaf sequencing is that the linac will be rotating around the patient in one continuous arc movement. This means the transition between apertures and neighbouring delivery angles must conform to leaf speed constraints as the linac will not be stopping during the treatment. Therefore, while accurate delivery of a single intensity matrix can always be achieved in IMRT, the problem is not as straightforward in AMRT and we have to take intensity matrices of neighbouring angles into account. Consequently, we want to minimise any deviations that may arise between the actual delivered intensity and the planned intensity, referred to as the *delivery error* from the AMRT leaf sequence measured in *monitor units*.

Sections 2 to 4 of this paper briefly describe two AMRT leaf sequencing methods, and Section 5 compares them in terms of computation speed, delivery error, and beam-on time. These methods sequence each MLC leaf pair independently, hence we consider only a single row of an intensity matrix in this paper. In some existing MLCs this is not the case, so Section 6 will discuss some possible MLC constraints that introduce leaf dependencies and whether these two methods can be adapted.

2 Method 1: Mixed Integer Programming

We have based our first method on the paper written by Zhu et al. (2010), who formulated the leaf sequencing problem using mixed integer programming. The following MIP model is constructed for each leaf pair.

Indices:

$k \in A$ = the index of the delivery angle, $A = \{1, 2, \dots, K\}$;

$t \in S$ = the index of the time step within a delivery angle, $S = \{1, 2, \dots, T_k\}$;

$n \in P$ = the position along the width of the MLC, $P = \{1, 2, \dots, N\}$.

Parameters:

range = equivalent to $N + 1$ (see Figure 2);

$f_{k,n}$ = the planned intensity to be delivered at bixel n of the k th angle.

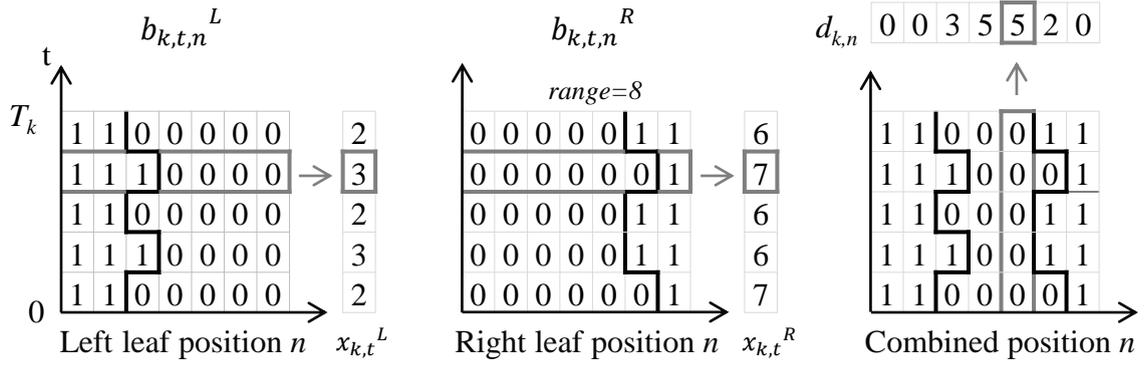


Figure 2: Mixed integer program representation, adapted from Zhu et al. (2010).

Variables:

- $b_{k,t,n}^L = 1$ if bixel n is blocked by the left leaf at k th angle, time t (0 otherwise);
- $b_{k,t,n}^R = 1$ if bixel n is blocked by the right leaf at k th angle, time t (0 otherwise);
- $x_{k,t}^L =$ left leaf edge position at k th angle, time t ;
- $x_{k,t}^R =$ right leaf edge position at k th angle, time t ;
- $d_{k,n} =$ the actual intensity delivered at k th angle, bixel n ;
- $od_{k,n} =$ amount by which $d_{k,n}$ exceeds $f_{k,n}$ at k th angle, bixel n ;
- $ud_{k,n} =$ amount by which $f_{k,n}$ exceeds $d_{k,n}$ at k th angle, bixel n .

Formulation:

$$\begin{aligned}
 & \text{minimise} && \sum_{k \in A, n \in P} ud_{k,n} + od_{k,n} \\
 s.t. & && f_{k,n} - ud_{k,n} + od_{k,n} = d_{k,n} && \forall k \in A, n \in P && \text{(Definitions)} \\
 & && \sum_{t \in S} b_{k,t,n}^L + \sum_{t \in S} b_{k,t,n}^R = T_k - d_{k,n} && \forall k \in A, n \in P \\
 & && x_{k,t}^L - \sum_{n \in P} b_{k,t,n}^L = 0 && \forall k \in A, t \in S \\
 & && x_{k,t}^R + \sum_{n \in P} b_{k,t,n}^R = range && \forall k \in A, t \in S \\
 & && b_{k,t,n}^L - b_{k,t,n+1}^L \geq 0 && \forall k \in A, t \in S, n = 1, \dots, N-1 && \text{(Contiguity)} \\
 & && b_{k,t,n+1}^R - b_{k,t,n}^R \geq 0 && \forall k \in A, t \in S, n = 1, \dots, N-1 \\
 & && x_{k,t}^L - x_{k,t}^R \leq -1 && \forall k \in A, t \in S && \text{(No overlap)} \\
 & && |x_{k,t}^R - x_{k,t+1}^R| \leq 1 && \forall k \in A, t = 1, \dots, T-1 && \text{(Speed)} \\
 & && |x_{k,t}^L - x_{k,t+1}^L| \leq 1 && \forall k \in A, t = 1, \dots, T-1 \\
 & && |x_{k,T_k}^L - x_{k+1,1}^L| \leq 1 && \forall k \in A \\
 & && |x_{k,T_k}^R - x_{k+1,1}^R| \leq 1 && \forall k \in A \\
 & && b_{k,t,n}^L \in \{0, 1\} && \forall k \in A, t \in S, n \in P && \text{(Binary)} \\
 & && b_{k,t,n}^R \in \{0, 1\} && \forall k \in A, t \in S, n \in P \\
 & && ud_{k,n}, od_{k,n} \geq 0 && \forall k \in A, n \in P && \text{(Non-negative)}
 \end{aligned}$$

The objective is to minimise the total deviation between planned and delivered intensities. The Definition constraints define relationships between the variables and

parameters. The Contiguity constraints ensure that contiguous leaves are connected to their respective MLC banks. The No overlap constraints prevent opposing left and right leaves from hitting each other. The Speed constraints limit the leaf movements to one bixel per time step, which also applies to adjacent delivery angle transitions.

The MIP is formulated for each leaf pair of the MLC, using the commercial optimisation package Gurobi Optimizer 5.0. The combined leaf motion sequences across all leaf pairs gives the final deliverable apertures for the AMRT treatment.

3 Method 2: Shortest Paths

The second method is based on a combination of the shortest path algorithm proposed by Wang et al. (2008), and the Constrained Coupled Path Planning (CCPP) algorithm by Chen, Luan, and Wang (2011). To find the optimal leaf trajectories for a single leaf pair, a directed acyclic graph G (see Figure 3) is constructed using three steps as follows:

1. Enumerate all feasible pairs of left-right leaf positions at every delivery angle. These positions are arranged into $K + 1$ vertex layers L_0, L_1, \dots, L_K , where the k th layer corresponds to the leaf positions at the k th angle.
2. A directed edge is constructed between every two vertices from adjacent vertex layers L_k and L_{k+1} . Each edge is weighted with the minimum delivery error associated with delivering the planned intensity f_{k+1} , starting and ending at the tail and head of the edge. This can be found using the CCPP algorithm developed by Chen, Luan, and Wang (2011), see Section 3.1.
3. The edges connecting to the source and sink vertices, s and t , are zero weighted. We then solve the s -to- t shortest path problem using Dijkstra's algorithm which will yield the optimal MLC leaf trajectory for the AMRT treatment with minimum overall delivery error for this leaf pair.

The algorithm is then repeated for each leaf pair in the MLC, and the individual results are combined to give the overall MLC leaf sequence for the AMRT treatment.

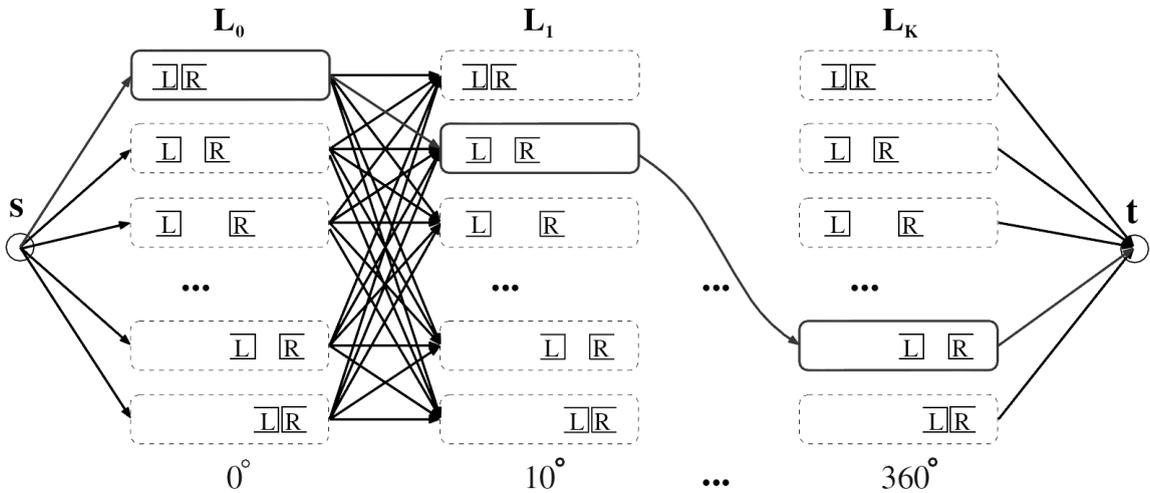


Figure 3: The directed acyclic graph G with shortest path (grey) (Wang et al. 2008).

3.1 Edge weights using Constrained Coupled Path Planning

Given the start and end positions, the left and right leaf trajectories over a specified beam-on time T can be represented by two non-crossing x - z monotone paths, p_l and p_r (black), see Figure 4. The vertical shaded segments (grey) enclosed between these paths gives the delivered intensity d . Let $f(n)$ and $d(n)$ denote the point-wise values of the planned and delivered intensities, respectively, of each bixel n across the leaf range. To find the optimal leaf sequence we have the objective function:

$$\text{Minimise } \sum_n |f(n) - d(n)| \quad (1)$$

To model the two paths, we can represent each of the enclosed vertical segments by its bottom and top co-ordinates (α, β) . We then construct a directed acyclic digraph (DAG), where the N vertex layers consist of an enumeration of feasible (α, β) co-ordinates at their respective bixel n across the leaf range. We have also extended the ideas presented in Chen, Luan, and Wang (2011) to include monotone decreasing paths as well as ‘wider’ paths i.e. $r_{start} > l_{end}$, where r_{start} is the starting right leaf position at $t = 0$, and l_{end} is the ending left leaf position at $t = T$.

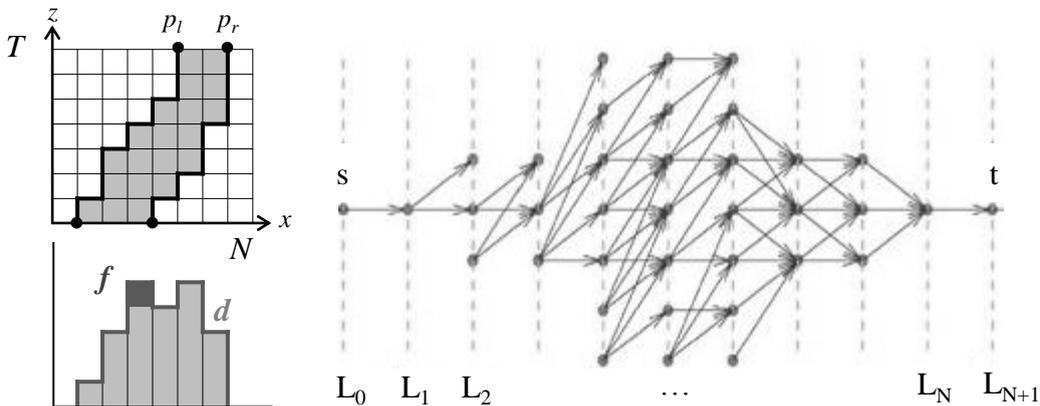


Figure 4: Constrained coupled path planning (Chen, Luan, and Wang 2011): a pair of leaf trajectories (top), delivered intensities d and planned intensities f (bottom), an example of a constructed DAG (right).

An edge connects two vertices from adjacent vertex layers L_n and L_{n+1} if the resulting leaf movement satisfies the maximum speed constraint, known as the c -step condition (see Chen, Luan, and Wang (2011) for more details). The edges are weighted with the bottleneck error of the tail node, $|f(n) - d(n)| \leq \Delta$, where Δ is a tolerance value used to limit the number of vertices generated in the DAG.

Two dummy vertices are also constructed: the source $s = (0, 0)$ in L_0 and sink $t = (T, T)$ in L_{N+1} . Edges from and to these vertices, respectively, have zero weight. Consequently, any s -to- t path in the DAG gives a sequence of vertical bars across the $T \times N$ grid to represent a non-crossing pair of x - z monotone leaf trajectories.

The total error of the leaf trajectories is given by the sum of the bottleneck errors $|f(n) - d(n)|$ in the s -to- t path. Therefore, the shortest s -to- t path in the network will define the optimal leaf trajectory (recall objective function (1)) with the minimum delivery error used to weight the edge in network G in Section 3. Chen, Luan, and Wang (2011) used a DAG with vertex weights and developed a special shortest path algorithm. However, since our DAG is edge weighted, the shortest path can be found using a standard shortest path algorithm such as Dijkstra’s.

4 Model parameters

We use the TNMU algorithm developed by Engel (2005) to find the unconstrained minimum beam-on time UT_k required for each intensity matrix. We then set $T_k = \max\{UT_k, 1\}$ to avoid $T_k = 0$ if there is no intensity to be delivered for the k th angle. Although it is unlikely that the optimal leaf sequence will achieve zero delivery error with the minimum unconstrained beam-on time, the resulting error can give an indication of how beam-on time is impacted by the constraints in AMRT.

The value of the bottleneck error tolerance Δ for CCP is hard to determine because large values will increase the number of feasible vertices in the DAG network and hence increase computation time, while small values may over-restrict the DAG network and result in high delivery error leaf sequences as the optimal solution.

As a result, we have used a variable Δ for each leaf pair by finding a guideline tolerance tol for each leaf pair m that reflects the unconstrained decomposition time for the leaf pair over all the K delivery angles. This is achieved by extracting the rows corresponding to the m th leaf pair (i.e. the m th rows) from all the intensity matrices to create a new matrix. We then use Engel's TNMU algorithm on this new matrix to find the unconstrained decomposition time tol for this leaf pair. This tol is set as a maximum allowable value for Δ and can then be scaled by some constant.

5 Results and comparisons

The following sections summarise the results obtained using an Intel Core i5 2520M 2.5GHz processor, with 4GB RAM running Windows 7 Professional (64-bit). We have conducted the various tests on a single leaf pair as the *relative* performance between the methods would have similar comparisons for multiple leaf pairs.

5.1 Speed and delivery error

We compared speed and delivery error over different leaf ranges by testing the two methods using random intensity profiles ranging from 2 bixels to 15 bixels wide. Top left of Figure 5 shows that whether or not the Shortest Paths method (*SP*) has a speed advantage greatly depends on the bottleneck tolerance Δ . We varied Δ between 100% of tol to 25% of tol , and found that a smaller Δ was very effective at reducing computation speed as the delivery errors of both methods were the same.

With more than one delivery angle, a lower Δ may result in a larger overall delivery error because a lower error at one angle may mean the leaf cannot get to a position to lower the error at the next angle. This was evident when we compared *MIP* and *SP* by varying the number of delivery angles from 2 to 7. The top and bottom right of Figure 5 show that a low Δ may compromise *SP*'s delivery error, but since *SP* has a clear advantage in speed, it can afford higher values for Δ . In this case, setting Δ to be higher than 50% of tol did not reduce the delivery error. However, *MIP*, although computationally very expensive, was always able to find a solution with the same or lower delivery error than *SP*.

5.2 Delivery error and total beam-on time

Intuitively, we expect the delivery error obtained from both methods to improve as total beam-on time increases from the unconstrained minimum. To assess how the

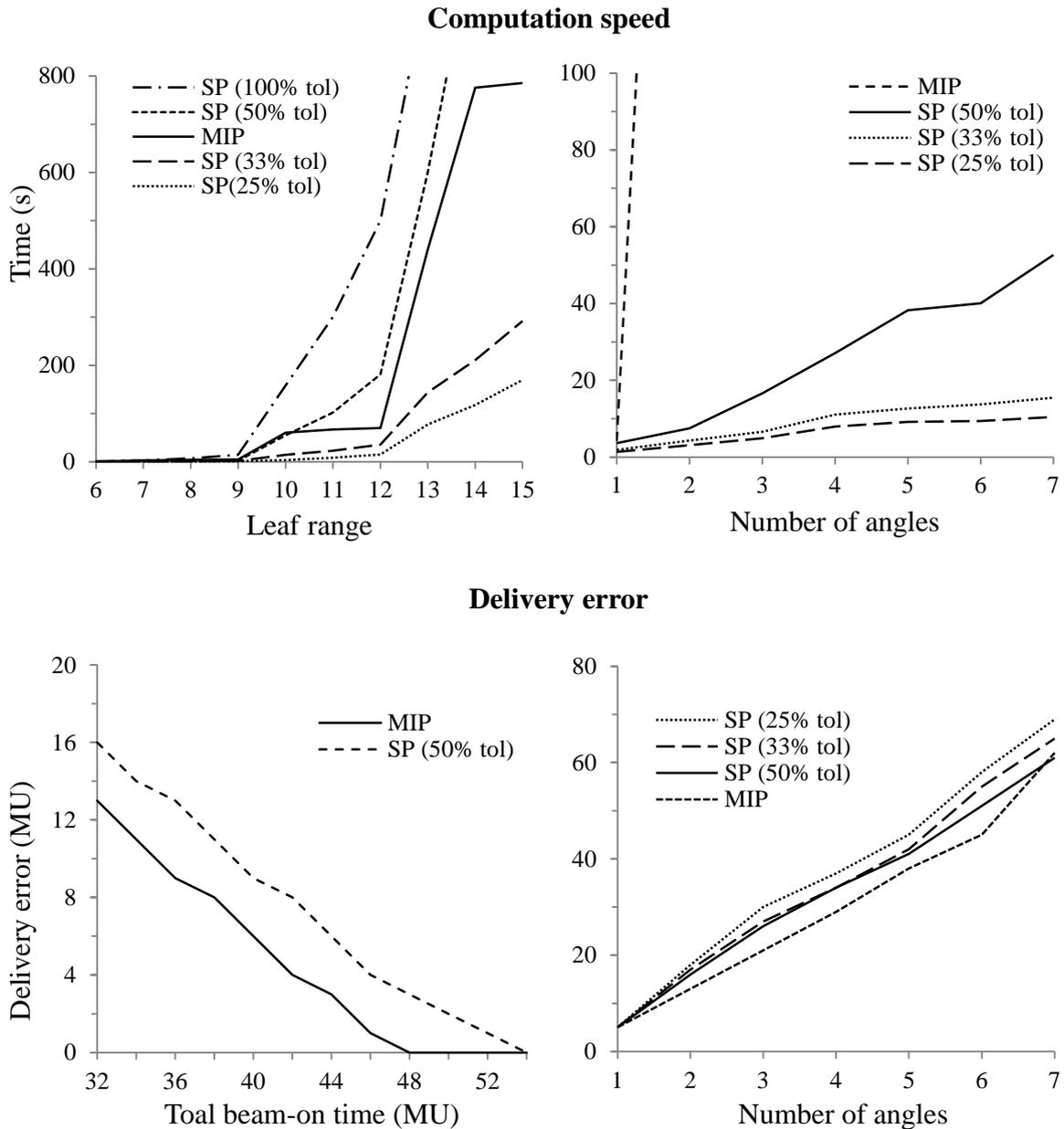


Figure 5: Results from computation speed and delivery error comparisons.

trade-off differs for the two methods, we took the first two angles of the data used in Section 5.1, and plotted the delivery errors against increasing total beam-on times from the unconstrained minimum until a zero-error solution was obtained. Δ was set to 50% of tol to obtain the lowest error solution achievable by SP . The bottom left graph of Figure 5 confirms that there is a trade off between the beam-on time and the delivery error, and suggest that the rate of the trade off is relatively similar between MIP and SP . The figure also shows that MIP consistently found optimal solutions with lower delivery error than SP , given a particular beam-on time.

5.3 Discussion

The advantage of SP is that solving a shortest path problem is much faster than solving a mixed integer program using branch and bound techniques like in the MIP method. However, the MIP method is usually better at finding a lower error leaf sequencing solution, given the same beam-on times. This is attributed to the fact

that the CCP algorithm in *SP* requires c -steep, x - z monotone leaf trajectories that satisfy a bottleneck error tolerance Δ . These restrictions do not apply to *MIP*.

The c -steep property related to the maximum speed constraint also implies that each uncovered area must be irradiated for at least c time steps. Since the ending leaf position of the previous delivery angle is the starting position for the next angle, this means the same leaf position is irradiated for at least $2c$ time steps during the angle transition. This ‘double-up’ is not mandatory for *MIP*. *MIP* solutions are also allowed to move both to the left and to the right of the MLC in the same delivery angle, and do not have a bottleneck error tolerance. *MIP* only focuses on minimising the total delivery error therefore *MIP* may violate Δ if the resulting total error is lower. These properties of *SP* may cause the algorithm to require a longer total beam-on time to obtain a zero-error solution, hence we observed that *MIP* tends to produce solutions with lower delivery error given the same beam-on times.

However, having a Δ may be desirable because a high delivery error (e.g. 10 MU) on one angle and a low delivery error (e.g. 1 MU) on another, is not necessarily better than having a medium delivery error (e.g. 6MU) on both angles. The first solution will have a lower total delivery error, so *MIP* will prefer this solution over the second but *SP* may find the second solution if the first one violates Δ . If there are vital healthy organs close to the targeted tumour, we may not wish to have a low delivery error at some places at the expense of high radiation exposure in other places.

6 MLC constraints

MLC-specific constraints can cause dependencies between different leaf pairs. These would need to be taken into account during constrained leaf sequence optimisation.

6.1 Inter-leaf motion and maximum leaf spread

The MLC may require that opposing left and right leaves of adjacent rows do not overlap, or that there must be a minimum gap δ between them (see top left Figure 6). This constraint prevents adjacent left-right leaves from hitting each other if the leaf movements are not perfectly level (Ehrgott, Hamacher, and Nußbaum 2008).

The maximum leaf spread constraint specifies a maximum distance at which leaves extending from the same leaf bank can be spread apart (Kamath et al. 2009). This constraint is present in some MLC systems, see bottom left Figure 6.

In principle it would be possible to expand the *MIP* model to include variables that represent all MLC leaf pairs, and then cross-constrain these variables to match the extra constraints. This would be reasonably straightforward as there is a variable representing each bixel of the irradiated area therefore each leaf pair in the new model would just require the same set of the variables as in the current formulation. Limiting the relative distances between all adjacent leaf pair positions should also be reasonably easy to model since there are specific variables in *MIP* that represent leaf pair positions. However if the problem is large, this will further increase the solve time of *MIP* which may not be a practical option.

Adapting the *SP* method may be less straightforward because the vertices in the shortest path network G represents leaf positions for a single leaf pair. To represent all leaf pairs, we will have to enumerate not just the openings for a single leaf pair but also the combinations of openings across multiple leaf pairs. For ex-

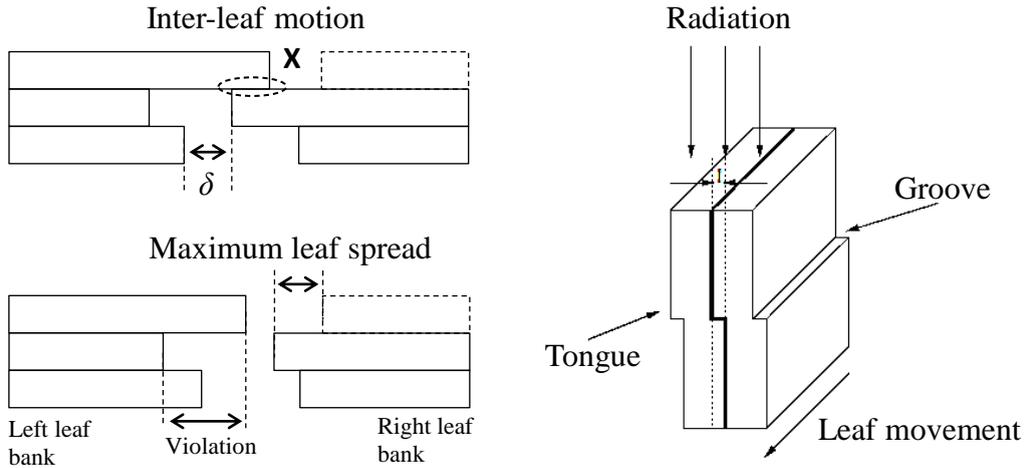


Figure 6: Inter-leaf motion constraint (top left), maximum leaf spread (bottom left), tongue and groove error (right, (Kamath et al. 2009)).

ample, if we have a 2×2 intensity matrix, each vertex layer in G for the first leaf pair would require the following leaf pair positions: (00), (01), (02), (11), (12), and (22). However, if we were to include the second leaf pair as well, then this enumeration must contain all combinations of feasible leaf openings for leaf 1 and leaf 2: (00₁ 00₂), (00₁ 01₂), (00₁ 02₂), (01₁ 00₂), \dots , etc. The numbers in brackets represent the left and right leaf positions, and the subscript indicates the leaf pair number. This would lead to an explosion of vertices, which may be impractical and inefficient.

Similarly, the CCP algorithm (which weighs the edges of network G) must also undergo major transformations. It will need to solve for two sets of coupled paths and find the optimal combination of these coupled paths across the leaf pairs while checking for violations of the inter-leaf and maximum leaf spread constraints. Hence, adapting the SP algorithm to incorporate these new constraints may be much more difficult compared to MIP . It may be more practical to consider post-optimisation adjustments where the leaves are optimised individually and then adjusted afterwards if some of the leaves violate the new constraints. However, the result of this adjustment may no longer be an optimal solution.

6.2 Tongue and groove error

Many MLC types have tongue and groove (TG) joints between neighbouring leaf pairs (Kamath et al. 2009), see right Figure 6. The advantage of having TG joints is so that radiation leakage between adjacent leaf pairs can be reduced. However, the thinner material in the TG causes uneven radiation leakage if a bixel immediately on either side is not covered by an adjacent leaf. This will cause inaccuracies (TG error) between the intended and actual delivered intensities.

Ehrgott, Hamacher, and Nußbaum (2008) mentioned that it may be possible to reduce this error by swapping the order of certain sets of leaf sequences within the same leaf pair. However, TG error often cannot be avoided (whether or not we optimise leaves independently) because as long as the leaves will form apertures, there will be some degree of tongue and groove error along the outline of these apertures where a thinner portion of the leaf is exposed to radiation. The maximum leaf spread constraint may be able to reduce this error since having a limited spread

will reduce the length of the leaf where the groove is not connected to its neighboring tongue and thus reduce the amount of radiation exposure to thinner leaf sections.

7 Conclusions

The *MIP* method generally gives optimal solutions with lower delivery error than *SP* for AMRT, but is much slower in terms of computation speed. The latter is understandable, since solving a shortest path problem is faster than solving a mixed integer program using branch and bound. *SP* has higher delivery error due to the x - z monotone, c -steep, and bottleneck error requirements, which are not present in *MIP*. However, a bottleneck error tolerance may be desirable as we may not want a low error solution (overall) at the expense of a high radiation exposure to any particular area, especially where there may be vital healthy organs nearby.

The addition of technological constraints such as inter-leaf motion and maximum leaf spread, can be modelled by solving for all the leaf pairs together, rather than individually as the two methods considered in this paper. Some tongue and groove error is usually unavoidable, although the maximum leaf spread constraint may be able to reduce this error. Both methods can potentially be modified to optimise for more than a single-leaf pair at a time. However, this may present memory and further computation issues as the size of the models will be much bigger. The practicality and efficiency of such modifications will need to be investigated.

Acknowledgments

I would like to thank my supervisors, Professor Matthias Ehrgott and Dr. Andrea Raith, for all their support and guidance throughout this project.

References

- Chen, D. Z., S. Luan, and C. Wang. 2011. "Coupled Path Planning, Region Optimization, and Applications in Intensity-modulated Radiation Therapy." *Algorithmica* 60 (1): 152–174.
- Ehrgott, M., H. W. Hamacher, and M. Nußbaum. 2008. "Decomposition of matrices and static multileaf collimators: a survey." In *Optimization in Medicine*, 25–46. Springer New York.
- Engel, K. 2005. "A new algorithm for optimal multileaf collimator field segmentation." *Discrete Applied Mathematics* 152:35 – 51.
- Kamath, S., S. Sahni, J. Palta, S. Ranka, and J. Li. 2009. "Algorithms for Sequencing Multileaf Collimators." In *Handbook of Optimization in Medicine*, 1–44. Springer US.
- Wang, C., S. Luan, C. Tang, D. Z. Chen, M. A. Earl, and C. X. Yu. 2008. "Arc-modulated radiation therapy (AMRT): a single-arc form of intensity-modulated arc therapy." *Physics in Medicine and Biology* 53 (22): 6291–6303.
- Zhu, X., D. Thongphiew, R. McMahon, T. Li, V. Chankong, F. Yin, and Q. J. Wu. 2010. "Arc-modulated radiation therapy based on linear models." *Physics in Medicine and Biology* 55 (13): 3873–3883.

An Evaluation Tool for Reservoir Management

Shane Dye, E Grant Read, Rosemary Read, and Stephen Starkey
Department of Management
University of Canterbury
New Zealand
shane.dye@canterbury.ac.nz

Abstract

We describe an evaluative tool based on Stochastic Constructive Dual Dynamic Programming (SCDDP) for the operation of a reservoir-based system under uncertain inflows to benefit multiple participants. The tool allows the evaluation of various market-based operating policies under uncertainty. The tool could evaluate hydroelectric reservoir planning, water resource management or a mixed use reservoir.

In operating a reservoir under uncertain inflows the most significant trade-off is between releasing water for immediate benefit and storing water for an uncertain future benefit. With multiple participants, information is needed about all participants' future requirements. The information requirements may be overwhelming if contingent on outcomes of uncertainties in the future.

The evaluative tool allows us to compare the benefits of alternative operating policies with different information requirements. The use of SCDDP allows direct evaluation of the probability distribution of various system outputs without resorting to Monte Carlo sampling.

Key words: Reservoir management, Hydro-generation, Markets.

Data Envelopment Analysis without Linear Programming

Matthias Ehrgott, Maryam Hassanasab, Andrea Raith
Department of Engineering Science
The University of Auckland
New Zealand
m.ehrgott@auckland.ac.nz

Abstract

Data envelopment analysis (DEA) is a very popular parameter free method for performance measurement of decision making units. Based on linear programming (LP), DEA is closely related to multi-objective linear programming (MOLP) in the sense that efficient decision making units represent efficient solutions of an MOLP. We exploit this relationship and apply the primal and dual variants of Benson's outer approximation algorithm for MOLP as presented in Ehrgott, Löhne, and Shao (2012) in order to solve DEA problems. We show that many of the LPs that need to be solved in these algorithms, when applied to DEA, reduce to trivial problems of finding the minima of finite sets. The geometric duality of multi-objective linear programming furthermore allows us to identify all efficient DMUs without solving a linear programme for every DMU using the dual outer approximation algorithm. Moreover, the primal outer approximation algorithm directly finds all hyperplanes defining the efficient frontier of the production possibility set. We demonstrate the efficiency of our algorithms on a number of DEA reference problems.

Key words: Data envelopment analysis, multi-objective linear programming, outer approximation algorithm, dual outer approximation algorithm.

1 Data Envelopment Analysis

Data envelopment analysis (DEA) is a linear programming based technique for performance measurement of comparable decision making units (DMUs). It was introduced by Charnes, Cooper, and Rhodes (1978) and in the last three and a half decades, hundreds of papers on the topic have been published. To assess the efficiency of a DMU, a fractional programming problem is solved that maximises the weighted sum of s outputs to the weighted sum of m inputs, subject to normalisation constraints. The variables in this fractional programme are the weights, which implies that each DMU is allowed to choose its most favourable weights. To formally define data envelopment analysis, we assume that DMU j is represented by a pair

of input and output vectors (x^j, y^j) . These define a production possibility set

$$T_\Lambda = \left\{ (x, y) \in \mathbb{R}^m \times \mathbb{R}^s : x \geq \sum_{j=1}^n \mu_j x^j, y \leq \sum_{j=1}^n \mu_j y^j, \mu \in \Lambda \right\}$$

of potential DMUs based on some assumptions on the returns to scale, which are incorporated in the definition of the set Λ . The most common return to scale models are constant returns to scale (Charnes, Cooper, and Rhodes 1978) and variable returns to scale (Banker, Charnes, and Cooper 1984). In this paper we will only consider variable returns to scale. The fractional programme mentioned above can be transformed into a linear programme, the dual of which is

$$\begin{aligned} & \min \theta & (1) \\ \text{s.t. } & \sum_{j=1}^n \mu_j x_i^j \leq \theta x_i^o \quad i = 1, \dots, m, \\ & \sum_{j=1}^n \mu_j y_r^j \geq y_r^o \quad r = 1, \dots, s, \\ & \mu \in \Lambda_V, \end{aligned}$$

with $\Lambda_V = \left\{ \mu \in \mathbb{R}^n : \sum_{j=1}^n \mu_j = 1, \mu \geq 0 \right\}$. (1) is commonly known as the envelopment form of an input oriented DEA model. (1) attempts to scale down the inputs x^o of DMU o without reducing its outputs y^o . If the optimal value of θ^* in (1) is 1, no such scaling down is possible, and DMU o is considered efficient, otherwise it is inefficient, indicating that the same output levels could have been achieved with lower inputs. Model (1) is of major interest in this paper.

Definition 1 (DEA Efficiency) *Point $(x, y) \in T_{\Lambda_V}$ is an efficient point if and only if the optimal value of problem (1) with $x^o = x$ and $y^o = y$ is 1 and all constraints are binding at all optimal solutions of the problem.*

2 DEA as Multi-objective Linear Programme

Figure 1 shows a plot of the inputs and outputs of 1000 DMUs with a single input (horizontal axis) and a single output (vertical axis). There are five efficient DMUs, and every point on the four line segments between them as indicated in Figure 1 corresponds to a DEA efficient point.

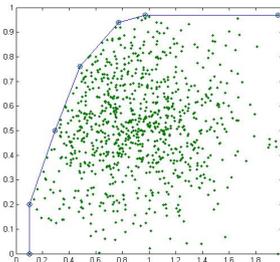


Figure 1: Production possibility set T_{Λ_V} of a variable returns to scale model.

From the definition of T_Λ and $\Lambda := \Lambda_V$ it is clear that the set of DEA efficient points is the “north-west” part of the boundary of the convex hull of all input-output vectors (x^j, y^j) of all DMUs $j = 1, \dots, n$. Therefore we have Theorem 1.

Theorem 1 *Every extreme point of T_{Λ_V} is the input-output vector (x^j, y^j) of an efficient DMU.*

Figure 1 illustrates that a DMU is efficient if it is not possible to increase its outputs without increasing its inputs, or vice versa, it is not possible to decrease its inputs without also decreasing its outputs. This observation alludes to the concept of efficiency or Pareto-optimality in multi-objective optimisation. In fact, it is possible to formulate DEA simultaneously for all DMUs as a multi-objective linear programme (MOLP), as shown by Lotfi et al. (2008). The MOLP formulation of a variable returns to scale DEA model is (2).

$$\begin{aligned}
 & \min x_1, \dots, x_m, -y_1, \dots, -y_s & (2) \\
 & \text{s.t.} \quad - \sum_{j=1}^n \mu_j x^j + x \geq 0, \\
 & \quad \quad \quad \sum_{j=1}^n \mu_j y^j - y \geq 0, \\
 & \quad \quad \quad \sum_{j=1}^n \mu_j = 1, \\
 & \quad \quad \quad \mu \geq 0.
 \end{aligned}$$

Theorem 2 (Lotfi et al. 2008) *Each efficient solution of (2) corresponds to a DEA efficient point and vice versa.*

3 Multi-objective Linear Programming

A multi-objective linear programme is an optimisation problem in n continuous variables with p linear objectives and m linear constraints as shown in equation (3)

$$\min\{Cx : Ax \geq b, x \in \mathbb{R}^n\}. \quad (3)$$

We define $\mathcal{X} := \{x \in \mathbb{R}^n : Ax \geq b\}$ as the feasible set in variable (decision) space and its image $\mathcal{Y} := \{Cx \in \mathbb{R}^p : x \in \mathcal{X}\}$ as the feasible set in objective (outcome) space. A feasible solution $\hat{x} \in \mathcal{X}$ is efficient if there is no other $x \in \mathcal{X}$ with $Cx \leq C\hat{x}$ and \mathcal{X}_E is the set of all efficient solution. If \hat{x} is efficient we call $C\hat{x}$ a non-dominated point. \mathcal{Y}_N is the set of all non-dominated points. The goal of multi-objective linear programming is to find all non-dominated points and for each $y \in \mathcal{Y}_N$ one $x \in \mathcal{X}_E$ such that $Cx = y$. We will assume that there is some $y \in \mathbb{R}^p$ such that $y \leq Cx$ for all $x \in \mathcal{X}$ and there are at least two $x^1, x^2 \in \mathcal{X}_E$ such that $Cx^1 \neq Cx^2$. Since the multi-objective linear programme considered in this paper is (2), it is clear that these assumptions are satisfied. There is a variety of algorithms to solve general MOLPs, see Ehrgott and Wiecek (2005) and references therein. The purpose of this paper is to investigate whether it is possible to make DEA procedures more efficient by applying MOLP algorithms. In particular, we are interested in the outer approximation algorithm of Benson (1998) and an improved version as well as a dual version proposed in (Ehrgott, Löhne, and Shao 2012). We will present the latter two algorithms in Section 4.

4 Benson's Primal and Dual Objective Space Algorithms

The version of Benson's Algorithm presented in Ehrgott, Löhne, and Shao (2012), like the original algorithm of Benson (1998), solves MOLP (3) in objective space by constructing a sequence of polytopes approximating the extended feasible set in objective space $\mathcal{P} := \mathcal{Y} + \mathbb{R}_{\geq}^p$ starting from $\mathcal{S}^0 := y^I + \mathbb{R}^p$. In each iteration the algorithm finds an extreme point y^k of the current polytope \mathcal{S}^k that does not belong to \mathcal{P} and computes the unique point q^k on the boundary of \mathcal{P} between y^k and an interior point \hat{p} of \mathcal{P} . It then constructs a hyperplane supporting \mathcal{P} at y^k by solving an LP. This hyperplane is used to update the approximation \mathcal{S}^k of \mathcal{P} to \mathcal{S}^{k+1} . The algorithm terminates as soon as all extreme points of \mathcal{S}^k belong to \mathcal{P} , in which case $\mathcal{S}^k = \mathcal{P}$ and the extreme points of \mathcal{S}^k are all non-dominated extreme points of \mathcal{Y} , as shown in (Ehrgott, Löhne, and Shao 2012). In order to describe the steps of the algorithm we define the dual pair of linear programmes

$$\begin{aligned} P_1(y) \quad & \min \{z : Ax \geq b, Cx - ez \leq y\}, \\ D_1(y) \quad & \max \{b^T u - y^T w : A^T u - C^T w = 0, e^T w = 1, u, w \geq 0\}. \end{aligned}$$

$P_1(y)$ has optimal value 0 if and only if y is a (weakly) non-dominated point of MOLP (3). In that case, an optimal solution (u^*, w^*) of $D_1(y)$ defines a supporting hyperplane $w^{*T}y = b^T u^*$ of \mathcal{P} at y .

Algorithm 1 (Benson's Algorithm)

- Init:** Compute $\hat{p} \in \text{int}\mathcal{P}$ and $y^I \in \mathbb{R}^p$ and define $\mathcal{S}^0 := y^I + \mathbb{R}^p \supset \mathcal{P}$.
Store vertex set $\text{vert}(\mathcal{S}^0) := \{y^I\}$, set $k = 0$ and go to **It k1**.
- It k1:** If $\text{vert}(\mathcal{S}^k) \subset \mathcal{P}$ go to **It k5**, otherwise choose $y^k \in \text{vert}(\mathcal{S}^k) \setminus \mathcal{P}$.
- It k2:** Find $0 < \alpha_k < 1$ such that $\alpha_k y^k + (1 - \alpha_k)\hat{p} \in \text{bd } \mathcal{P}$.
Set $q^k = \alpha_k y^k + (1 - \alpha_k)\hat{p}$.
- It k3:** Set $\mathcal{S}^{k+1} = \mathcal{S}^k \cap \{y \in \mathbb{R}^p : (w^k)^T y \geq b^T u^k\}$,
where (u^{kT}, w^{kT}) is an optimal solution to $D_1(q^k)$.
- It k4:** Find $\text{vert}(\mathcal{S}^{k+1})$, set $k = k + 1$ and go to **It k1**.
- It k5:** $\mathcal{Y}_N = \mathcal{P}_N = \mathcal{S}_N^k$ and $\mathcal{Y}_{NE} = \text{vert}(\mathcal{S}^k)$.

We illustrate the steps of Algorithm 1 with a small example. Let

$$C = \begin{pmatrix} 3 & 1 \\ -1 & -2 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & -1 \\ -3 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -3 \\ -6 \\ 0 \\ 0 \end{pmatrix}$$

be the data of an MOLP.

Figure 2 shows the iterations of Algorithm 1 on the example. The leftmost plot shows the feasible set \mathcal{Y} in objective space and its extension \mathcal{P} . The ideal point and the initial approximation \mathcal{S}^0 as well as the interior point \hat{p} are added in the second plot. The third plot shows the line connecting y^I and \hat{p} as a broken line. It defines a boundary point q^1 at which solving $D_1(q^1)$ finds a supporting hyperplane to \mathcal{P} . Adding the hyperplane to the description of \mathcal{S}^1 defines two new extreme points, only one of which does not belong to \mathcal{P} . Hence the second iteration (fourth plot) chooses this point and finds the second supporting hyperplane. No further infeasible extreme point is generated, hence the algorithm stops.

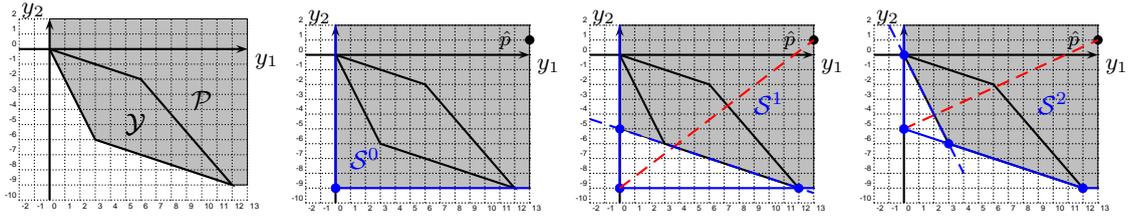


Figure 2: The iterations of the primal algorithm.

To explain the dual variant of the algorithm, we first need to summarise the main concepts of geometric duality for MOLPs as developed by Heyde and Löhne (2008). Geometric duality associates a dual MOLP to every MOLP of the form (3). This is a multi-objective optimisation problem in which maximisation is defined with respect to cone $\mathcal{K} := \mathbb{R}_{\geq} e^p = \{y \in \mathbb{R}^p : y_1 = \dots = y_{p-1} = 0, y_p \geq 0\}$. The dual MOLP is defined by (4),

$$\max_{\mathcal{K}} \{D(u, \lambda) : (u, \lambda) \in \mathbb{R}^m \times \mathbb{R}^p, (u, \lambda) \geq 0, A^T u = C^T \lambda, e^T \lambda = 1\}, \quad (4)$$

where $D(u, \lambda) := (\lambda_1, \dots, \lambda_{p-1}, b^T u)^T$ is the \mathbb{R}^p -valued linear objective function. In analogy to $\mathcal{P} := C(\mathcal{X}) + \mathbb{R}_{\geq}^p$ we define $\mathcal{D} := D(\mathcal{U}) - \mathcal{K}$, where \mathcal{U} is the feasible set of the dual MOLP (4).

For MOLP data

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 3 \\ 4 \\ 0 \\ 0 \end{pmatrix},$$

Figure 3 illustrates both \mathcal{P} and \mathcal{D} .

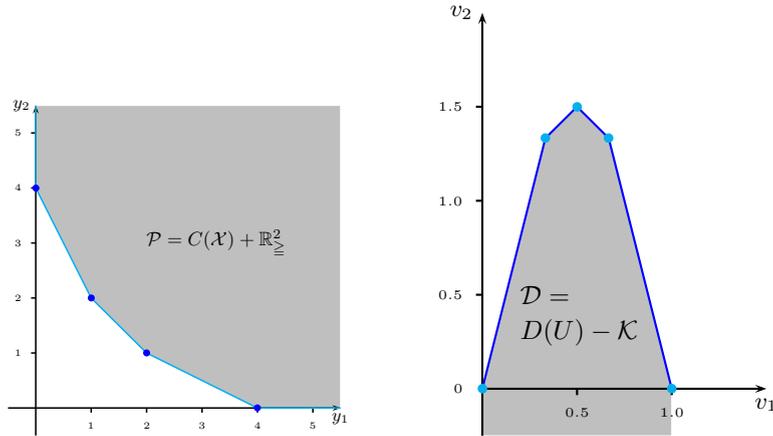


Figure 3: The extended primal and dual feasible sets \mathcal{P} and \mathcal{D} .

The relationships between \mathcal{P} and \mathcal{D} , are described by a coupling function $\varphi(y, v)$ with $y, v \in \mathbb{R}^p$,

$$\varphi(y, v) := \sum_{i=1}^{p-1} y_i v_i + y_p \left(1 - \sum_{i=1}^{p-1} v_i \right) - v_p. \quad (5)$$

We observe that for $x \in \mathcal{X}$ and $(u, \lambda) \in \mathcal{U}$ it holds that $\varphi(Cx, D(u, \lambda)) = \lambda^T Cx - b^T u$, i.e. the duality gap between the objective values of a weighted sum scalarisation of MOLP (3) and its dual.

Defining

$$\begin{aligned}\lambda(v) &:= \left(v_1, \dots, v_{p-1}, 1 - \sum_{i=1}^{p-1} v_i \right)^T, \\ \lambda^*(y) &:= (y_1 - y_p, \dots, y_{p-1} - y_p, -1)^T, \\ H(v) &:= \{y \in \mathbb{R}^p : \lambda(v)^T y = v_p\}, \\ H^*(y) &:= \{v \in \mathbb{R}^p : \lambda^*(y)^T v = -y_p\},\end{aligned}$$

we can now state the geometric duality result for $\mathcal{F}^* \subset \mathcal{D}$ and $\Psi(\mathcal{F}^*) := \bigcap_{v \in \mathcal{F}^*} H(v) \cap \mathcal{P}$.

Theorem 3 ((Heyde and Löhne 2008)) *Ψ is an inclusion reversing one-to-one map between the set of all proper \mathcal{K} -maximal faces of \mathcal{D} and the set of all proper weakly non-dominated faces of \mathcal{P} and the inverse map is given by*

$$\Psi^{-1}(\mathcal{F}) = \bigcap_{y \in \mathcal{F}} H^*(y) \cap \mathcal{D}.$$

Moreover, for every \mathcal{K} -maximal face \mathcal{F}^* of \mathcal{D} it holds that $\dim \mathcal{F}^* + \dim \Psi(\mathcal{F}^*) = p - 1$.

According to Theorem 3, in Figure 3, the four extreme points of \mathcal{P} correspond, via geometric duality, to the four facets of \mathcal{D} and the five extreme points of \mathcal{D} correspond to the five facets of \mathcal{P} . Theorem (3) forms the basis of the dual variant of Algorithm 1. The idea of the dual algorithm is to perform an outer approximation of \mathcal{D} following the same ideas as in the primal algorithm. Details of the algorithm, shown as Algorithm 2, are, however, slightly different. The pair of linear programmes needed is

$$\begin{aligned}P_2(v) & \quad \min \{ \lambda(v)^T Cx : x \in \mathbb{R}^n, Ax \geq b \}, \\ D_2(v) & \quad \max \{ b^T u : u \in \mathbb{R}^m, u \geq 0, A^T u = C^T \lambda(v) \}.\end{aligned}$$

$P_2(v)$ is a weighted sum version of the primal MOLP (3), where the weight vector λ is defined by a point v in \mathcal{D} . Its optimal solution together with the function φ defines a supporting hyperplane of \mathcal{D} at the intersection of the line connecting infeasible extreme point s^k with interior point \hat{d} of \mathcal{D} .

Algorithm 2 (Dual Variant of Benson's Algorithm)

- Init:** For $\hat{d} \in \text{int } \mathcal{D}$ find optimal solution x^0 of $P_2(\hat{d})$.
Set $\mathcal{S}^0 := \{v \in \mathbb{R}^p : \lambda(v) \geq 0, \varphi(Cx^0, v) \geq 0\}$; $k := 1$.
- It k1:** If $\text{vert}(\mathcal{S}^{k-1}) \subset \mathcal{D}$ stop, otherwise choose $s^k \in \text{vert}(\mathcal{S}^{k-1}) \setminus \mathcal{D}$.
- It k2:** Find α^k with $v^k := \alpha^k s^k + (1 - \alpha^k) \hat{d} \in \max_{\mathcal{K}} \mathcal{D}$.
- It k3:** Compute an optimal solution x^k of $P_2(v^k)$.
- It k4:** Set $\mathcal{S}^k := \mathcal{S}^{k-1} \cap \{v \in \mathbb{R}^p : \varphi(Cx^k, v) \geq 0\}$.
- It k5:** Set $k := k + 1$ and go to **It k1**.

We illustrate the steps of Algorithm 2 for the same data used in Figure 3. The first plot in Figure 4 shows the initial approximation \mathcal{S}^0 and interior point \hat{d} . The light gray line is of course the boundary of \mathcal{D} . An infeasible extreme point s^1 of \mathcal{S}^0 is chosen, the intersection point v^k of the line connecting s^1 with \hat{d} is found and a supporting hyperplane of \mathcal{D} at v^k is constructed by solving $P_2(v^k)$. Adding this hyperplane to the description of \mathcal{S}^0 identifies two extreme points of \mathcal{S}^1 . The remaining plots show iterations 2 to 4 of the algorithm. At completion of iteration 4 there are no further infeasible extreme points and the algorithm stops.

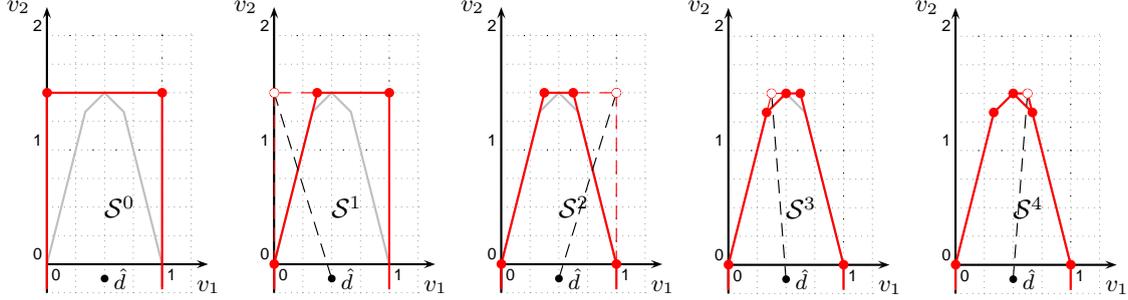


Figure 4: The iterations of the dual algorithm.

5 Applying the Primal and Dual Algorithms to DEA

We are now ready to apply Algorithms 1 and 2 to the MOLP formulation (2) of DEA. To that end, we first write down the very special matrix structure of C , A and b for this problem, which is given by

$$C = \begin{bmatrix} I_{m \times m} & 0_{m \times s} & 0_{m \times n} \\ 0_{s \times m} & -I_{s \times s} & 0_{s \times n} \end{bmatrix}, \quad A = \begin{bmatrix} I_{m \times m} & 0_{m \times s} & -X_{m \times n} \\ 0_{s \times m} & -I_{s \times s} & Y_{s \times n} \\ 0_{1 \times m} & 0_{1 \times s} & 1_{1 \times n} \\ 0_{1 \times m} & 0_{1 \times s} & -1_{1 \times n} \\ 0_{n \times m} & 0_{n \times s} & I_{n \times n} \end{bmatrix}, \quad b = \begin{bmatrix} 0_{m+s \times 1} \\ 1 \\ -1 \\ 0_{n \times 1} \end{bmatrix}.$$

Exploiting this structure, it turns out that the dual MOLP (4) of (2) is

$$\begin{aligned} \max_{\mathcal{K}} \quad & \lambda_1, \dots, \lambda_{m+s-1}, u_{m+s+1} - u_{m+s+2} & (6) \\ \text{s.t.} \quad & u_i = \lambda_i \quad i = 1, \dots, m+s, \\ - \sum_{i=1}^m u_i x_i^j + \sum_{r=1}^s u_{m+r} y_r^j + u_{m+s+1} - u_{m+s+2} + u_{m+s+j} & = 0 \quad j = 1, \dots, n, \\ & \sum_{i=1}^{m+s} \lambda_i = 1, \\ & u, \lambda \geq 0. \end{aligned}$$

Setting $u_0 := u_{m+s+1} - u_{m+s+2}$, removing u_{m+s+j} for $j = 1, \dots, n$ and substituting

λ_i with u_i for $i = 1, \dots, m + s$ this becomes

$$\begin{aligned} & \max_{\mathcal{K}} u_1, \dots, u_m, u_{m+1}, \dots, u_{m+s-1}, u_0 & (7) \\ \text{s.t.} \quad & - \sum_{i=1}^m u_i x_i^j + \sum_{r=1}^s u_{m+r} y_r^j + u_0 \leq 0 \quad j = 1, \dots, n, \\ & \sum_{i=1}^{m+s} u_i = 1, \\ & u \geq 0. \end{aligned}$$

We now look at the details of the primal algorithm. We first need to find the ideal point. In DEA this is a hypothetical DMU composed of the lowest input and highest output values among all existing DMUs, i.e. $y_i^I := \min\{x_i^j : j = 1, \dots, n\}$ for $i = 1, \dots, m$ and $y_i^I := \max\{y_i^j : j = 1, \dots, n\}$ for $i = m + 1, \dots, s$. Next, we need an interior point \hat{p} of \mathcal{P} . In DEA we can define this point by the hypothetical DMU made up of the highest inputs and lowest outputs, i.e. $\hat{p}_i := \max\{x_i^j : j = 1, \dots, n\}$ for $i = 1, \dots, m$ and $\hat{p}_i := \min\{y_i^j : j = 1, \dots, n\}$ for $i = m + 1, \dots, s$. In each iteration, some vertex $s^k \in \mathcal{S}^k \setminus \mathcal{P}$ has to be selected. According to Theorem 2, this can be done by choosing any vertex of $\mathcal{S}^k \setminus \{(x^j, -y^j) : j = 1, \dots, n\}$. Then a boundary point q^k of \mathcal{P} is needed by solving the optimisation problem

$$\max \{ \alpha : x \in \mathcal{X}, \alpha s^k + (1 - \alpha) d^N \geq Cx \}, \quad (8)$$

which is done by solving an LP. It now remains to solve the LP $D_1(v)$

$$\begin{aligned} & \max - \sum_{i=1}^m u_i x_i - \sum_{r=1}^s u_{m+r} y_r + u_0 & (9) \\ \text{s.t.} \quad & - \sum_{i=1}^m u_i x_i^j + \sum_{r=1}^s u_{m+r} y_r^j + u_0 \leq 0 \quad j = 1, \dots, n, \\ & \sum_{i=1}^{m+s} u_i = 1, \\ & u \geq 0. \end{aligned}$$

which cannot be simplified further. Hence, in each iteration it is necessary to solve two linear programmes, one for finding a boundary point and one for finding a supporting hyperplane. At the end of the algorithms we know all non-dominated extreme points of \mathcal{Y} , hence with Theorem 1 all efficient DMUs. Moreover, we obtain an equation for every facet defining the non-dominated set, hence we have a complete description of the DMU efficient frontier of the production possibility set T_{Λ_V} . This allows us to compute any other quantities of interest in DEA, such as the efficiency scores of all inefficient DMUs as well as their targets and peers.

Finally, let us look at the details of the dual Algorithm 2. To start with, define $\gamma := \min\{-y_r^j : r = 1, \dots, s, j = 1, \dots, n\}$ as the lowest of any outputs of any DMU. Clearly, $\gamma - 1$ is below any convex combination of the entries of (x^j, y^j) for any DMU. Hence by definition of \mathcal{D} , $\hat{d} := (\frac{1}{m+s}, \dots, \frac{1}{m+s}, \gamma - 1)$ is an interior point of \mathcal{D} . The

main step of the algorithm requires the solution of $P_2(v)$, i.e.

$$\begin{aligned}
\min \quad & \sum_{i=1}^m v_i x_i - \sum_{r=1}^{s-1} v_{m+r} y_r - \left(1 - \sum_{i=1}^{m+s-1} v_i\right) y_s \\
\text{s.t.} \quad & - \sum_{j=1}^n \mu_j x_i^j + x_i \geq 0 \quad i = 1, \dots, m, \\
& \sum_{j=1}^n \mu_j y_r^j - y_r \geq 0 \quad r = 1, \dots, s, \\
& \sum_{j=1}^n \mu_j = 1, \\
& \mu \geq 0.
\end{aligned} \tag{10}$$

It is not difficult to see that this can actually be done without solving any LP.

Theorem 4 *Given $(x^j, y^j) \in \mathbb{R}^p$ for $j = 1, \dots, n$ and $v \in \mathbb{R}^p$, let*

$$\begin{aligned}
u(j) &:= \left\{ \sum_{i=1}^m v_i x_i^j - \sum_{r=1}^{s-1} v_{m+r} y_r^j - \left(1 - \sum_{i=1}^{m+s-1} v_i\right) y_s^j \right\}, \\
j^*(v) &:= \operatorname{argmin}\{u(j) : j := 1, \dots, n\}, \\
u_0^*(v) &:= u(j^*(v)).
\end{aligned}$$

Vector $(x^{j^(v)}, y^{j^*(v)}, e^{j^*(v)})$ is an optimal solution of $P_2(v)$ in (10), where $j^*(v)$ is an existing DMU and the optimal value of $P_2(v)$ is equal to $u_0^*(v)$.*

The remaining step is to find a vertex $s^k \in \mathcal{S}^k \setminus \mathcal{D}$. To do that we can once again exploit Theorem 4 and the fact that $s^k \in \mathcal{D}$ if and only if $s_{m+s}^k \leq u_0^*(s^k)$, the optimal value of $P_2(s^k)$, which is a consequence of the geometric duality relationship between \mathcal{P} and \mathcal{D} . Otherwise $v^k := (s_1^k, \dots, s_{m+s-1}^k, u_0^*(s^k))$ is the boundary point. Since Algorithm 2 delivers the same results as Algorithm 1, namely all non-dominated extreme points of \mathcal{P} and equations for all facets of \mathcal{P} , all DEA related comments made for Algorithm 1 apply to Algorithm 2, too. The significance of what we have shown is that applying Algorithm 2 to DEA does not require the solution of any linear programmes: *DEA without linear programming*.

6 Numerical Results

To conclude the paper we present some numerical results. The first three columns of Table 1 shows the number of DMUs, inputs, and outputs for 13 instances. Note that line 1 refers to the data plotted in Figure 1. We solved all instances using the traditional DEA method of solving one LP for every DMU (DEA), the primal Algorithm 1 (PBDEA) and the dual Algorithm 2 (DBDEA). All algorithms were implemented in MATLAB using CPLEX as LP solver. The computation times and the number of LPs solved are reported in Table 1. The standard approach was only able to solve the four smallest instances within one hour. The primal algorithm could handle six problems in that time. Clearly, the dual algorithm, which solved all instances in less than ten minutes and all but one in less than 30 seconds, far outperforms the other methods.

Name	Number			Time(s)			Number of LPs		
	DMU	In	Out	DEA	PBDEA	DBDEA	DEA	PBDEA	DBDEA
dea_1	1000	1	1	3760.00	711.45	2.22	1000	8	0
PE1	8	2	1	0.02	0.04	0.02	8	10	0
PE2	4	2	1	0.01	0.02	0.02	4	8	0
N1 2i1o	500	2	1	3418.00	1669.00	0.29	500	156	0
N2 2i1o	500	2	1	–	3694.00	0.74	500	277	0
N3 2i1o	500	2	1	–	1509.00	0.09	500	121	0
N1 2i2o	500	2	2	–	–	0.44	500	–	0
N2 2i2o	500	2	2	–	–	1.66	500	–	0
N3 2i2o	500	2	2	–	–	0.84	500	–	0
N1 2i3o	500	2	3	–	–	14.19	500	–	0
N2 2i3o	500	2	3	–	–	23.26	500	–	0
N3 2i3o	500	2	3	–	–	14.08	500	–	0
N3 3i3o	500	3	3	–	–	439.80	500	–	0

Table 1: Run times and number of LPs solved for three algorithms.

7 Conclusion

In this paper we have considered an MOLP formulation of DEA and applied outcome space algorithms to solve the resulting MOLP. We have shown that the special structure of the MOLP formulation allows significant simplification of the steps of these algorithms. The primal algorithm can be expected to solve approximately twice as many LPs as there are efficient facets rather than one LP for every DMU. Moreover, we have shown that the dual variant of Benson’s algorithms eliminates the need to solve any linear programmes in order to compute all efficient DMUs as well as the complete efficient production frontier. In the future, we plan to develop open source software based on our findings, which we hope will become a useful resource for practitioners and researchers of data envelopment analysis. We will also extend our computational results to include large reference instances from the literature.

References

- Banker, R. D., A. Charnes, and W. W. Cooper. 1984. “Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis.” *Management Science* 30 (9): 1078–1092.
- Benson, H. P. 1998. “An Outer Approximation Algorithm for Generating All Efficient Extreme Points in the Outcome Set of a Multiple Objective Linear Programming Problem.” *Journal of Global Optimization* 13 (1): 1–24.
- Charnes, W.W. Cooper, and E. Rhodes. 1978. “Measuring the Efficiency of Decision Making Units.” *European Journal of Operational Research* 2 (6): 429–444.
- Ehrgott, M., A. Löhne, and L. Shao. 2012. “A dual variant of Benson’s “outer approximation algorithm” for multiobjective linear programming.” *Journal of Global Optimization* 52 (4): 757–778.
- Ehrgott, M., and M.M. Wiecek. 2005. “Multiobjective programming.” Chapter 17 of *Multicriteria Decision Analysis: State of the Art Surveys*, edited by J. Figueira, S. Greco, and M. Ehrgott, Volume 78 of *International Series in Operations Research & Management Science*, 667–722. Springer New York.
- Heyde, F., and A. Löhne. 2008. “Geometric duality in multiple objective linear programming.” *SIAM Journal on Optimization* 19 (2): 836–845.
- Lotfi, F. Hosseinzadeh, A. A. Noora, G. R. Jahanshahloo, J. Jablonsky, M. R. Mozaffari, and J. Gerami. 2008. “An MOLP based procedure for finding efficient units in DEA models.” *Central European Journal of Operations Research* 17 (1): 1–11.

Future Focused Network Modelling at New Zealand Post

Michelle Goodall
New Zealand Post
michelle.goodall@nzpost.co.nz

Grant Robinson
New Zealand Post
grant.robinson@nzpost.co.nz

Abstract

Currently in the New Zealand Post network, when a letter is posted, it goes through many processes before it is delivered at a box lobby or by a postie. These processes include origin and destination mail sorting centres, either a machine or manual sort process, and a transportation network. Other than adding automation, this model has remained essentially unchanged for 120 years.

Over the last 10 years along with other postal organisations around the world, NZ Post has seen significant changes to customer behaviour driven partly by technological advances and the pressures caused by the recession. The result has been a reduction to letter volumes and an increase in parcel volumes processed through the network.

With postal volumes and products changing, New Zealand Post needs to better understand the optimal configuration of sites and machines to assist with the planning of the physical postal network over the next 5-10 years.

The business has applied a hill-climbing algorithm to a mathematical representation of the postal network, enabling the optimal shape of the network under the forecast scenarios to be determined.

1 Introduction

From the moment a letter is posted, to the point it is delivered, it goes through many processes, including those at an origin mail centre, and those at the destination mail centre. The current postal network of mail centres has grown organically over the last 120 years, and with significant changes to the business environment, New Zealand Post as a business has recognised that it may not be able to remain the same in the future.

An optimisation model was created to enable the understanding of an optimal network configuration under different scenarios. The model calculates the flow of mail through the network from origin to destination, including both machine and manual processing streams. Using a heuristic approach, a lowest cost solution is identified, balancing transport, property, resource and machine costs.

The modelling has allowed senior leaders the visibility of future network requirements under different mail volume and product mix segmentations, and is enabling robust and mathematically-based future-focused decision making.

2 The Postal Network

2.1 Postcodes

Geographically, New Zealand is defined by approximately 1800 urban delivery, PO Box and Private Bag, and Rural Delivery postcodes. The primary purpose of these postcodes is to assist in the efficient and accurate delivery of mail. For modelling purposes, postcodes have been used as the lowest geographical breakdown of the country, for both the origin and destination of each mail item.

2113 (Papakura, Auckland) —————> 5028 (Tawa, Wellington)

Figure 1. An example of the origin and destination of a mail item with postcodes.

2.2 The Journey of a Letter

When a mail item is posted into a street receiver (mail box), it will first be picked up by a courier, and taken to the courier hub for consolidation. Cages of mail are then trucked to the Origin Mail Centre. At the mail centre, each bag of mail will be tipped, with machinable mail items going through a Bar Coding Machine (BCM), and all mail that cannot be read or processed by a machine going to manual sort. For various reasons, including adherence to addressing and layout standards, machinable items may be rejected from the BCM, and require a manual sort.

Manual sort is carried out on a 55-hole sorting case, with sorters sorting to a fixed configuration, which at the Origin Mail Centre includes a pigeon hole for all Destination Mail Centres, as well as a pigeon hole for local secondary sort breakdowns and any high volume local Box Lobbies.

Continuing with the Auckland to Wellington example, the mail item will leave Auckland Mail Centre having been through either a BCM or manual sort. It will then be transported on the line-haul network to Wellington Mail Centre.

At the Wellington Mail Centre, the mail item will be sorted to local destination, either through the Bar Code Sorting (BCS) Machine or through the manual sort process. The mail will then be transported locally to the relevant Delivery Branch, where it will be sorted to round order, then delivered by a Postie.

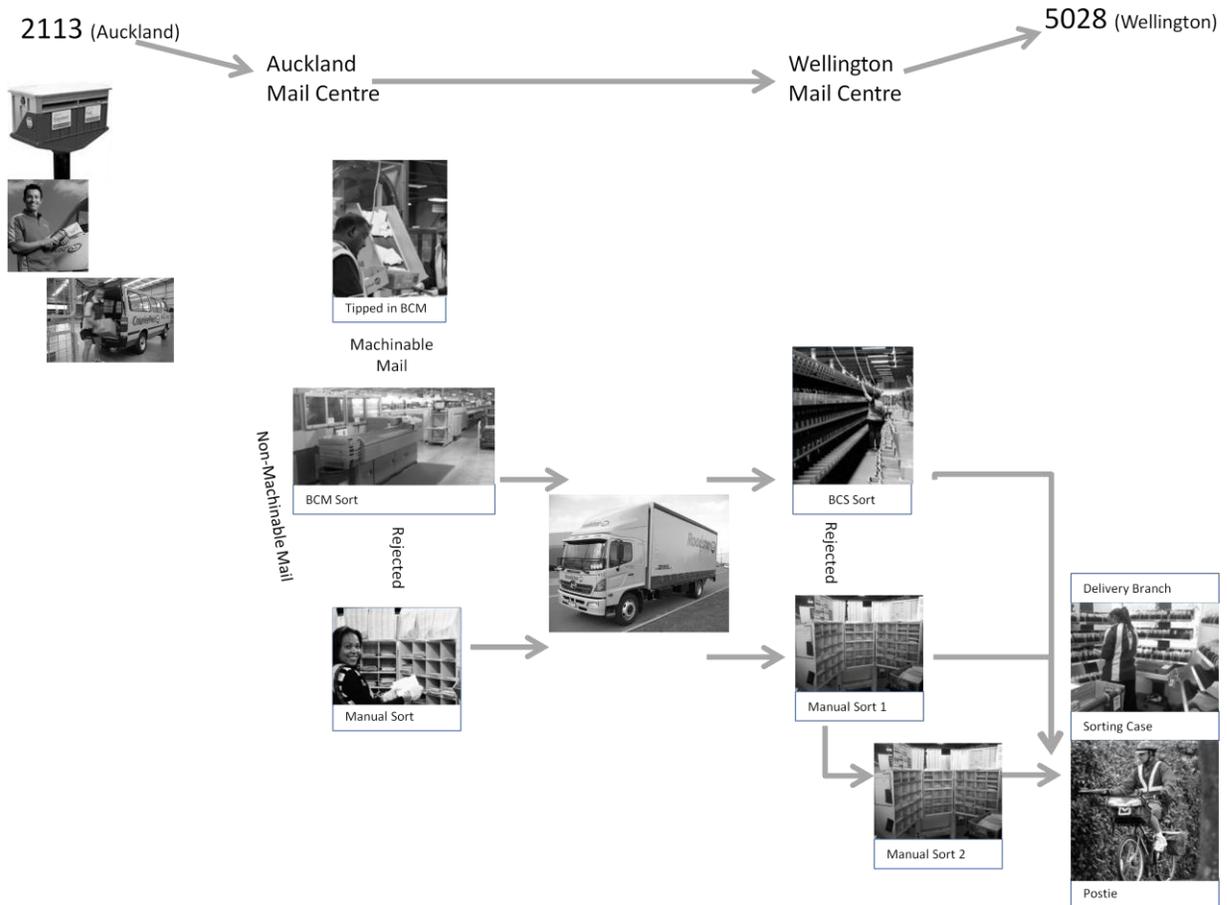


Figure 2 The Journey of a Letter

The New Zealand Post Network that currently consists of six major automated sites, two manual sites and 18 smaller destination sites, has remained relatively unchanged over the last 120 years, apart from the addition of automation into the process.

2.3 The Current Business Climate

Over the past decade, NZ Post's total mail volumes have fallen from 1.1 billion items in 2002, to 834 million items in the 2012 financial year. - Reported on Stuff.co.nz on the 28th of October 2012

New Zealanders are changing the way that they interact with each other. Fewer people are sending letters to each other, and more of us are opting to get bank statements and bills online. At the same time web based businesses such as Trademe and Amazon have increased the number of parcels travelling around the country.

New Zealand Post has spent more than a hundred years growing and changing based on an increasing demand for letters. We now have to challenge some of the fundamental building blocks of the business and consider the future requirements of the network to ensure we remain relevant to our customers.

3 The Model

3.1 Modelling Objectives

Structured as a stream under a larger programme of work, this modelling was set some clear objectives – to design and develop a model capable of enabling the understanding of the optimal configuration of the postal network under different scenarios.

The questions for which the answers were required included:

- Where should mail sorting centres be located, and how many should there be?
- Should each mail sorting centre have mail sorting machines or should they be entirely manual?
- Which postcodes should make up the catchment area of which mail centres?
- What is the difference in cost for an optimised network compared to the current network as volumes and product mix changes?
- How much volume will be processed at each site, and what will the resource requirements be?
- Will we still be able to meet our defined service standards for delivering mail?

With the answers to these questions we will know the optimal shape of the network under the scenarios that we run.

3.2 The Parameters

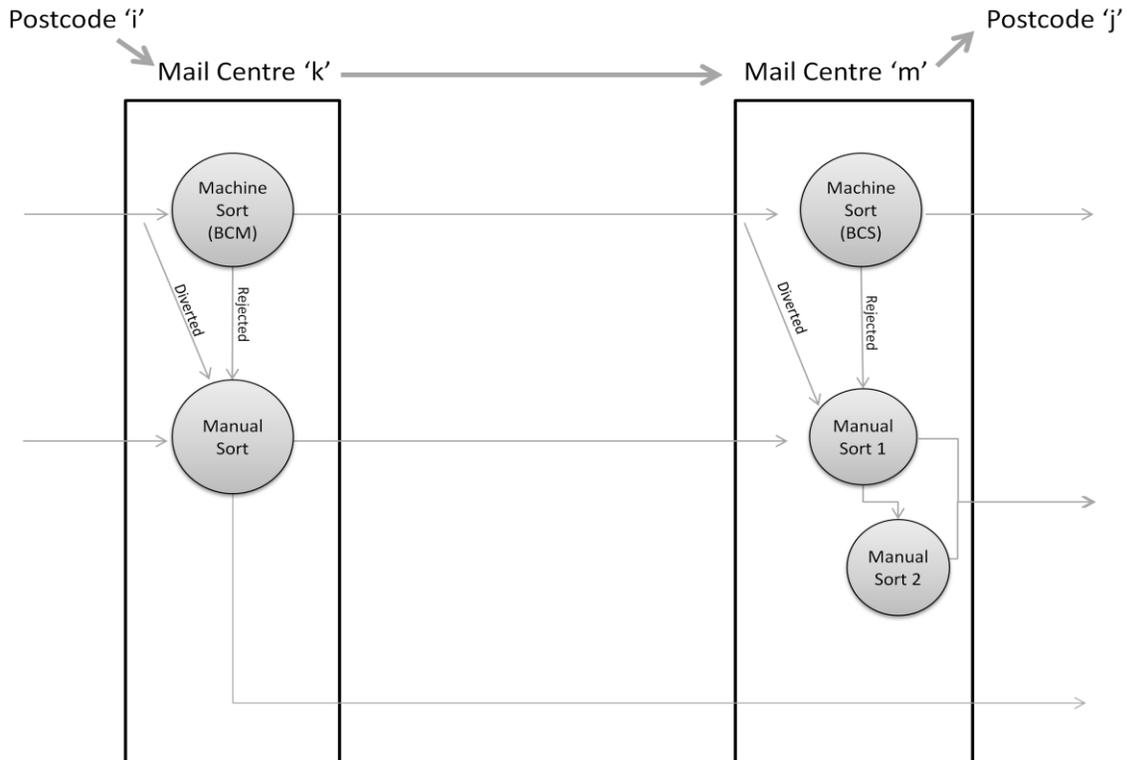


Figure 3. The Journey of a Letter as per the modelling parameters, with the Machine and Manual Sort Relationship

For modelling purposes, a generic ‘journey of a letter’ has been used. Mail from Origin Postcode ‘*i*’, gets processed at Origin Mail Centre ‘*k*’, before being transported to Destination Mail Centre ‘*m*’ and delivered to Destination Postcode ‘*j*’. This modelling includes the allocation of Postcodes to Mail Centres, but does not include any hubbing process before the Origin Mail Centre, or the Delivery process after the Destination Mail Centre.

3.3 The Journey

Under initial scoping, the complexity of this model was severely underestimated. The initial project plan allowed for two months to define, build and validate the full optimisation model, however the further this work progressed, the more understanding was gained of the complexities involved.

The model has been formulated as a quadratically constrained quadratic program, with an objective function that minimises total network cost. Total network cost is broken down into the cost of transport, based on items per kilometre; the cost of property, calculated as a fixed cost, plus a variable cost dependent on the number of machines, and the volume processed within the mail centre; the cost of machines, both the fixed cost of technical support and the variable cost of electricity and parts; and the cost of resource, for both manual processing and the running of machines.

The calculation of each of these costs is based on the volumes in each part of the network, at each step of the process, as per the following diagram.

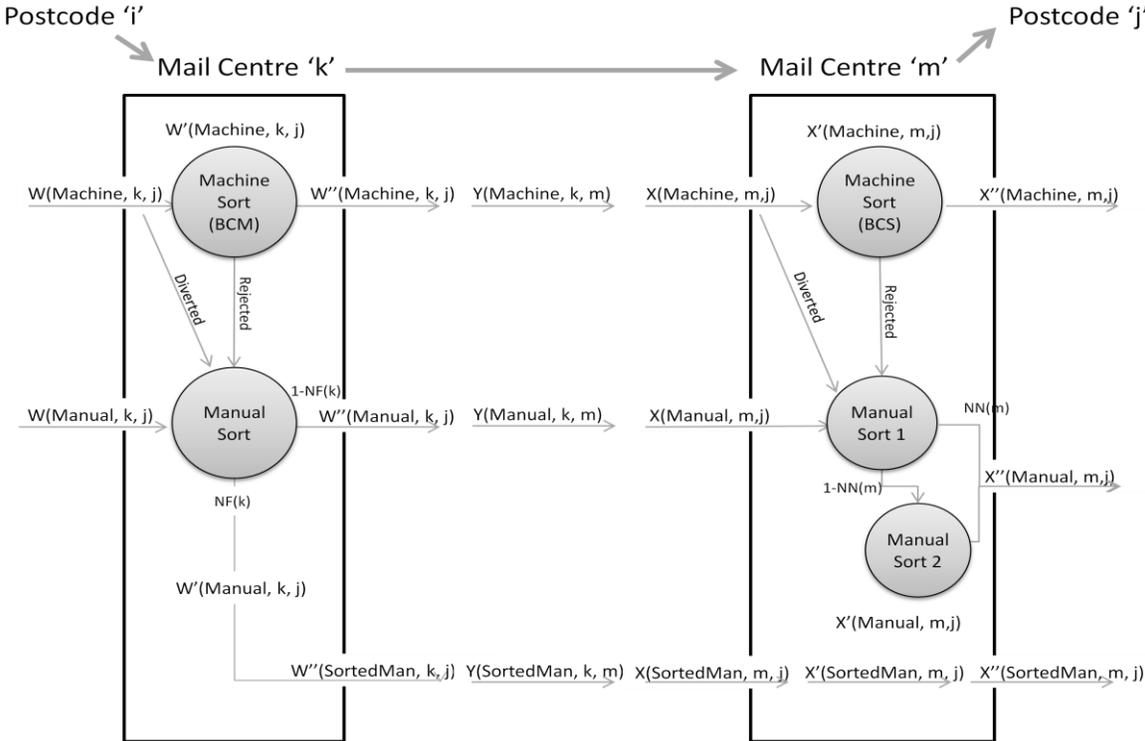


Figure 4 The Journey of a letter as per the modelling parameters.

3.3.1 The challenge of manual sort passes

One of the challenges that arose was around the number of sort passes that a manual item would have to go through at the origin Mail Centre, and at the destination Mail Centre. To demonstrate: A manual sort case has 55 pigeon holes. In Gisborne, there are a total of 48 postie rounds, rural delivery rounds and box lobbies. Therefore all final destinations can fit on the sorting case, and all mail items can be delivered on one pass through a sorting case.

However in Dunedin, there are 83 final destinations, which means that 54 of these 83 final destinations will be sorted on first pass, with all other final destinations being sorted to the remaining pigeon hole, then going on to a second pass through another sort case. To complicate the calculation of the percentage of mail that gets sorted on first pass, the distribution of mail to final destinations is not even - the Dunedin box lobby will receive a higher volume of mail than one postie round in the Dunedin delivery branch.

This continues to grow in complication, as the number of final destinations increases. In Auckland, there are over 1000 final destinations, which are broken into 21 second pass cases. This means that only 33 final destinations in Auckland can be sorted on first pass. However these 33 final destinations include some of the largest Box Lobbies in the country, resulting in around 20% of the Auckland mail being sorted on first pass.

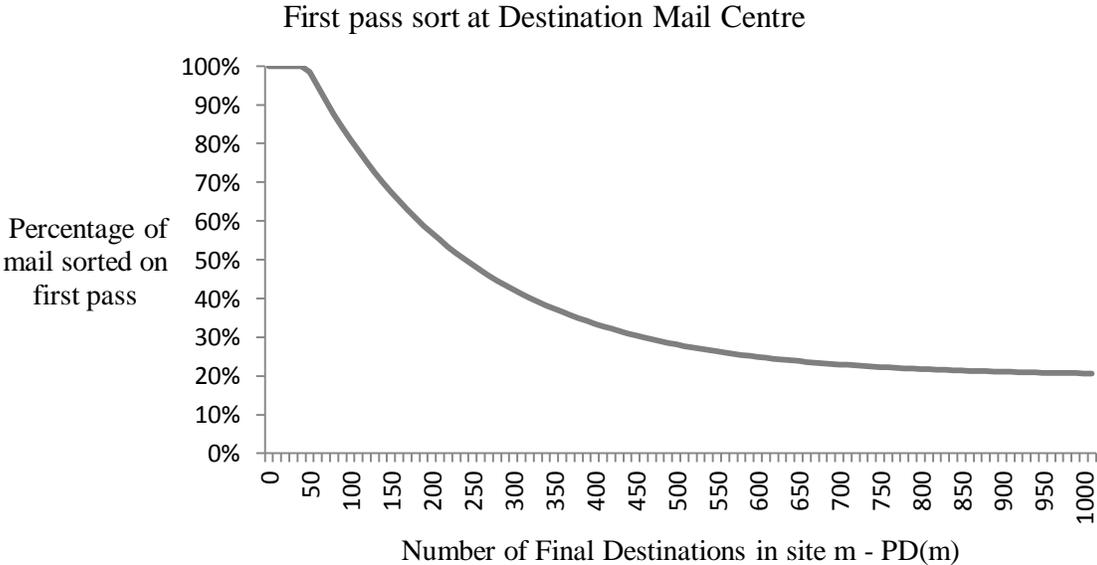


Figure 5 First Pass Destination Sort

The function $NN(m)$ is a piecewise function for which we have used multivariate non-linear regression to approximate the value for when the number of Final Destinations in a destination site is greater than 55.

3.3.2 Solving the Model

Several attempts were made to solve the larger problem, including the Excel based ‘What’s Best’ with nonlinear and Global options, and the solvers CPLEX, MINOS and KNITRO on the back of Ampl coding. Unfortunately due to the size, complexity and non-linearities, none of these solvers could find even an initial feasible solution. At this point, the decision was made to take a heuristic approach, using a basic local search heuristic.

4 Operations Research Solution

4.1 The Algorithm

The algorithm that was used was fairly simple, and followed the following steps:

1. Create an initial feasible solution by assigning each origin postcode to an origin mail centre, and each destination postcode to a destination mail centre.

There were a number of different options used for the assignment of postcodes to mail centres for the initial solution, including random allocation, allocation based on the current network configuration, and the two extreme options of one site in the centre of the country with all postcodes assigned to it, and one site in each postcode.

2. Calculate the cost of sorting all the mail in this current configuration, placing mail sorting machines where there is enough volume to justify it.
3. Look at each mail centre configuration and examine the cost implications if:
 - a. That mail centre acquired a postcode from a neighbouring mail centre
 - b. That mail centre split into two mail centres
 - c. The mail centre moved within the catchment
 - d. That mail centre merged with another mail centre
4. Adopt the change that provides the most favourable improvement and abandon all others
5. Repeat steps 4 & 5 until there are no further cost reductions.

As this progressed we could follow the output, and observed a series of evolutionary steps progressing until the volume in a mail centre justified a mail sorting machine, or it fell below the threshold that would make a mail sorting machine viable.

Total Network Cost across Iterations

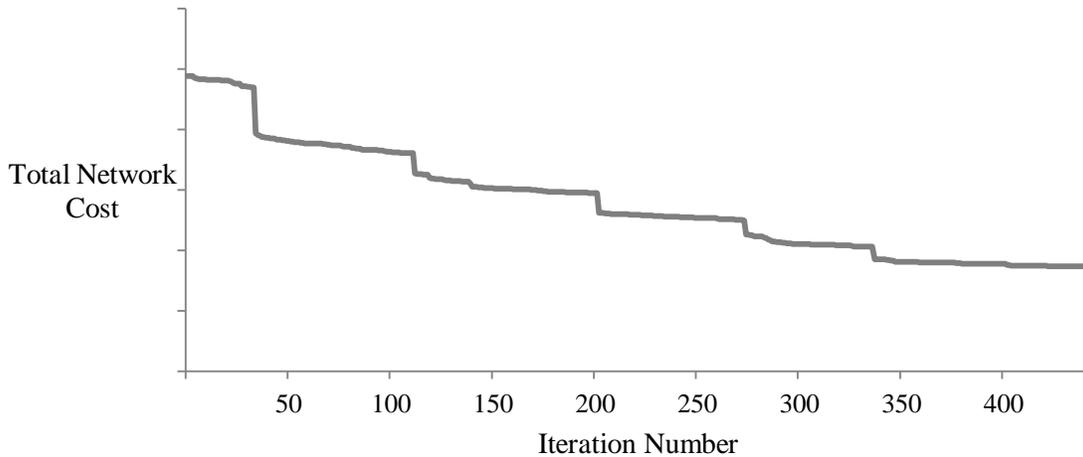


Figure 6 Cost as the Model Progresses

4.2 Output

The modelling produced several interesting outputs, showing the inherent relationships between different factors.

4.2.1 Effect of Automation

One interesting output was that mail sorting machines change the configuration of the network. As mail volumes in any particular mail centre decrease there is a tipping point where it becomes more expensive to keep a machine sorting process. At which point transportation savings overwhelm the labour savings and what was once a large (machine) mail sorting centre handling all of a regions mail suddenly becomes several smaller manual centres.

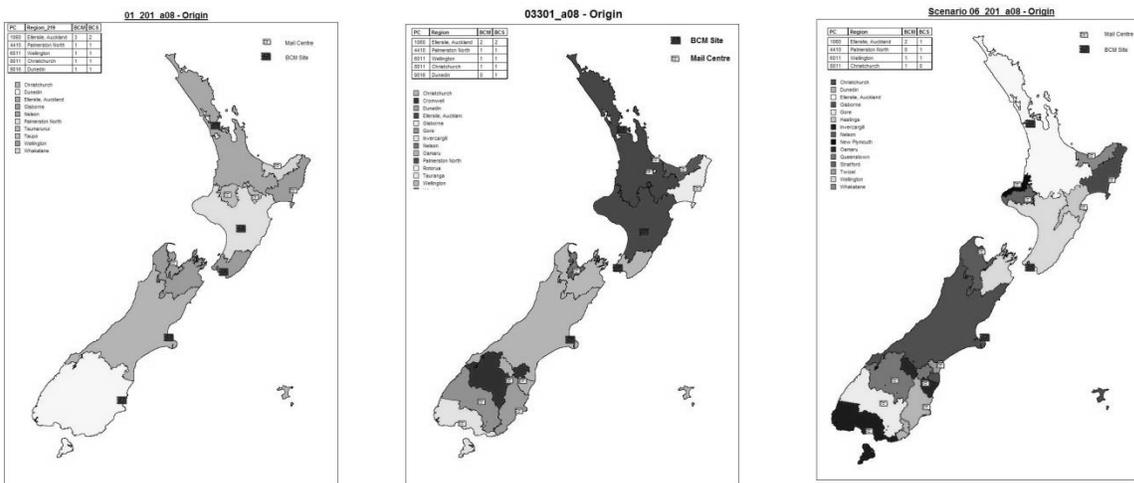


Figure 7. Progressively more fragmented Mail Catchments

4.2.2 Labour and Transport cost interaction

Another output was a tension between the cost of labour and the cost of transport. As the cost of labour increases, the model attempts to centralise the network to make use of the cost savings of mail sorting machines. At the same time, as the cost of transportation increases, the model attempts to push the national transport network out, so that larger, cheaper trucks can be used between mail centres.

4.2.3 Confirmation of current configuration

As well as providing configurations for future scenarios, the model reinforced some of the intuitive design choices which New Zealand Post has made over the years, for example the modelling of our transportation network as a main trunk with several branches, and that for some larger towns which are distant from the main trunk, such as Gisborne and Nelson, it makes sense to have a smaller local mail sorting centre. These have been a part of our network for some time, but before this was modelled, there was no formal mathematical justification for this configuration.

This last point is particularly valuable, because it gives us confidence that our current network is reasonably close to the optimal network at our current volumes.

4.3 Progress to date

The initial modelling occurred last year and the results of several different scenarios were delivered to the business. As with any results the answers have prompted more questions. We have been asked to revisit the modelling, modifying the code so that we can take account of some specific questions, such as:

- We have a proposed new product which we would be processed differently through the network, how would that affect the model?
- New Zealand Post is pioneering a project to pay bills online (<http://www.youpost.co.nz/>), given that this project may reduce the number of bills that move through our network, what will the impact be on the network?
- Fuel has increased out of step with the CPI in recent years, if this continues, what will the impact be on the shape of the network?

Some of the assumptions that were used are being challenged, and these are being explored as areas to develop new products and processes.

The initial model has been modified to enable understanding around this new set of questions from the business, and outputs are being used to provide direction to some of the large programmes of work around New Zealand Post.

5 Conclusion

The optimisation modelling that has been carried out over the last two years has provided New Zealand Post with a robust tool for enabling senior leaders to understand the future required configuration of the network under different volume and product mix scenarios. It

has also given us insight into the underlying tensions and balances which make up the core of the New Zealand Post network. The objectives that were initially set have been achieved, and key process optimisation recommendations have been made.

The business is facing the changing environment and is stepping up to the opportunities that come with it by adopting a rigorous, mathematically-based approach to new situations and scenarios.

The New Zealand Post network is significantly driven by the mail volumes that enter, and flow through it, and as these volumes change over time, it is important that the business understands the optimal configuration of the network. Understanding the costs that drive the business, and the way in which the network configuration impacts on these costs enables New Zealand Post to make effective future-focused decisions.

Using L1-Regression to Estimate a Monotone Two-Piece Linear Relationship Between Two Angular Variables

Petros Hadjicostas

School of Mathematics, Statistics and Operations Research
Victoria University of Wellington
New Zealand
petros.hadjicostas@msor.vuw.ac.nz

Abstract

The Daniels-Guilbaud-Fisher-Lee-Shieh circular rank correlation coefficient can be used to measure the strength of the relationship between two angular variables. Given n pairs of angles between 0 and 2π , the coefficient is invariant to the choice of origin on the unit circle (used for measuring angles). Assuming that the data for one variable are ordered in increasing order, this coefficient equals one (respectively, minus one) if for some $m \leq n$ we can increase the first $m - 1$ data in the first variable by 2π , while keeping the other $n - m + 1$ data fixed, and we can find a strictly increasing (respectively, decreasing) function that interpolates these transformed data. Inspired by this result, we define the relationship between two angular variables to be a perfect monotone two-piece linear association if the coefficient equals ± 1 and the interpolating function is the union of two parallel half-lines. We derive conditions for this to occur. Given n pairs of angles, we then find the best $k \in \{2, \dots, n - 2\}$ so that the first k pairs of data and the last $n - k$ pairs follow linear models and such that the sum of the absolute residuals is minimised subject to the above conditions.

1 Introduction

Let θ and ϕ be two angular variables such as peak times (i.e., times of a 24-hour day that correspond to maximum pressure) of successive measurements of blood pressure, converted into angles; see Downs (1974) and Fisher and Lee (1982,1983). Alternatively we may consider wind directions at two different times in a given day at a weather station; see pp. 149-150 in Fisher (1995). One may then collect data

over a period of n days for the same weather station, or collect data for n weather stations on a single day¹ (but at two different times):

$$(\theta_1, \phi_1), (\theta_2, \phi_2), \dots, (\theta_n, \phi_n). \quad (1)$$

We assume $0 \leq \theta_i, \phi_i < 2\pi$ for $i = 1, 2, \dots, n$.

There are many ways to measure the association between these two angular variables and we refer to the excellent book by Fisher (1995) for more details. We mention, for example, one of the earliest nonparametric correlation coefficients that measure the association between two such variables, the Daniels (1950) circular rank correlation coefficient (valid for $n \geq 3$):

$$r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \binom{n}{3}^{-1} \sum_{1 \leq i < j < k \leq n} \delta_{\theta, \phi}(i, j, k), \quad (2)$$

where

$$\begin{aligned} \delta_{\theta, \phi}(i, j, k) := & \operatorname{sgn}(\theta_i - \theta_j) \operatorname{sgn}(\theta_j - \theta_k) \operatorname{sgn}(\theta_k - \theta_i) \\ & \times \operatorname{sgn}(\phi_i - \phi_j) \operatorname{sgn}(\phi_j - \phi_k) \operatorname{sgn}(\phi_k - \phi_i), \end{aligned}$$

with $\operatorname{sgn}(x) = 1$ if $x > 0$, -1 if $x < 0$, and 0 if $x = 0$. (Here $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ and $\boldsymbol{\phi} := (\phi_1, \dots, \phi_n)^T$ are vectors in \mathbb{R}^n .)

Notice, however, that formula² (2) is due to Fisher and Lee (1982), who have been unaware of Daniels' coefficient (introduced by him thirty years earlier). The equivalence between Daniels' original definition and Fisher and Lee's formula was proven, for example, by Shieh (1990) and Shieh *et al.* (1994, Appendix 4). Daniels (1950), Guilbaud (1980) and Shieh (1990) have shown (independently of one another) that the above circular coefficient can be written as a linear combination of Kendall's tau and Spearman's rho rank correlation coefficients. See also Monjardet (1997, 1998) for a generalization of the above coefficient in more abstract settings.

One of the reasons that the above coefficient is appropriate (to a certain extent) to measure the association between the angular variables θ and ϕ is that it is invariant to the choice of the origin on the unit circle used for measuring angles, i.e.,

$$r_{U,n}[(\boldsymbol{\theta} + x\mathbf{1}_n) \pmod{2\pi}, (\boldsymbol{\phi} + y\mathbf{1}_n) \pmod{2\pi}] = r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi})$$

for any $x, y \in [0, 2\pi)$. (Here $\mathbf{1}_n := (1, 1, \dots, 1)^T \in \mathbb{R}^n$.) Daniels' rank correlation coefficient has been generalized by Jupp (1987) for spherical and hyperspherical data, but in this paper we will concentrate only on circular data.

¹It is important to note here that the angles $\theta_1, \dots, \theta_n$ must be comparable and so must the angles ϕ_1, \dots, ϕ_n . The same reference point must be used for each weather station when collecting the θ_i data, and if someone were to call each of the stations and mention a specific angle, that information (number) should mean exactly the same thing to all the personnels of all the stations.

²Daniels' original formula for $r_{U,n}$ and the given formula by Fisher and Lee essentially ignore possible ties among the θ ranks and among the ϕ ranks. This is not correct. The formulas must be corrected using the theory of Critchlow (1985), but we will not concern ourselves with this problem in this paper.

Note that, for circular data, it is imperative to use *circular* correlation coefficients, not the usual linear ones (Pearson’s or Spearman’s). For example, as Downs (1974) notes, a circular correlation coefficient for peak times of successive measurements of blood pressure, converted into angles, should (ideally) give a value close to one, but a linear one would most probably not give such a value. (Downs (1974) uses parametric correlation coefficients and he compares the circular ones with the linear ones.)

In Cheng and Hadjicostas (2012), the following theorem (that has an elementary proof) is stated:

Theorem 1 *Assume $n \geq 3$ and let $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^T$ be two vectors in \mathbb{R}^n such that $0 \leq \eta_i, \omega_i < 2\pi$ for $i = 1, 2, \dots, n$ with $\omega_i \neq \omega_j$ for $i \neq j$ and $\eta_1 < \eta_2 < \dots < \eta_n$. Then $r_{U,n}(\boldsymbol{\omega}, \boldsymbol{\eta}) = 1$ if and only if there is $m \in \{1, 2, \dots, n\}$ and a strictly increasing function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that*

$$\eta_i = g(\omega_i) \quad \text{for } i = m, m+1, \dots, n; \text{ and} \quad (3)$$

$$2\pi + \eta_i = g(\omega_i) \quad \text{for } i = 1, \dots, m-1. \quad (4)$$

(If $m = 1$, the second set of equalities must be omitted from the statement of this theorem.)

In the previous theorem, in the case $r_{U,n}(\boldsymbol{\omega}, \boldsymbol{\eta}) = -1$, the phrase “strictly increasing” should be replaced with the phrase “strictly decreasing.”

Fisher and Lee (1983) mentioned that a natural way of defining a linear relationship between these two angular variables is to write $\theta \equiv \phi + \alpha_0 \pmod{2\pi}$ for positive association, and $\theta \equiv -\phi + \alpha_0 \pmod{2\pi}$ for negative association, for some arbitrary angle α_0 . They call such a dependence between θ and ϕ as *toroidal-linear*. Mardia (1975, Section 6)—see also, Fisher and Lee (1982)—suggests a more general formula for a possible linear relationship between θ and ϕ :

$$l\theta \pm s\phi + \psi_0 = 0 \pmod{2\pi}, \quad (5)$$

where l and s are positive integers and ψ_0 is an unknown constant angle.

In this paper, however, instead of studying such “linear relationships” (on the torus), we will “define” (what we call) a *perfect monotone two-piece linear relationship* between the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ in $[0, 2\pi)^n$ (rather than between the angular variables θ and ϕ that generated the components of each vector, respectively). Inspired by Theorem 1, we use the following definition in this paper:

Definition 1 *Assume $n \geq 4$ and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^T$ be two vectors in \mathbb{R}^n (corresponding to the first components and the second components, respectively, of the angular data in (1)) such that $\phi_1 \leq \phi_2 \leq \dots \leq \phi_n$ (with $\phi_1 < \phi_n$). These two vectors are said to have a **perfect monotone two-piece linear relation** (with the variable θ as the response and the variable ϕ as a covariate) if there is an integer $k \in \{2, \dots, n-2\}$ and constants $\gamma_0, \delta_0, \beta \in \mathbb{R}$ such that*

$$\theta_i = \gamma_0 + \beta\phi_i \quad \text{for } i = 1, \dots, k;$$

$$\theta_i = \delta_0 + \beta\phi_i \quad \text{for } i = k+1, \dots, n,$$

and (at least) one of the following two conditions holds:

1. $\beta \geq 0$ and either $\beta(\phi_{k+1} - \phi_k) \geq \gamma_0 - \delta_0$ or $\beta(\phi_n - \phi_1) \leq \gamma_0 - \delta_0$.
2. $\beta \leq 0$ and either $\beta(\phi_{k+1} - \phi_k) \leq \gamma_0 - \delta_0$ or $\beta(\phi_n - \phi_1) \geq \gamma_0 - \delta_0$.

It is clear that our idea of “linearity” differs from that of Mardia (1975) and Fisher and Lee (1982, 1983)—see equation (5) in this paper. In the future, it will be interesting to examine the case when β is restricted to be a rational number.

It is also clear from Definition 1 that there is no loss of generality in assuming $\phi_1 \leq \phi_2 \leq \dots \leq \phi_n$ since we can always re-arrange the n pairs in (1) so that these inequalities hold. The following result explains why we need conditions 1 and 2 in the above definition:

Lemma 1 *Assume $n \geq 4$ and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)^T$ be two vectors in \mathbb{R}^n such that $\phi_1 < \phi_2 < \dots < \phi_n$. Assume there is an integer $k \in \{2, \dots, n-2\}$ and constants $\gamma_0, \delta_0, \beta \in \mathbb{R}$ such that $\gamma_0 \neq \delta_0$,*

$$\theta_i = \gamma_0 + \beta\phi_i \quad \text{for } i = 1, \dots, k; \quad \text{and} \quad (6)$$

$$\theta_i = \delta_0 + \beta\phi_i \quad \text{for } i = k+1, \dots, n. \quad (7)$$

Then:

1. $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$ if and only if $\beta > 0$ and either $\beta(\phi_{k+1} - \phi_k) > \gamma_0 - \delta_0$ or $\beta(\phi_n - \phi_1) < \gamma_0 - \delta_0$.
2. $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -1$ if and only if $\beta < 0$ and either $\beta(\phi_{k+1} - \phi_k) < \gamma_0 - \delta_0$ or $\beta(\phi_n - \phi_1) > \gamma_0 - \delta_0$.

Proof. If $\beta = 0$, then $\theta_i = \gamma_0$ for $i = 1, \dots, k$ and $\theta_i = \delta_0$ for $i = k+1, \dots, n$, in which case $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 0$. Thus, we assume $\beta \neq 0$. In addition, if $\theta_i = \theta_j$ for $1 \leq i \neq j \leq n$, then some of the terms $\delta_{\theta,\phi}(i, j, k)$ in the sum in (2) would be zero, and thus (in such a case) $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi})$ can neither be equal to 1 nor to -1 . Hence we assume $\theta_i \neq \theta_j$ for $i \neq j$. We prove only the first claim of the lemma (for the second one can be proven in a similar way).

(a) Assume first $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$. By Theorem 1, there is $m \in \{1, 2, \dots, n\}$ and a strictly increasing function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that equalities (3) and (4) hold with $\eta_i := \phi_i$ and $\omega_i := \theta_i$. This means $\theta_1 < \theta_2 < \dots < \theta_{m-1}$ and $\theta_m < \theta_{m+1} < \dots < \theta_n$. In view of (6) and (7) and the assumption $\phi_1 < \phi_2 < \dots < \phi_n$, these inequalities imply $\beta > 0$. If $\beta(\phi_{k+1} - \phi_k) \leq \gamma_0 - \delta_0$, then $\theta_{k+1} < \theta_k$, i.e., it is not the case that $\theta_1 < \theta_2 < \dots < \theta_n$, which implies $m > 1$.

In the case $m > 1$, if it were true that $\theta_1 < \theta_n$, then $2\pi + \phi_1 = g(\theta_1) < g(\theta_n) = \phi_n$, a contradiction since we consider all angles to be in $[0, 2\pi)$. Thus,

$$\delta_0 + \beta\phi_n = \theta_n < \theta_1 = \gamma_0 + \beta\phi_1, \quad (8)$$

from which we conclude that $\beta(\phi_n - \phi_1) < \gamma_0 - \delta_0$.

(b) Conversely, assume $\beta > 0$. It follows from equations (6) and (7) that $\theta_1 < \theta_2 < \dots < \theta_k$ and $\theta_{k+1} < \dots < \theta_n$.

If $\beta(\phi_{k+1} - \phi_k) > \gamma_0 - \delta_0$, then $\theta_k < \theta_{k+1}$, and thus $\theta_1 < \theta_2 < \dots < \theta_n$. In such a case, we clearly have $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$.

Assume $\beta(\phi_n - \phi_1) < \gamma_0 - \delta_0$. Clearly (8) holds as well, i.e.,

$$\theta_{k+1} < \dots < \theta_n < \theta_1 < \dots < \theta_k.$$

Since also $\phi_1 < \dots < \phi_n$, it is a simple exercise to show that (in this case) $r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$. \square

Remark 1 Despite the strictness of the inequalities for the parameters γ_0 , δ_0 and β in Lemma 1, we have included *limiting cases* in Definition 1 (i.e., we have included the cases $\beta = 0$; $\beta(\phi_{k+1} - \phi_k) = \gamma_0 - \delta_0$; and $\beta(\phi_n - \phi_1) = \gamma_0 - \delta_0$), because we will be performing linear programming in the next section and constraints on the parameters (“decision variables”) must include the equality cases as well. In addition, the case where the n pairs of angular data in (1) are on a straight line (i.e., the case $\gamma_0 = \delta_0$) is also included in Definition 1 because it is also a limiting case. It is clear that when there are ties in the data or when the limiting cases hold, Daniels’ circular rank correlation coefficient may not necessarily equal ± 1 .

2 Estimating a monotone two-piece linear relation using L1-regression

There are many methods in the literature for performing a circular-circular regression (i.e., a simple regression where both the response θ and the covariate ϕ are angular): see, for example, Fisher (1995), Fisher and Lee (1992), Kato *et al.* (2008), and Mardia and Jupp (2000). In this paper, however, we will assume that θ and ϕ have a monotone two-piece linear relationship, given by

$$\begin{aligned} \theta_i &= \gamma_0 + \beta\phi_i + \epsilon_i & \text{for } i = 1, \dots, k; \\ \theta_i &= \delta_0 + \beta\phi_i + \epsilon_i & \text{for } i = k + 1, \dots, n, \end{aligned}$$

for some integer $k \in \{2, \dots, n - 2\}$ (assuming $n \geq 4$). Here $(\epsilon_i : i = 1, \dots, n)$ are random errors, while we also assume that either condition 1 or condition 2 of Definition 1 holds. These conditions impose restrictions (constraints) on the regression parameters γ_0 , δ_0 and β .

It is not our purpose here to compare our elementary model and method of estimation to the existing methods or to perform a statistical analysis (hypothesis testing and confidence intervals for the regression coefficients) by using an assumed distribution on the errors ϵ_i . Our main objective is more of an educational value: to illustrate how linear programming techniques can be used to estimate the parameters γ_0 , δ_0 , and β and how to determine the “best” integer k for the above monotone two-piece linear model between angular variables θ and ϕ .

According to the article “Method of least absolute values” in the *Encyclopedia of Statistical Sciences*—see Harter (1985)—the history of the minimization method

known as *L1-estimation* or *L1-regression* goes back to the Italian/Croatian physicist and astronomer R. J. Boscovitch in the middle of the 18th century and the French scientist P. S. Laplace at the end of the 18th century. In the Operations Research literature, one of the earliest papers that uses linear programming for solving problems of this kind is Charnes *et al.* (1955).

In our optimization program we seek to *minimize* (over all $k \in \{2, \dots, n-2\}$)

$$z_k = \sum_{i=1}^k |\theta_i - \gamma_0 - \beta\phi_i| + \sum_{i=k+1}^n |\theta_i - \delta_0 - \beta\phi_i|$$

subject to at least one of the following conditions:

1A: $\beta \geq 0$ and $\beta(\phi_{k+1} - \phi_k) \geq \gamma_0 - \delta_0$.

1B: $\beta \geq 0$ and $\beta(\phi_n - \phi_1) \leq \gamma_0 - \delta_0$.

2A: $\beta \leq 0$ and $\beta(\phi_{k+1} - \phi_k) \leq \gamma_0 - \delta_0$.

2B: $\beta \leq 0$ and $\beta(\phi_n - \phi_1) \geq \gamma_0 - \delta_0$.

This minimization problem can be solved using standard Linear Programming techniques; e.g., see pp. 221-226 in Chvátal (1983).

Given $n \geq 4$ and $\boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^n$ with $\phi_1 \leq \phi_2 \leq \dots \leq \phi_n$ (and $\phi_n > \phi_1$), the above minimization program has to be run $4(n-3)$ times: four times (once for each condition) for each $k \in \{2, \dots, n-2\}$. We choose the model (i.e., we choose k, γ_0, δ_0 and β) with the smallest sum of absolute residuals, z_k .

It is clear that the above minimization problem always has a feasible solution (just let $\beta = 0$ and $\gamma_0 = \delta_0$). Unfortunately (for given $k \in \{2, \dots, n-2\}$), the minimization problem does not always have a unique solution. If k^* is an integer in $\{2, \dots, n-2\}$ for which the smallest z_k occurs, for each of the four conditions (1A, 1B, 2A, 2B), the set of all optimal solutions $(\gamma_0^*, \delta_0^*, \beta^*)$ corresponding to this k^* most probably forms a *closed bounded convex set* (with possible exceptions in some extreme configurations of the data in (1)). The most appropriate solution, in such a case, is *the center of mass* of this closed bounded convex set (which in general requires the computation of all the vertices of this convex set).

As an example, assume that $n = 6$ and the data are

$$(0, \pi/4), (\pi/4, \pi/2), (2\pi/3, 2\pi/3), (3\pi/4, 5\pi/6), (5\pi/6, \pi/4), (7\pi/8, 2\pi/3). \quad (9)$$

In other words,

$$\boldsymbol{\phi} = (0, \pi/4, 2\pi/3, 3\pi/4, 5\pi/6, 7\pi/8) \quad \text{and} \quad \boldsymbol{\theta} = (\pi/4, \pi/2, 2\pi/3, 5\pi/6, \pi/4, 2\pi/3).$$

The Daniels' circular rank correlation coefficient between these two angular vectors is

$$r_{U,n}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{3}{10}.$$

Table 1 contains the results from running $4(n-3) = 12$ linear programs for the various conditions and the various values of $k \in \{2, 3, 4\}$. We see that the smallest

z_k (i.e., the smallest sum of absolute values of the residuals) corresponds to the case of Condition 1B with $k^* = 4$, $\gamma_0^* = 16\pi/57$, $\delta_0^* = -83\pi/228$ and $\beta^* = 14/19$.

The last column of Table 1 contains the Daniels rank correlation coefficient between the estimated (predicted) vector $\hat{\theta}$ and the covariate vector ϕ . Unless $z_k = 0$ (and we do not have limiting cases), we do not expect $r_{U,n}(\hat{\theta}, \phi)$ to be equal to ± 1 .

Table 1: Minimization results for the data in (9)

Condition	k	z_k	$\hat{\gamma}_0$	$\hat{\delta}_0$	$\hat{\beta}$	$r_{U,n}(\hat{\theta}, \phi)$
1A	2	$5\pi/6$	$\pi/4$	$2\pi/3$	0	0
1B	2	$7\pi/6$	$\pi/2$	$\pi/2$	0	0
2A	2	$7\pi/6$	$\pi/2$	$\pi/2$	0	0
2B	2	$5\pi/6$	$\pi/4$	$2\pi/3$	0	0
1A	3	$155\pi/192$	$\pi/4$	$19\pi/96$	$5/8$	$4/5$
1B	3	$31\pi/30$	$2\pi/5$	$\pi/20$	$2/5$	$4/5$
2A	3	$7\pi/6$	$\pi/2$	$\pi/2$	0	0
2B	3	π	$\pi/2$	$2\pi/3$	0	0
1A	4	$145\pi/192$	$\pi/4$	$19\pi/96$	$5/8$	$4/5$
1B	4	$127\pi/228$	$16\pi/57$	$-83\pi/228$	$14/19$	$4/5$
2A	4	$7\pi/6$	$\pi/2$	$\pi/4$	0	0
2B	4	$7\pi/6$	$\pi/2$	$\pi/2$	0	0

3 Discussion

The method used in this paper assumes that the (θ_i, ϕ_i) data in (1) form at most two almost parallel clusters (bands) of data and the above L1 minimization method estimates the intercepts of the two parallel lines and the common slope. It is clear that the method will be more appropriate if the data have a large Daniels' rank correlation coefficient (in absolute value).

As pointed before, one problem with this estimation method is that the estimates of the parameters γ_0, δ_0 and β are not unique, and once the best model has been identified, one has to use an efficient algorithm to find the vertices of the (convex) solution space (in \mathbb{R}^3) and then find the center of mass of this solution space.

Another problem that might arise by using the above L1-estimation method is that the minimum sum of absolute values of the residuals may (potentially) be achieved when $\beta = 0$, which is *not* a desirable solution to the problem (see the last two columns in Table 1). Further research is needed to examine if and when this problem can occur.

The *most serious problem*, however, with this analysis is that the solutions to the problem are *not* necessarily invariant to the choice of origin on the unit circle

(used) for measuring angles. The method needs to be modified appropriately (if possible!) to take into account this requirement.

References

- Charnes, A., W. W. Cooper, and R. O. Ferguson. 1955. "Optimal estimation of executive compensation by linear programming." *Management Science* 1:138-151.
- Cheng, D. and P. Hadjicostas. 2012. "Right-invariant metrics applied to rank correlation coefficients," submitted for publication.
- Chvátal, V. 1983. *Linear Programming*. W. H. Freeman and Company, New York, NY.
- Critchlow, D. E. 1985. *Metric methods for analyzing partially ranked data*. Springer-Verlag, Berlin.
- Daniels, H. E. 1950. "Rank correlation and population models (with discussion)." *Journal of the Royal Statistical Society—Series B* 12:171-191.
- Downs, T. D. 1974. "Rotational angular correlations," in: *Biorhythms and human reproduction*. (Eds. M. Ferin, F. Halberg, R. Richart, and R. Vande Wiele), Chapter 7, pp. 97-104. Wiley, New York, NY.
- Fisher, N. I. 1995. *Statistical analysis of circular data*. Cambridge University Press, New York, NY.
- Fisher, N. I. and A. J. Lee. 1982. "Nonparametric measures of angular-angular association." *Biometrika* 69:315-321.
- Fisher, N. I. and A. J. Lee. 1983. "A correlation coefficient for circular data." *Biometrika* 70:327-332.
- Fisher, N. I. and A. J. Lee. 1992. "Regression models for an angular response." *Biometrics* 48:665-677.
- Guilbaud, G. Th. 1980. "Relation entre les deux coefficients de corrélation de rangs." *Mathématiques et Sciences Humaines* 72:45-59.
- Harter, H. L. 1985. "Method of least absolute values," in: *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson and C. B. Read, eds.), pp. 462-464, John Wiley and Sons, New York, NY.
- Jupp, P. E. 1987. "A non-parametric correlation coefficient and a two-sample test for random vectors or directions." *Biometrika* 74:887-890.
- Kato, S., K. Shimizu, and G. S. Shieh. 2008. "A circular-circular regression model." *Statistica Sinica* 18:633-645.
- Mardia, K. V. 1975. "Statistics of directional data (with discussion)." *Journal of the Royal Statistical Society, Series B* 37:349-393.
- Mardia K. V. and P. E. Jupp. 2000. *Directional Statistics*, John Wiley and Sons, Chichester.

- Monjardet, B. 1997. "Concordance between two linear orders: The Spearman and Kendall coefficients revisited." *Journal of Classification* 14:269-295.
- Monjardet, B. 1998. "On the comparison of the Spearman and Kendall metrics between linear orders." *Discrete Mathematics* 192:281-292.
- Shieh, S. R. 1990. *Some extensions of U- and V-statistics*. Ph.D. thesis, University of Wisconsin-Madison, Madison, Wisconsin, USA.
- Shieh, G. S., R. A. Johnson, and E. W. Frees. 1994. "Testing independence of bivariate circular data and weighted degenerate U-statistics." *Statistica Sinica* 4:729-747.

Optimizing Clothing Catalogues for EziBuy

Oliver Hinder
Department of Engineering Science
University of Auckland
New Zealand
ohin004@auckland.ac.nz

Abstract

Ezibuy is a company that sells clothing by mailing customers catalogues. An important decision they need to make is which clothing catalogue should be sent to which customers. At the moment this decision is made manually. This project investigates using optimisation to make this decision with the objective of maximizing short term profit. The report is split into two main sections: modelling sales profit and optimisation.

In the section on modelling sales profit, we build a model to explain the catalogue browsing process. We attempt to fit parameters to this model using data from EziBuy, but have little success. Consequently, we suggest a model which only uses historical catalogues.

In the section on optimisation, we consider the problem of choosing which catalogues to send to which customers. The optimisation model require a list of possible catalogues, with corresponding costs and sales predicted by the models. The optimisation model is tested using Ezibuy's 2011 catalogues. We compare the performance of the optimisation model with simple rules of thumb and find that the optimisation yields significantly better results.

1 Introduction

1.1 Background

An effective and common marketing strategy is to separate customers into segments and target each segment with different products and prices. If products are more relevant customers are more likely to respond. If pricing reflects customer budgets it is possible to extract more profit from each customer.

In this project, we are interested in applying these ideas to EziBuy, a New Zealand clothing catalogue company. EziBuy sells clothes through a mail order system. When a customer registers with EziBuy they are sent clothing catalogues. If a customer likes a particular item they have seen in a clothing catalogue they phone EziBuy and purchase the product. The product is then sent to the customer in the mail. EziBuy specialises primarily in woman's clothing and includes a wide range of brands such as Capture, Urban and Essentials. Many of EziBuy's customers have specific brand preferences. For example, there are groups of customers that particularly like European fashion, some customers like Homeware and Home Decor, others like plus size clothing. If customers are sent large catalogues that mostly contain irrelevant items they are likely to lose

interest and will not view the items that they want to buy. Furthermore, large catalogues are expensive to print and send. Therefore it makes sense to send customers smaller targeted catalogues.

A *segment* is a group of customers which have very similar buying preferences. We will send all customers in the same segment the same catalogues. We will use the word *collection* to describe a set of catalogues sent to the customer over a short time interval. We assume customer behaviour over this short interval is independent of events outside of this interval. For example, if this short interval was a week, this implies that whether a customer purchases an item this week is not influenced by the collection they were sent last week. We ignore the impact of other EziBuy advertising channels, including online promotions. The purpose of this project is to maximize the profit made within this short interval.

The printers that EziBuy use require that pages of catalogues are printed in 32 page *sections*. The same section can be reused in different catalogues. The content of each catalogue is defined by these sections. A *part* is an abstract term that refers to anything that is required to be completed before the collection can be made. For example, a part could be the cover page of a catalogue or a section in a catalogue.

An *item* is an article of clothing that is placed in the catalogue for sale to the customer. The *margin* on an item is the sale price minus the cost of producing the item. The *sales profit* is the sum of the margins over all the items sold. The *profit from sending a collection* is the sales profit from sending the collection minus the cost of sending the collection. The *cost of sending a collection* incorporates postage and printing costs, but ignores *fixed costs*.

The fixed costs of producing a collection can be split into *setup costs* and *design costs*. Setup costs are incurred for each different part produced. For example, in order for a section to be produced the printer needs to be setup. *Design costs* are the costs associated with photographing the item on a model and placing the items on the page. Both these costs can be shared between different collections. A *brand* is a set of items whose design cost can be recycled between collections. A solution to this problem chooses which collections to allocate to which customers. The *total profit* of a solution is the total profit from sending the collections minus the design costs minus the setup costs.

An obvious approach to allocating catalogues to customers would be to send each customer a personalised catalogue to extract maximum expenditure from each customer. Unfortunately, this would require significant setup and design costs. Alternatively, one might produce only one catalogue and send it to every customer. Although design costs and setup costs would be low, the profit from sending the collection would be lower. This is because customers are less likely to purchase from a catalogue that contains mostly irrelevant items. Furthermore the collection is more likely to be larger, so the cost of printing each page will be higher. Clearly there exists a solution which is a trade-off between personalising the catalogues and using only one catalogue which gives the maximum profit. This report focuses on finding this optimal solution.

1.2 Prior Research

We are aware of one paper by Steinbach, Karaypis, & Kumar in 2000, that attempts to develop optimisation models to decide which products should be put in catalogues to maximize short term profit. They make very strong assumptions about the nature of catalogue sales which we believe are not robust. Their core assumption is that the profit

of sending a segment a particular item is constant. This means that if we send a customer a catalogue with two shirts then we expect twice the sales than if we only send them a catalogue that contains one shirt. This assumption is poor because it neglects that similar items are substitutes, so adding more of the same item will not increase sales. Additionally, their optimisation model requires that the catalogues produced must be all the same size and ignores any design costs. In this report, we aim to develop optimisation models that do not require these restrictive assumptions.

2 Modeling Sales Profit

The purpose of this section is to discuss possible models used to predict sales profit when different collections are sent to a customer. We develop a model that explains the tendency of smaller catalogues to perform proportionally better than larger catalogues. The parameters of these models are estimated from historical data. Unfortunately, the models do not fit the data well, so we suggest a simple historical model.

2.1 Catalogue Browsing Models

Suppose customers randomly flick through pages in the collection C which is a set of items with cardinality n . They have started the browsing process. The probability they continue the browsing process after viewing a page is w . Each page contains exactly one item. Suppose the expected sales profit of a customer viewing item i is v_i . This yields expected sales profit of:

$$V(C) = E_C[v](1 + w + w^2 + \dots + w^n) = E_S[v] \frac{1 - w^{n+1}}{1 - w}$$

Now suppose we model a collection as a vector of brands (q_1, \dots, q_k) where q_b corresponds to the number of pages of brand b in the collection. Each page contains exactly one item. Suppose that for each brand b comprises of a set of items A_b . Each additional item of brand b is randomly chosen from the remaining items.

Let α_b be the expected sales profit when a customer views a random page of brand b . This implies that the expected sales profit can be expressed as:

$$E_C[v] = \sum_b \alpha_b \frac{q_b}{n} = \sum_b \alpha_b \frac{q_b}{q_1 + \dots + q_k}$$

This model encapsulates the idea that making the catalogue larger will have diminishing marginal returns, because customers will lose interest in the browsing process and not view items they would otherwise purchase. The biggest flaw of this model is that it does not take into account that there will be a substitution effect between similar items. Consequently, it will suggest a collection which comprise only of the brand with the largest α_b .

Before we find parameters for the catalogue browsing model we need to reformulate the model so we can use regression. To do this we assume that error in sales profit is normally distributed with constant variance.

Parameters:

w is the probability a customer exits the browsing process. It measures the amount we penalise large collections.

Let α_b be the expected sales profit when a customer views a random pages from brand b .

Let Y_i we the observed sales profit of collection i .

Let Y_{ib} be the recorded sales profit of collection i accruing from brand b .
Let q_{ib} be the number of pages of brand b in collection i .
Let n_i be the number of pages in the collection i .
 $\varepsilon_i, \varepsilon_{ib}$ which are independent random variables with distribution $N(0,1)$.
 σ, σ_b is the variance corresponding to these errors.

If we directly use the catalogue browsing model we get the following formula for the total sales profit of collection i :

$$Y_i = \frac{1 - w^{n_i}}{n_i(1 - w)} \sum_b \alpha_b q_{ib} + \sigma \varepsilon_i \quad (1)$$

The disadvantage this formulation is that there is a large number of brands and a relatively small number of catalogues available. This means that we may over fit the models to the data. So instead of directly fitting the total sales profit we could take advantage of the fact that the catalogue browsing model implies that the sales profit for a particular brand b within a collection C is:

$$Y_{ib} = \alpha_b \frac{1 - w^{n_i}}{n_i(1 - w)} q_{ib} + \sigma_b \varepsilon_{ib} \quad (2)$$

Giving total sales profit of:

$$Y_i = \sum_b Y_{ib}$$

We fit this model to the 2011 EziBuy catalogues and consistently find that the optimum w is very close to one. This is result was unexpected, since it implies that adding more pages to the catalogue does not result in diminishing marginal returns. We believe this is because the data has not been obtained from controlled experiments, so we can only establish correlation not causation. There are many possible explanations for why the large catalogues are performing better than our models would suggest. Firstly, EziBuy is more likely to send large catalogue separately, but small catalogues together. When a customer is sent multiple small catalogues together it will mimic a large catalogue. Consequently, the customer will spend less on the small catalogue than they would have if it had been sent by itself. The next possible explanation is that EziBuy are more likely to send larger catalogues to higher value customers. This will cause the larger catalogues to appear to perform better than smaller catalogues.

2.2 Historical models

In this model, we assume that if we send the same collection to the same segment we will get the same response. Firstly, this implies that customer expenditure on collections is independent of earlier events. Therefore customer behaviour is not influenced by recent collections sent to the customer. It also means that we must ignore fluctuations in demand from seasonal changes. It is easy to think of instances where this assumption is violated. For example, we would expect customers to buy cardigans more frequently in winter.

3 Optimisation model

In this section, we investigate an optimisation model to decide which collections to send to which customers in order to maximize profit. This formulations will be tested by using data from historical catalogues. This is based on the assumption that if we send a historical catalogue to the same segment we will get the same response as we did previously.

Parameters

C is the set of possible collections which can be used.

P is the set of parts that can be used.

B is the set of all available brands.

S is the set of segments.

$v_{sc} \in [0, \infty)$ is the expected profit from sending collection c to segment s .

$\eta_{cb} \in [0, \infty)$ is the number of pages of brand b required for collection c .

$\alpha_p \in [0, \infty)$ is the setup cost associated with part p .

$\beta_b \in [0, \infty)$ is the design cost per page of brand b .

$\Omega_{cp} \in \{0,1\}$ is whether part p is required for collection c to be built.

Variables

$x_{sc} \in \{0,1\}$ is whether segment s receives collection c .

$y_p \in \{0,1\}$ is whether part p is used.

$u_c \in \{0,1\}$ is whether collection c is used.

$q_b \in [0, \infty)$ is the number of pages of brand b that must be designed.

3.1 Core Assumption

The most fundamental assumption made by this optimisation models is that we only maximize profit over a short interval. To avoid this assumption the model would require quite a different formulation. This assumption implies that:

- The impact of collections that have been sent to the customer in the past is negligible. This assumption is likely to be worse the shorter the time interval, since if a customer bought a shirt two days ago it makes it unlikely they will buy shirts today.
- We ignore the future value from sending a customer a collection which may prompt them to become a more valuable customer. For example, sending new customers collections may make them become a regular customer and therefore more valuable.
- There is no future value from designing items. This is unlikely to be true since the cost of designing a page of catalogue depends on whether the items on the page have been used in previous catalogues. If the item has been used in previous catalogues the page design can be recycled so it costs less to produce each page.

3.2 Profit from Sending a Collection

In order to make the optimisation possible we must split the customers into groups, known as segments. The expected profit from sending a collection to a segment (v_{sc}) is the profit from sales to the segment minus the printing cost from sending the collection to the segment. v_{sc} excludes any fixed costs. We assume that there is a finite list of collections and each segment can receive at most one collection. These concepts are captured by the following formulas.

Total profit from allocating collections segments is:

$$\sum_{s \in S} \sum_{c \in C} v_{sc} x_{sc} \quad (3)$$

Since each segment s can be allocated at most one collection, we have:

$$\sum_{c \in C} x_{sc} \leq 1, \forall s \in S \quad (4)$$

The following table is an example of the values that v_{sc} could take. Highlighted in grey is the largest non-negative element in each row. If there were no other constraints then the optimal solution would be to set $x_{sc} = 1$ for these cells and $x_{sc} = 0$ otherwise.

		Collections			
		c_1	c_2	c_3	c_4
Segment	s_1	3	5	1	4
	s_2	5	6	7	3
	s_3	6	3	3	4
	s_4	6	2	1	2
	s_5	3	4	4	5

Table 1 – Expected profit from sending a collection to a segment, showing the optimal solution ignoring fixed costs, shaded in grey.

However, when there are setup and design costs this problem is not straightforward. In Table 1 the solution used four different collections. If there are setup costs and design costs this is unlikely to be optimal. One could avoid using c_3 with a loss of 2 units of profit by choosing c_1 for segment s_2 . This would be worthwhile if the setup and design costs saved were greater than 2.

3.3 Design costs

The design cost includes photographing the item on the model and laying out the items on the page. We assume that there is a constant cost β_b of designing a page of brand b . We also assume that we can recycle designs between collections. For example, if there are two collections c_1 and c_2 which use 5.5 pages and 6.3 pages of brand 1 respectively, then we will need to design a total of 6.3 pages of brand 1 for cost of $6.3 \times \beta_1$. In other words, we need to design the maximum number of pages of brand b over the collections used:

$$q_b = \max_{c \in C} \eta_{cb} u_c$$

If we formalise these ideas then we get the following expressions for the design costs.

The total design costs:

$$\sum_{b \in B} \beta_b q_b \quad (5)$$

For each collection c used we must have designed at least η_{cb} pages of brand b :

$$\eta_{cb} u_c \leq q_b, \forall b \in B, \forall c \in C \quad (6)$$

If an item has already been used in a previous catalogue then the design cost is significantly lower since we can re-use the layout and photographs. Therefore we will split each brand into two dummy brands ‘new’ items and ‘old’ items. ‘New’ items have a higher design cost per page than ‘old items’. For example, in the optimisation the brand Sara will be split into ‘new Sara’ and ‘old Sara’, which will be treated as two separate brands with different design costs per page.

3.4 Setup costs

In order for a collection to be sent first the required parts must be made. Each of these parts has a setup cost. Parts include costs like the setup cost of a 32 page section at the printer. Some of these parts may be shared amongst different collections. This setup cost is the same irrespective of how many copies of the collection are produced.

We will now express the setup costs by the following equations.

Setup costs are the sum of setup costs for each part p used:

$$\sum_{p \in P} \alpha_p y_p \quad (7)$$

Collection c can only be used if the required parts p have been used:

$$\Omega_{cp} u_c \leq y_p, \forall c \in C, \forall p \in P \quad (8)$$

3.5 Formulation

In this section we combine the prior equations to complete the formulation.

Profit = Profit from sending segments collections (3) – part setup costs (7) – design costs (5):

$$\max \sum_{s \in S} \sum_{c \in C} v_{sc} x_{sc} - \sum_{p \in P} \alpha_p y_p - \sum_{b \in B} \beta_b q_b \quad (9)$$

$$\sum_{c \in C} x_{sc} \leq 1, \forall s \in S \quad (4)$$

$$x_{sc} \leq u_c, \forall s \in S, \forall c \in C \quad (5)$$

$$\Omega_{cp} u_c \leq y_s, \forall c \in C, \forall p \in P \quad (8)$$

$$\eta_{cb} u_c \leq q_b, \forall b \in B, \forall c \in C \quad (6)$$

3.6 Parameter Input

The lists of collection used were each of the historical catalogues from 2011. In order to determine the sales profit of sending a collection to a segment we made the same assumptions outlined in section 2.2. The printing costs were estimated to be \$0.02 per page. Setup costs are \$4000 per catalogue and \$4000 for each 32 page section used in each catalogue. Brands were split into new and repeat items. New items require a photo-shoot with a model so they are more expensive to design per page. New items cost \$3500 per page and \$500 per page for old items.

4 Results

All the problem instances in could be solved quickly to optimality, using Gurobi, with up to thirty segments and five hundred collections in less than two minutes.

We use two different rules of thumb to find solutions. The personalised collection is selected by choosing the catalogue with the maximum v_{sc} for each segment. The one collection solution is chosen by selecting the best collection single collection to send to all segments.

	Personalised Collection	One Collection	Optimisation
Profit from Sending	\$1,545,672	\$1,014,556	\$1,312,654
Design costs	-\$942,533	-\$88,359	-\$166,157
Setup costs	-\$725,125	-\$22,250	-\$70,250
Total Profit	-\$121,986	\$903,948	\$1,076,247

Table 2 - Profit of solutions found by different methods.

Table 2 above shows profit of the solutions found by the different methods. The personalised catalogue for each customer produces the most profit from sending but has such high design and setup costs that it is not even profitable. On the other hand, only using one collection has the least profit from sending, but the lowest design and setup costs, consequently it produces net profit of \$900,000. The optimisation model provides a significant improvement on these two rules of thumb, producing over \$150,000 in additional profit.

5 Conclusions

In the section on sales models, we investigated models for predicting customer behaviour. The catalogue browsing model predicts that larger catalogues should perform worse than smaller catalogues. Unfortunately, when we attempt to fit parameters to this model these patterns do not appear. We believe this is because the data was not sourced from a controlled experiment. Therefore, we recommend that EziBuy should conduct a controlled experiment to gather the data to give accurate estimates of consumer response to different sizes and styles of catalogue.

In the section on optimisation, we create an optimisation model. The optimisation model is tested using historical catalogues. We find that the objective value from the optimisation is significantly better than the objective found when using a simple rule of thumb.

A natural extension of this work would be to test the optimisation models with collections that contain more than one catalogue. This would require a significant amount of data analysis to extract this information from the data. Another possible extension is to implement this optimisation model within EziBuy. This would require getting more accurate values of costs and product margins. Furthermore, an effective user interface for the optimisation model would need to be developed. Ideally, this would be integrated with the EziBuy database. Finally, the core assumption of this work was that we were maximizing short term profit. A new direction for this research could be extending the optimisation models so that they can take into account the future value of sending a collection to a customer.

Acknowledgments

I would like to thank my supervisor Andy Philpott and EziBuy decision support manager Catherine Hicks.

6 References

Steinbach, M., G. Karaypis, and V. Kumar. 2000. "Efficient Algorithms for Creating Product Catalogs." Technical Report, University of Minnesota.

HVDC Roles in the Economic Operation of the New Zealand Electricity Market

Vladimir Krichtal

vladimir.krichtal@transpower.co.nz

Conrad Edwards

conrad.edwards@transpower.co.nz

both from System Operations at Transpower New Zealand

Abstract

The paper presents a MILP formulation for the controllable unit commitment HVDC model. This formulation allows modelling reserve transfer between two AC systems via an HVDC link, creating a national reserves market. Round power refers to having the two poles transfer power in opposite directions to allow overall smooth control of HVDC power flow for lower rates of transfer. Testing of the model shows that the “round power” mode can be optimal for reserve sharing via the HVDC link for some market conditions.

Key words: MILP, HVDC, reserve transfer, national reserve market, round power.

1 Introduction

New Zealand’s electricity system is undergoing major structural changes. The high voltage direct current (HVDC) inter-island link, with the construction of a modern Pole 3, the decommissioning of the old Pole 1, and the installation of a new control system, can be operated in more sophisticated modes including “round power”.

The existing dispatch model assumes that both poles transfer power in the same direction. This operation minimises HVDC losses, but not necessarily total generation and reserve costs. This paper explores how overall cost savings could be achieved.

The HVDC link in New Zealand connects the two HVAC electrical systems on the North and South Island, each of which has separate reserves requirements. Islands are interconnected in terms of energy but not interconnected in terms of sharing reserves (both instantaneous reserves and frequency keeping). While the control system of the old HVDC configuration of Pole 1 and Pole 2 allowed sharing instantaneous and regulating reserve in physical operation, this was and is not reflected in the market rules or the scheduling, pricing and dispatch (SPD) model that produces least cost dispatch. Cost saving may be able to be achieved with the planned introduction of a national reserves market, which would allow the sharing of reserves between the two islands using the HVDC in a controllable configuration.

Some research of this area was reported in Krichtal 2006. In that paper the reserve transfer model was implemented into the SPD model and simulation results compared with benchmark results. The HVDC configuration and other system parameters were not changed during the simulation. These simulations captured the effect of reserve sharing with an already chosen and fixed HVDC configuration: the HVDC configuration was set outside

SPD and not optimised, in particular when total HVDC flow was close to the dead zone of 30 MW for each pole's operation.

Continuity of overall HVDC operation close to and through the dead zone can be achieved via "round power" mode with forced constraints on each pole's power flow, but this will increase loss costs. This paper explores and demonstrates how round power mode can be feasible and economic in operation in the presence of a reserve sharing/transfer model.

Further, existing schedules are calculated as static optimisation problems, independently run for each time interval. This paper explores also whether it is worth introducing a dynamic optimisation schedule with a start-up cost for each pole.

2 Market model with HVDC poles control.

For purposes of this research we do not need to model entire AC systems. So, North Island and South Island AC networks are aggregated into Haywards (HAY) and Benmore (BEN) nodes respectively. Only one reserve class is defined.

MODEL: DEMMDC

Variables, parameters and constraints are written for each study interval t unless otherwise specified.

Sets and indices

T	: study time interval $t \in T$,
OF	: set of energy and reserve offers $i \in OF$,
MN	: set of nodes $n \in MN$,
GU	: set of generation units $u \in GU$,
LN	: set of all power lines $l \in LN$,
POL	: set of DC poles $p \in POL$,
RZ	: set of reserve zones $rz \in RZ$.

Parameters

$OF_{i,n,t}^E$: mapping of energy offer i at node n ,
$OF_{i,u,t}^{GU}$: mapping of energy offer i at unit u ,
$OF_{i,u,t}^{RU}$: mapping of reserve ¹ offer i at unit u ,
$OF_{i,n,t}^R$: mapping of reserve offer i at node n ,
$PLSR_{i,rz}$: mapping of PLSR offer i in reserve zone rz ,
$ILR_{i,rz}$: mapping of ILR offer i in reserve zone rz .
$LN_{l,n}^{TO}$: conventional mapping of line l to node n ,
$LN_{l,n}^{FR}$: conventional mapping of line l from node n ,
$PL_{p,l}$: mapping of DC line l to the pole p ,
$URZ_{u,rz}$: mapping of generation unit u to reserve zone rz ,
$RZ_{rz,n}$: mapping of node n to reserve zone rz ,
$DCML_{rz,t}$: total HVDC modulation limit at reserve zone rz ,
$Of_{i,t}^{UP}$: upper limit for offer i ,
$Of_{i,t}^{PR}$: price (\$/MW) for offer i ,
$Fl_{l,t}^{UP}$: upper flow limit at line l ,

¹ Reserve refers to instantaneous reserves, of two types: PLSR = partially loaded spinning reserve linked to the unit generators, and ILR = interruptible load reserves linked to loads.

$Fl_{l,t}^{UPOV}$: contingency overload flow limit at line l ,
$Fl_{l,t}^{LO}$: lower flow limit at line l ,
$Ld_{n,t}^F$: fixed load at node n ,
$DCStUpPr_{p,t}$: HVDC pole's start-up price for pole p ,
$UCap_u$: capacity limit of generation unit u .

Variables

$Of_{i,t}$: value of cleared energy or reserve offer i .
$Fl_{l,t}$: power flow at line l ,
$LossL_{l,t}$: line's losses at line l ,
$LossN_{n,t}$: node's losses at node n ,
$FlSend_{l,t}$: flow at sending end of line l ,
$FlRec_{l,t}$: flow at receiving end of line l ,
$DCB_{l,t}$: binary variable showing operational status of HVDC line l ,
$DCPlIn_{p,t}$: binary variable showing operational status of HVDC pole p ,
$DCStUpCost_{p,t}$: HVDC pole's start-up cost for pole p .

All variables are nonnegative and limited until otherwise specified.

Constraints

Line losses models as quadratic function of flow:

$$(1) LossL_{l,t} = R_l \times (Fl_{l,t})^2, l \in LN$$

Line's flow at sending end:

$$(2) FlSend_{l,t} = Fl_{l,t} + 0.5 \times LossL_{l,t}, l \in LN$$

Line's flow at receiving end:

$$(3) FlRec_{l,t} = Fl_{l,t} - 0.5 \times LossL_{l,t}, l \in LN$$

Line's losses at node n , time t are set by halves to the connected nodes:

$$(4) LossN_{n,t} = 0.5 \times \left(\sum_{l \in LN_n^{TO}} LossL_{l,t} + \sum_{l \in LN_n^{FR}} LossL_{l,t} \right), n \in MN$$

Generation injection at node n :

$$(5) Gn_{n,t} = \sum_{i \in OF_{n,t}^E} Of_{i,t}, n \in MN$$

Generation of unit u :

$$(6) Gu_{u,t} = \sum_{i \in OF_{u,t}^{GU}} Of_{i,t}, u \in GU$$

Net node injection at node n :

$$(7) Nbi_{n,t} = \sum_{l \in LN_n^{TO}} Fl_{l,t} - \sum_{l \in LN_n^{FR}} Fl_{l,t}, n \in MN$$

Energy conservation equation at node n with node losses set as a load:

$$(8) Nbi_{n,t} + Gn_{n,t} - Ld_{n,t}^F - LossN_{n,t} = 0, n \in MN$$

HVDC configuration constraints

The HVDC configuration is presented in Figure 1. It consists of two physical poles (branches) set in parallel between two AC nodes. Each pole's power flow in every direction has positive lower and upper operation limit. There is a dead band $Fl_{l,t}^{LO}$ near the zero point. Operationally each HVDC pole is a directed line, so changing direction requires changing the HVDC pole's configuration. So, each pole's power flow is modelled by two lines in opposite directions. In the model if first line is operational, the second has zero power flow. Here the South Island node is Benmore (BEN or B) and the North Island node is Haywards (HAY or H):

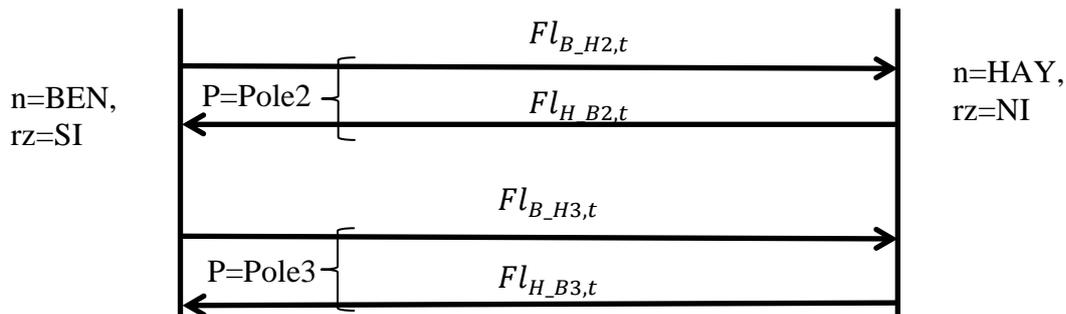


Figure 1 – HVDC model network

For each pole there are three operational possibilities:

1. Pole does not operate: $Fl_{FR,t} = Fl_{BW,t} = 0$.
2. Pole operates in forward conventional direction:
 $Fl_{BW,t} = 0, Fl_{FR,t}^{LO} \leq Fl_{FR,t} \leq Fl_{FR,t}^{UP}$
3. Pole operates in backward conventional direction:
 $Fl_{FR,t} = 0, Fl_{BW,t}^{LO} \leq Fl_{BW,t} \leq Fl_{BW,t}^{UP}$

These three possibilities can be modelled by MIP constraints

$$(9) Fl_{l,t} \geq Fl_{l,t}^{LO} \times DCB_{l,t}, l \in LN$$

$$(10) Fl_{l,t} \leq Fl_{l,t}^{UP} \times DCB_{l,t}, l \in LN$$

$$(11) DCPoleIn_{p,t} = \sum_{l \in PL_{p,l}} DCB_{l,t}$$

$$(12) DCPoleIn_{p,t} \leq 1.$$

where (9) is lower limit, (10) is upper limit, and (11,12) show that no more than one direction of each pole can be operational.

Risk-reserves constraints

Loss of a generation unit or total HVDC poles are considered in the model as a risk. Risks are rare enough to consider the events as independent, so instantaneous reserves from the entire system can be used to cover any risk, given the HVDC's capacity limits.

Partially loaded spinning reserves:

$$(13) \text{PLSRPool}_{rz,t} = \sum_{i \in \text{PLSR}_{i,rz}} Of_{i,t}, rz \in RZ$$

Energy plus PLSR reserves from the same unit should be less than the unit's capacity:

$$(14) \sum_{i \in OF_{u,t}^{GU}} Of_{i,t} + \sum_{i \in OF_{i,u,t}^{RU}} Of_{i,t} \leq UCap_{u,t}, u \in GU$$

Interruptible reserves:

$$(15) \text{ILRPool}_{rz,t} = \sum_{i \in \text{ILR}_{i,rz}} Of_{i,t}, rz \in RZ,$$

Total instantaneous reserves consist of PLSR and ILR:

$$(16) \text{RPool}_{rz,t} = \text{ILRPool}_{rz,t} + \text{PLSRPool}_{rz,t}, rz \in RZ$$

Generator risk is unit generation plus reserves from the same unit:

$$(17) GR_{u,t} = Gu_{u,t} + \sum_{i \in OF_{u,t}^{RU}} Of_{i,t}, u \in GU$$

HVDC risk is calculated as a sum of receiving minus sum of sending flows:

$$(18) \text{DCR}_{rz,t} = \sum_{n \in RZ_{rz,n}} \left(\sum_{l \in LN_{l,n}^{TO}} Fl_{l,t}^{Rec} - \sum_{l \in LN_{l,n}^{FR}} Fl_{l,t}^{Sent} \right), rz \in RZ,$$

When HVDC in operation there is one or two poles sending power in the same or opposite directions. Reserve transfer capacity for each HVDC pole is illustrated in Figure 2.

Reserve transfer in forward direction:

$$(19) \text{ResTr}_{l,t}^{FW} \leq Fl_{l,t}^{UPOV} \times DCB_{l,t} - Fl_{l,t}, l \in LN$$

Reserve transfer in backward direction:

$$(20) \text{ResTr}_{l,t}^{BW} \leq Fl_{l,t}, l \in LN$$

Reserve import into zone:

$$(21) \text{RImZn}_{rz,t} = \sum_{l \in LN_{l,n}^{TO} \& n \in RZ_{rz,n}} \text{ResTr}_{l,t}^{FW} + \sum_{l \in LN_{l,n}^{FR} \& n \in RZ_{rz,n}} \text{ResTr}_{l,t}^{BW},$$

$$rz \in RZ$$

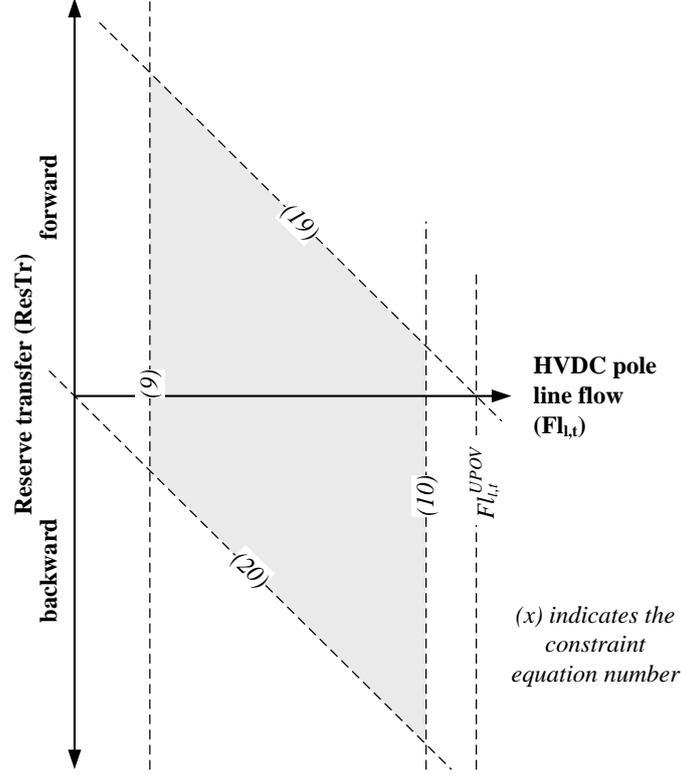


Figure 2 – HVDC line reserve transfer feasible area

Reserve import into zone rz should be less than the total HVDC modulation level:

$$(22) \quad RImZn_{rz,t} \leq DCML_{rz,t}$$

Reserve collected inside zone rz plus reserve imported should be bigger than the generation risk:

$$(23) \quad RPool_{rz,t} + RImZn_{rz,t} \geq \sum_{u \in URZ_{u,rz}} GR_{u,t}, rz \in RZ.$$

Reserve collected inside zone rz should be bigger than HVDC risk (reserves transfer is not counted):

$$(24) \quad RPool_{rz,t} \geq DCR_{rz,t}, rz \in RZ$$

Generation unit ramp-up:

$$(25) \quad (Gu_{u,t} - Gu_{u,t-1}) \leq URU_u$$

Generation unit ramp-down:

$$(26) \quad (Gu_{u,t-1} - Gu_{u,t}) \leq URD_u$$

There may be cost associated with switching on an HVDC pole from non-operational (cold) to operational (hot) stages. Changing pole's direction when pole is in operation and switching pole off has no cost.

$$(27) \quad DCStUpCost_{p,t} \geq DCStUpPr_{p,t} \times (DCPlIn_{p,t} - DCPlIn_{p,t-1}),$$

Objective function

The objective is to minimise reserve, energy and HVDC start-up costs, summed over study time interval T :

$$(28) \text{ Minimise } \sum_{t \in T} (\sum_{i \in OF} Of_{i,t} \times Of_{i,t}^{PR} + \sum_{p \in POL} DCStUpCost_{p,t})$$

The model defined in equations (1) to (28) is a mixed integer non-linear program (MINLP) and has a non-convex feasible decision space. After approximating quadratic losses by piecewise linear functions with mixed integer linear program (MILP) constraints (as is done in New Zealand’s dispatch and pricing model), the model is solved for a global optimum using a MILP solver.

3 Numerical results

We study the economic effect of HVDC configuration switching using the DEMMDC model described above, with only two generators at each island, one reserve type, each generator is a risk, net HVDC flow is a risk, and for three periods.

Test 1

The goal is to show that the reserve transfer model can be more economical in “round power” mode. This test includes all model features except generation ramping constraints and start-up costs, and is run for one period. Results of basic case without reserve transfer (No RTR) are compared to the case with reserve transfer (RTR). The results are shown in Tables 1, 2 and 3 below:

Table 1 – Costs

	SysCost	EnCost	ResCost
No RTR	28369.21	20108.59	8260.62
RTR	20170.93	19381.40	789.53

Table 2 – Energy and reserves at zones (North Island and South Island)

		Gn	RPool	RImZn	Ln	Nb	EnPrice
No RTR	NI	450.00	300.00	0	300.00	147.28	63.23
	SI	355.43	300.00	0	500.00	147.28	68.00
RTR	NI	409.54	110.46	189.53	300.00	104.76	40.00
	SI	400.00	189.53	105.23	500.00	104.76	48.04

Table 3 – DC Flow, losses

		DC Line	DCB	Fl	LossL	FIlo	FIUp
No RTR		B_H2	0	0	0	30	210
		B_H1	0	0	0	30	210
		H_B1	1	75.38	2.84	30	210
		H_B2	1	71.90	2.58	30	210
RTR		B_H2	0	0	0	30	210
		B_H1	1	30	0.45	30	210
		H_B1	0	0	0	30	210
		H_B2	1	134.76	9.08	30	210

The case with reserve transfer (RTR) is comparing with the case without (No RTR). Energy and reserve costs are decreased with RTR (see Table 1), and energy prices are decreased at both nodes with RTR (see Table 2).

HVDC pole's flow at one direction changed to pole's flow in different direction (see Table 3) forcing HVDC round power mode. Result of this simulation shows 'Round power' mode of HVDC configuration can be more economical in RTR than the mode with both pole sending power in the same direction in No RTR. This is achieved despite the new HVDC configuration in RTR producing 4 MW more losses. The cost of these losses is less than the cost of reserves saved.

Test 2

We run the model for three periods. Again, the first run is without reserves transfer (No RTR) and the second with reserve transfer (RTR). All the model features are used including dynamic HVDC start-up costs and generation ramping constraints.

Table 4 – Costs in three periods

		SysCost	EnCost	SUpCost	ResCost
No RTR	t1	34800.00	23000.00	0	11800.00
	t2	28843.76	19653.74	200	8990.02
	t3	16593.34	14785.97	0	1807.37
RTR	t1	24062.61	23036.83	200	825.78
	t2	20170.94	19381.40	0	789.53
	t3	14033.71	13333.71	0	700.00

Table 5 – Energy and reserves in three periods

		EnZn	RezZn	ResImZone	DemZn	EnZoneIm	
SI	No RTR	t1	400.00	300.00	0	400.00	0
		t2	393.38	300.00	0	500.00	108.07
		t3	300.00	300.00	0	100.00	-195.20
	RTR	t1	400.00	200.00	100.00	400.00	0.46
		t2	400.00	189.53	105.23	500.00	104.76
		t3	366.68	50.00	180.00	100.00	-258.34
NI	No RTR	t1	500.00	300.00	0	500.00	0
		t2	409.53	300.00	0	300.00	-108.07
		t3	309.53	300.00	0	500.00	195.23
	RTR	t1	500.92	100.00	200.00	500	-0.46
		t2	409.53	110.46	189.53	300	-104.76
		t3	250.00	250.00	50.00	500	258.34

Table 6 – DC flow, losses in three periods

		t=1		t=2		t=3	
		BrFlow	BrLosses	BrFlow	BrLosses	BrFlow	BrLosses
No RTR	B_H2			0	0	96.92	4.69
	B_H1			0	0	98.31	4.83
	H_B1			53.84	1.44	0	0
	H_B2			54.22	1.47	0	0
RTR	B_H2	0	0	0	0	129.23	8.35
	B_H1	39.38	0.77	30.00	0.45	129.11	8.33
	H_B1	0	0	0	0	0	0
	H_B2	30.00	0.45	134.76	9.08	0	0

HVDC poles were switched off before starting the simulation. Total cost and reserve costs are decreased in every time period. Energy cost increased slightly in t=1 and decreased significantly in other times (see Table 4). In the first run (No RTR) HVDC stopped at t=1 and had the same operational directions at t=2, t=3. In the second run with reserve transfer (RTR) the HVDC was operational at periods t=1,2,3. So, without reserve transfer it was not economical to run the HVDC link at t=1: it is switched off. With RTR, in periods t=1,2 HVDC configuration switched from running poles in one direction to round power mode, substantially reducing reserves and overall cost (see Table 6).

4 Conclusions

This study has demonstrated that, using a relatively simple model, the “round power” mode may under certain conditions be optimal for reserve sharing via the HVDC link. This result may assist in the consideration of any market implementation.

Acknowledgements

The authors acknowledge Kieran Devine and Doug Goodwin at Transpower for supporting this research.

5 References

Krichtal V, 2006, “National Instantaneous Reserve Market in the New Zealand Wholesale Electricity”, proceedings of the 41st Annual Conference, ORSNZ’06.

Elective Course Student Sectioning at Danish High Schools

Simon Kristiansen* and Thomas R. Stidsen

Operation Research, DTU Management, Technical University of Denmark
Produktionstorvet, 426 B, DK-2800, Kgs Lyngby, Denmark

*sikr@dtu.dk

Abstract

The *Elective Course Student Sectioning* (ECSS) serves as a preprocessing planning problem for the actual High School Timetabling Problem at Danish high schools.

Each year the students in the Danish high schools request for some elective courses which they would like to participate in next to their mandatory courses. The main problem of the ECSS is of fulfilling as many student requests as possible while minimizing the number of classes created while respecting certain resource limitations. While assigning the students to classes it is attempted to incorporate some fairness to the distribution between classes.

The ECSS has been formulated as an Integer Programming model and is solved using Adaptive Large Neighborhood Search. Computational results are established using 50 real life instances from Denmark and the each solution is compared to an upper bound found using Gurobi. It is shown that the ALNS algorithm in average provides results within 1% of optimum.

The algorithm has been implemented in the cloud-based software high school administration system Lectio, and is hence available for more than 200 Danish high schools.

Key words: Student Sectioning, Elective Course Planning, Adaptive Large Neighborhood Search, Educational Timetabling, High School Planning, Metaheuristics

1 Introduction

Elective Course Student Sectioning (ECSS) is a recurrent planning problem at the Danish high schools and functions as a preprocessing planning problem for the High School Timetabling. Each year the students requests some elective courses, and the problem is then to assign these requests to some course classes and assign these to some time slots. The problem is very important for the students as they often have selected the courses upon the requirements for their future university education. However for the high schools there exists a cost of approximately NZ\$40.000 p.a. for each created class. As the high schools are paid upon graduated students it is

necessary to fulfill as many student requests while minimizing the number of assigned course classes.

Figure 1 illustrates a typically schedule for a high school with four time slots each day. The gray time slots are reserved for the elective courses whereas the white colored time slots are used for the mandatory courses. As not all students have requests the same amount of elective course, the time slots used for the elective courses are usually placed at the beginning or at the end of a day to minimize the possibility of creating idle time slots for a student.

	Monday	Tuesday	Wednesday	Thursday	Friday
8:15 9:45	Time1			Time3	
10:00 11:30					
Lunch break					
12:00 13:30					
13:45 15:15		Time2		Time4	Time5

Figure 1: An example of a weekly schedule with four time slots each day. Five time slots (gray colored) is reserved for the elective courses whereas the mandatory courses is placed in remaining time slots (white colored)

As the schedule can be divided into a mandatory and an elective part we will not consider the mandatory courses in the rest of this paper. Assigning mandatory course classes to time slots are known as the regular High School Timetabling Problem (see e.g. (Sørensen, Kristiansen, and Stidsen 2012))

When assigning students to elective course classes it is desirable to have some kind of fairness incorporated in the distribution. It is quite often that more students have requested a given course than it is allowed to fill in one course class due to limitation on the class size. I.e. more than one class are needed to fulfill all the student requests for the given course. When having two classes of same course, two subjects are looked upon for determining the fairness of the distribution of students. Firstly, the representation of common classes in each course class. Each student is assigned a common class when they start at a high school. Within these common classes most of the mandatory courses are taught. This gives some advances in collaboration between the students as they are quite familiar and it makes it easier for collaborations between mandatory course classes if all the students attending the two classes are exactly the same. Thus, it is also an advantage to have as few common classes represented in each elective course class. Secondly, it is preferred to have an even distribution of the students in the classes of same course. This is to ease the workload of the teachers and to make sure that all students has the possibility to get help during lectures. E.g. if 40 students have requested a given course and having an upper class size of 28, it is more fair to have a distribution of 20-20 instead of 28-12.

Student Sectioning is one of the less studied planning problems within educational timetabling (Pillay 2010) and those which exists are usually based on University Student Sectioning (Erben and Keppler 1996; Müller and Murray 2010). University Student Sectioning is often concentrated on a single or a few universities and is hence very adapted to these. The ECSS is focused on all the Danish high schools, i.e. more

than 200 users, hence it has to be very generalized and suitable for many different kind of setups. The last couple of years more researched has been focused on High School Timetabling (HSTT) and the *third International Timetabling Competition* (ITC2011) was treating the HSTT (see e.g. (Post et al. 2012)). However these do not consider the sectioning of students into classes. (Kristiansen, Sørensen, and Stidsen 2011) is concerned the ECSS, this was however only a test of a previous heuristic which was proven not to be sufficient and it lacked the fairness distribution of students.

This paper is written in collaboration with the small Danish software company MaCom A/S which main product is the cloud-based high school administration software Lectio. Lectio is used by the vast majority of the high schools in Denmark.

2 Integer programming model

ECSS is now formulated as an IP model. A high school have a set of students \mathcal{S} , a set of offered courses \mathcal{E} , a set of classes \mathcal{C} and a set of time slots \mathcal{T} . The parameter $D_{c,e} \in \{0, 1\}$ indicates if class c is teaching course e and $R_{e,s} \in \{0, 1\}$ indicates whether student s has requested course e , or not. For each class there exists an upper bound on the class size given by $U_c \in \mathbb{R}^+$, and the maximum number of classes which can be created is given by $P \in \mathbb{R}^+$. Each course belongs to one of the course subjects given by $f \in \mathcal{F}$. $K_{c,f} \in \{0, 1\}$ denotes whether course e is teaching subject f and $B_{f,t} \in \mathbb{R}^+$ denotes the maximum number of classes of subject f there can be assigned to time slot t . Each student is assigned a common class, $q \in \mathcal{Q}$, denoted by $I_{s,q} \in \{0, 1\}$. The parameters $A_{c,s}$ and A_c indicate whether student s is locked to course class c , and if course class c is a locked class, respectively.

The decision whether student s is assigned course class c in time slot t is given by the binary variable $x_{s,c,t} \in \{0, 1\}$, while the binary decision variable $y_{c,t} \in \{0, 1\}$ determine whether course class c is assigned to time slot t . The binary variable $z_{c,q} \in \{0, 1\}$ takes value 1 if common class q is represented in course class c , zero otherwise. Variable $w_{c,c'} \in [0, 1]$ is used for determine the difference in students between the course classes c and c' . The objectives are weighted in respect to each other using the weights $\alpha_{c,s}$, β_c , γ and δ .

$$\max \sum_{c,t,s} \alpha_{c,s} \cdot x_{s,c,t} - \sum_{c,t} \beta_c \cdot y_{c,t} - \gamma \cdot \sum_{c,q} z_{c,q} - \delta \cdot \sum_{c,c'} U_c \cdot w_{c,c'} \quad (1)$$

$$\text{s.t.} \quad \sum_c x_{s,c,t} \leq 1 \quad \forall t, s \quad (2)$$

$$\sum_{c,t} D_{c,e} \cdot x_{s,c,t} \leq R_{e,s} \quad \forall e, s \quad (3)$$

$$\sum_t y_{c,t} \leq 1 \quad \forall c \quad (4)$$

$$\sum_t x_{s,c,t} \leq U_c \quad \forall c, t \quad (5)$$

$$\sum_{c,t} y_{c,t} \leq P \quad (6)$$

$$x_{s,c,t} \leq y_{c,t} \quad \forall c, t, s, A_{c,s} = 0 \quad (7)$$

$$x_{s,c,t} = y_{c,t} \quad \forall c, t, s, A_{c,s} = 1 \quad (8)$$

$$\sum_c K_{c,f} \cdot y_{c,t} \leq B_{f,t} \quad \forall f, t, B_{f,t} > 0 \quad (9)$$

$$y_{c,t} + y_{c',t} \leq 1 \quad \forall c, c', t, J_{c,c'} = 1 \quad (10)$$

$$y_{c,t} \leq D_{c,e} \cdot h_{e,t} \quad \forall c, e, t \quad (11)$$

$$h_{e,t} = h_{e,t'} \quad \forall e, e', t, L_{e,e'} = 1 \quad (12)$$

$$\sum h_{e,t} \leq 1 \quad \forall e, e', t, L_{e,e'} = 1 \quad (13)$$

$$\sum_{s,q}^e I_{s,q} \cdot x_{s,c,t} \leq z_{c,q} \quad \forall c, s, q, A_c = 0 \quad (14)$$

$$\sum_{t,s} x_{s,c,t} - \sum_{t,s} x_{s,c',t} \leq g_{c,c'} \cdot U_c \quad \forall c, c', D_{c,e} = D_{c',e} = 1, \\ A_c = A_{c'} = 0, \rho(c) < \rho(c') \quad (15)$$

$$\sum_{t,s} x_{s,c',t} - \sum_{t,s} x_{s,c,t} \leq g_{c,c'} \cdot U_c \quad \forall c, c', D_{c,e} = D_{c',e} = 1, \\ A_c = A_{c'} = 0, \rho(c) < \rho(c') \quad (16)$$

$$\sum_t (y_{c,t} + y_{c',t}) - 2 + g_{c,c'} \leq w_{c,c'} \quad \forall c, c', D_{c,e} = D_{c',e} = 1, \\ A_c = A_{c'} = 0, \rho(c) < \rho(c') \quad (17)$$

$$x_{s,c,t} \in \{0, 1\} \quad (18)$$

$$y_{c,t} \in \{0, 1\} \quad (19)$$

$$z_{c,q} \in \{0, 1\} \quad (20)$$

$$w_{c,c'} \in [0, 1] \quad (21)$$

$$g_{c,c'} \in [0, 1] \quad (22)$$

$$h_{e,t} \in \{0, 1\} \quad (23)$$

As mentioned the objective of ECSS (1) is of fulfilling as many granted requests while minimizing the number of assigned course class, the number of common classes represented in each class and the difference in students between assigned course classes of same course. The constraints of (2) ensure that no students are assigned more than one course class in each time slot. Constraints (3) ensure that students only can be assigned to courses which they have requested. The constraints of (4) ensure that no course classes are assigned more than one time slot. Constraints (5) set the upper bound on the number of students in a given course class, whereas constraints (6) ensures that the total number of classes assigned to a time slots does not exceed maximum. Constraints (7) and (8) are the connection between the variables $x_{s,c,t}$ and $y_{c,t}$, and make sure that students cannot be assigned classes which is not assigned to a time slot. Constraints (8) ensure that if a course class has locked students, only these students can be assigned to the class, and that all the locked students are assigned the class if the class is assigned a time slot.

The resource restrictions on the number of subjects in each time slot are constrained by (9). Constraints (10) ensure that classes which cannot be placed in the same time slot are not done so, whereas constraints (11),(12) and (13) are used to ensure that course classes of courses which should be held in same time slot are satisfied. The slack variable $h_{e,t} \in \{0, 1\}$ takes the value 1 if course e is represented in time slot t .

Constraints (14) are counting the number of common classes used in each course class. Finally, constraints (15), (16) and (17) are used to determine the value of $w_{c,c'}$ used for the equal distribution of the students in assigned classes of same course. The slack variable $g_{c,c'} \in [0, 1]$ is created to assist the determination of $w_{c,c'}$. Let $\rho(j)$ denote the ordinal number of j .

3 Solution methods

It has been chosen to attempt *Adaptive Large Neighborhood Search* (ALNS) to establish solutions to the ECSS. The concept of ALNS (Pisinger and Ropke 2005) is to extend the large neighborhood heuristic by Shaw (Shaw 1997) by allowing the use of multiple removal and insertion heuristics. For each iteration a removal and an insertion heuristic are chosen upon some performance indicators which are updated after each iteration. ALNS has been successfully applied for various subjects (Kristiansen et al. 2012; Laporte, Musmanno, and Vucelja 2010; Steg and Schröder 2008) A pseudo code for the ALNS is presented in Algorithm 1.

Algorithm 1: Adaptive Large Neighborhood Search

Input: a feasible solution $x_{s,c,t}$

- 1 solution $x^{best} = x$; $\pi = (1, \dots, 1)$
- 2 **repeat**
- 3 select a removal $d \in \Omega^-$ and an insertion heuristic $r \in \Omega^+$ using π
- 4 $x' = r(d(x))$
- 5 **if** $c(x') > c(x^{best})$ **then**
- 6 $x^{best} = x'$
- 7 **if** $accept(x', x)$ **then**
- 8 $x = x'$
- 9 update π
- 10 **until** *stop-criterion met*
- 11 **return** x^{best}

Some main elements for the ALNS are now described.

- *Adaptive search strategy:* ALNS is using a scoring scheme which is governing the choice of the removal and insertion heuristics. The scoring used in this article is based on (Muller, Spoorendonk, and Pisinger 2011) where the performance of the heuristics is tracked by the percentage-wise gap between the current solution and the new solution. The search is divided into segments of N_{it} consecutive iterations. Let π_h^i be the measure of performance of heuristic h in segment i . In the first segment all the heuristics are given the same weight. The probability of choosing heuristic h in segment i is given by $\frac{\pi_h^i}{\sum_{\bar{h}} \pi_{\bar{h}}^i}$. After each segment the weights of the heuristics are updated according to the scoring scheme.

$$\pi_h^{i+1} = \eta \cdot \frac{\bar{\pi}_h^i}{a_h^i} + (1 - \eta)\pi_h^i \quad (24)$$

where $\eta \in [0, 1]$ is the reaction factor and $\bar{\pi}_h^i$ is the observed weight of heuristic h in segment i . $\bar{\pi}_h^i$ is updated after each iteration using the following scaling parameter

$$gap = \frac{c(x') - c(x)}{c(x)} \quad (25)$$

$$\bar{\pi}_h^i = \bar{\pi}_h^i + 5^{\min(\sigma \cdot gap, 1)} \quad (26)$$

where σ is a tuning parameter.

- *Acceptance criteria:* The acceptance criteria are based on the acceptance criteria in *Simulated Annealing*. I.e. given a current solution x , a new solution x' is accepted if $c(x') > c(x)$ and it is accepted with the probability $\exp(-\frac{c(x')-c(x)}{T})$ otherwise. $T > 0$ is the current *temperature*. The temperature starts at T_{start} and is multiplied by d_{SA} , where $0 < d_{\text{SA}} < 1$ is the *cooling rate*. T_{start} is selected using a *temperature control parameter* w_{SA} , where $0 < w_{\text{SA}} < 1$, such that a solution is accepted with the probability of 0.5 if the solution is w_{SA} percent worse than the initial solution.
- *Stopping criteria:* The selection of removal and insertion heuristics is continued until one of the following two stopping criteria is met; (1) the running time exceed the maximum running time of 60 seconds; or (2) the number of iterations without any improvements in the objective reaches 1,000.

3.1 Removal and insertion heuristics

This section describe all the removal and insertion heuristic used for the ECSS. Furthermore some coupled methods are introduced. Let $m \in \mathbb{N}$ be the number of classes which should be removed from a solution x in a removal heuristic, and let $\bar{\mathcal{C}} \subseteq \mathcal{C}$ be the set of unassigned classes which could be assigned in an insertion heuristic.

3.1.1 Random removal

This simple removal heuristic removes m course classes with students from the current solution. It is obvious that this heuristic tends to generate poor solutions, but it helps diversifying the search.

3.1.2 Shaw removal

The idea behind the Shaw removal heuristic is to remove part of the current solution which is somehow related, as it is expected that it should be reasonably easy to shuffle between similar items ((Shaw 1997) and (Ropke and Pisinger 2006)). The relatedness between two classes of the current solution of ECSS is calculated using a relatedness measure based on the number of students the two classes could share in reality, i.e.

$$M(c, c') = \frac{|\mathcal{S}_c \cap \mathcal{S}_{c'}|}{\min(|\mathcal{S}_c|, |\mathcal{S}_{c'}|)} \quad \text{where } D_{c,e} = D_{c',e'} = 1 \quad (27)$$

where \mathcal{S}_e is the set of students which has requested course e . The more students they share the more related the two classes are.

Algorithm 2: Shaw removal heuristic

Input: A feasible solution $x_{s,c,t}$, $m \in \mathbb{N}$, $p_{\text{shaw}} \in \mathbb{R}^+$

- 1 class: $c =$ a randomly selected class with students from $x_{s,c,t}$
- 2 set of classes : $D = \{c\}$
- 3 **while** $|D| < m$ **do**
- 4 $\hat{c} =$ randomly selected class from D
- 5 $L =$ all classes from $x_{s,c,t}$ not in D , sorted by similarity to \hat{c}
- 6 choose a random number $b^{p_{\text{shaw}}} \in [0, 1[$
- 7 $l =$ element number $b^{p_{\text{shaw}}} \cdot |L|$
- 8 $D = D \cup L[l]$
- 9 remove the classes with students in D from $x_{s,c,t}$

Two Shaw heuristics has been implemented for the ECSS. One ordered by decreasing similarity (removing the ones which are most related) and one sorted by increasing similarity (of those related, remove those which are less related).

3.1.3 Basic greedy

One of the insertion heuristics for the ALNS is a basic greedy algorithm. A move of classes is of assigning a class with as many students as possible to a time slot. For the basic greedy, it assigns the class with the highest contribution to the objective. The process is then repeated until no classes with a positive contribution can be assigned. The initial solution for the ECSS is created using a basic greedy algorithm.

3.1.4 Regret- k heuristic

The *Regret* heuristic tries to improve the myopic behavior of greedy heuristics. The heuristic calculates a regret value equal to the cost difference between solutions in which a class is assigned to its best or at its second best position. The class with highest regret value is the class which will be regretted most if it is not assigned to its best position. The concept can be extended such that it considers the k -best positions. Formally, let $o_{\bar{c}}^k$ denote the regret value by inserting class c into the k^{th} best position. I.e. the regret value of \bar{c} , $r_{\bar{c}}$, is given by

$$r_{\bar{c}} = \sum_{h_2}^k (o_{\bar{c}}^1 - o_{\bar{c}}^k) \quad (28)$$

In each iteration the heuristic chooses to insert class \bar{c} according to $\max_{\bar{c} \in \bar{\mathcal{C}}} \{r_{\bar{c}}\}$ It has been chosen to use Regret-2, -3 and -4 as repair methods for the ECSS.

3.2 Coupled heuristics

The previous mentioned removal and insertion heuristics can all be connected in an iteration of the ALNS. However it can be an advantage to create some coupled heuristics, i.e. one removal and one insertion heuristic which can only be used together. For the ECSS three coupled heuristics have been created, all concerning only the students and the fair distribution between classes of the same course. Let $\hat{\mathcal{C}} \subseteq \mathcal{C}$ be the set of course classes for which more than one class is created for the given course. The following coupled constraints have been created.

- Remove all students from classes in $\hat{\mathcal{C}}$ and insert them greedily.
- Remove all students from classes in $\hat{\mathcal{C}}$ and insert them greedily based on their common classes
- Remove all students from classes in $\hat{\mathcal{C}}$ which are located in the same time slot and insert them greedily.

4 Results

The ALNS algorithm for the ECSS has been implemented in the cloud-based high school administration system Lectio and is hence available for use for approximately 200 different high schools in Denmark. This gives the possibilities for a huge amount of data for testing and future research.

The algorithm has been tested on instances from 50 different high schools. The datasets are selected randomly and is believed to cover all possible setups for the ECSS. The runtime for the ALNS is 60 seconds, selected upon conversation with MaCom A/S and the users of Lectio. Each data set is run 10 times in order to reduce eventual influence of stochastic behavior. The running time for Gurobi is 1 hour as we want to have good upper bound for the instances. The percentage difference between the average solution found using ALNS and the upper bound from Gurobi is calculated by $\frac{UB-\bar{x}}{UB}$. Both the ALNS and the Gurobi implementation of ECSS was coded in C# 4.0 and all tests are performed using NUnit 2.6 on a machine with an Intel i7-930@2.8GHz CPU and 12GB of RAM under a Windows operating system. No parallelization has been implemented for improvements.

The ALNS contained 9 free parameters which were tuned using *iterative F-race*. (Balaprakash, Birattari, and Stützle 2007).

Table 1 shows that the ALNS in average finds solutions within 1% from optimum, which is quite satisfying results. Although ALNS has a far lower running time than Gurobi, it finds solutions very close to the solutions from Gurobi in many cases. For the largest problem *Slagelse* it finds a solution much better.

Table 1: *ALNS for the ECSS on 80 datasets compared with an upper bound using Gurobi 5.0.1. For each dataset is listed the number of students “ $|\mathcal{S}|$ ”, number of requests “ $|\mathcal{R}|$ ” and number of courses “ $|\mathcal{E}|$ ”, which indicates the size of the given instance. For Gurobi is listed the final objective value, “ x ”, the best bound “ UB ” and the reported gap between these. For the ALNS, the mean performance of the algorithm over 10 runs, “ \bar{x} ” and column “ σ ” is the standard deviation. Finally column “ $Gap(\%)$ ” is the percentage difference between ALNS and Gurobi.*

	S	R	E	Gurobi 5.01			ALNS		
				x	UB	Gap[%]	\bar{x}	σ	Gap[%]
Aabenraa	20	20	3	1630.0	1630.0	0.0	1630.0	0.0	0.0
Aalborg	212	539	16	79570.0	79570.0	0.0	79010.0	0.0	0.7
Aarhus	332	463	27	54794.0	55088.0	0.5	54765.6	3.6	0.6
Alssund	183	338	17	32829.0	32929.0	0.3	32824.4	2.8	0.3
Bagsvaerd	56	75	10	10450.0	10450.0	0.0	10450.0	0.0	0.0
Bornholm	525	965	41	114877.0	115368.0	0.4	114828.8	20.2	0.5
CPHWEST	249	426	32	48405.0	48499.0	0.2	48405.0	0.0	0.2
Detfrie	112	112	3	10839.0	10840.0	0.0	10825.0	0.0	0.1
Erhvervsskolerne	219	365	23	39315.0	39315.0	0.0	39315.0	0.0	0.0
Esbjerg	595	789	34	89972.0	90670.0	0.8	90014.6	21.9	0.7
EUCNORD	335	735	45	95080.0	95088.0	0.0	94911.0	97.5	0.2

Continued on next page

Table 1 – continued from previous page
Gurobi 5.01

	\mathcal{S}	\mathcal{R}	\mathcal{E}	Gurobi 5.01			ALNS		
				x	UB	Gap[%]	\bar{x}	σ	Gap[%]
Falkoner	297	456	34	41395.0	41746.0	0.9	41374.4	4.0	0.9
Fjerritslev	456	822	71	113706.0	113717.0	0.0	113700.6	1.9	0.0
Frederikssund	294	473	24	68950.0	68950.0	0.0	68950.0	0.0	0.0
Gladsaxe	1038	1510	89	220100.0	220100.0	0.0	220100.0	0.0	0.0
Greve	306	892	31	99588.0	100594.0	1.0	98869.0	144.4	1.7
Gribskov	394	648	32	72092.0	72098.0	0.0	71485.1	215.4	0.9
GU-Aasiaat	71	82	12	11820.0	11820.0	0.0	11820.0	0.0	0.0
Haderslev	447	1034	37	112058.0	113385.0	1.2	110688.2	1537.6	2.4
Herlev	502	1336	56	157826.0	159125.0	0.8	156145.8	1120.7	1.9
Hoeje-Taastrup	233	380	17	35825.0	35828.0	0.0	35789.8	27.2	0.1
HTXSukkertoppen	332	339	12	35439.0	35581.0	0.4	35409.0	0.0	0.5
Koebenhavns aabne	289	816	31	101240.0	101249.0	0.0	100845.0	38.3	0.4
Koege Gymnasium	369	546	31	79360.0	79360.0	0.0	79360.0	0.0	0.0
Koege Handelsgym	76	76	12	10440.0	10440.0	0.0	10440.0	0.0	0.0
Mariagerfjord	365	521	24	76070.0	76070.0	0.0	75670.0	0.0	0.5
Marselisborg	760	1248	42	140605.0	141664.0	0.8	140518.8	29.5	0.8
Middelfart	390	1332	61	170243.0	170515.0	0.2	169764.0	324.7	0.4
Munkensdam	482	6456	231	349400.0	349400.0	0.0	349400.0	0.0	0.0
NZahles	189	271	20	31955.0	31958.0	0.0	31955.0	0.0	0.0
Noerresundby	582	1027	45	118115.0	118915.0	0.7	118028.4	12.5	0.8
Nordfyns	381	600	36	68884.0	68906.0	0.0	68868.8	27.1	0.1
Oeregaard	547	826	27	96468.0	97167.0	0.7	96419.0	17.0	0.8
Risskov	539	784	38	91968.0	92610.0	0.7	91943.2	5.0	0.7
Roedovre	350	868	34	100899.0	101331.0	0.4	100874.6	10.3	0.5
Roskilde Katedralskole	383	1145	40	126611.0	128351.0	1.4	126346.0	238.5	1.6
Roskilde Tekniske	199	381	16	52950.0	52950.0	0.0	52950.0	0.0	0.0
Rysensteen	285	570	19	56218.0	56263.0	0.1	55794.3	93.2	0.8
Skanderborg	245	439	18	41808.0	41999.0	0.5	41790.4	9.6	0.5
Slagelse	1272	2220	57	162785.0	234527.0	44.1	229262.8	214.3	2.3
Soenderborg	284	728	27	82981.0	83643.0	0.8	82356.1	222.6	1.6
Solroed	451	727	49	81281.0	81796.0	0.6	81259.4	17.0	0.7
Struer	553	810	47	114970.0	114970.0	0.0	114970.0	0.0	0.0
Taarby	298	760	29	110985.0	110985.0	0.0	110490.0	0.0	0.5
Varde	229	675	30	97870.0	97870.0	0.0	94381.0	1606.0	3.7
Vejen	382	586	29	66128.0	66213.0	0.1	66047.0	80.2	0.3
Viborg	589	1036	43	117796.0	118660.0	0.7	117712.4	44.3	0.8
Viborg Handelsgym	359	593	21	62360.0	62743.0	0.6	62333.6	7.0	0.7
Viby	232	472	18	47448.0	47867.0	0.9	47448.0	8.6	0.9
Vordingborg	411	1288	53	162133.0	162601.0	0.3	162021.5	74.2	0.4
Average	374.0	812.0	35.9			1.2			0.6
Max	1272.0	6456.0	231.0			44.1			3.7

5 Conclusion

In this paper it has been shown how the Elective Course Student Sectioning can be formulated as an IP model and how ALNS has proven successful in establish solutions to the problem. The ALNS algorithm has been implemented in the cloud-based high school software system Lectio and is hence available for more than 200 different high schools in Denmark. It has been shown that the ALNS in average finds solution within 1% from the optimum, which is very satisfying results. The average is taken over 50 real-life instances. The main subject for future research is to use a MIP solver as an insertion method for the ALNS to get closer to optimal solutions. This could also improve the fairness of the distribution of the students.

References

- Balaprakash, P., M. Birattari, and T. Stützle. 2007. “Improvement strategies for the F-Race algorithm: sampling design and iterative refinement.” *Proceedings*

- of the 4th international conference on Hybrid metaheuristics, HM'07. Berlin, Heidelberg: Springer-Verlag, 108–122.
- Erben, W., and J. Keppler. 1996. “A genetic algorithm solving a weekly course-timetabling problem.” In *Practice and Theory of Automated Timetabling*, edited by Edmund Burke and Peter Ross, Volume 1153 of *Lecture Notes in Computer Science*, 198–211. Springer Berlin / Heidelberg.
- Kristiansen, S., M. Sørensen, T.R. Stidsen, and M.B. Herold. 2012. “The Consultation Timetabling Problem at Danish High Schools.” *Journal of Heuristics*, vol. To appear.
- Kristiansen, Simon, Matias Sørensen, and Thomas R. Stidsen. 2011. “Elective course planning.” *European Journal of Operational Research* 215 (3): 713 – 720.
- Laporte, G., R. Musmanno, and F. Vocaturro. 2010. “An Adaptive Large Neighbourhood Search Heuristic for the Capacitated Arc-Routing Problem with Stochastic Demands.” *Transportation Science* 44 (1): 125–135.
- Müller, T., and K. Murray. 2010. “Comprehensive approach to student sectioning.” *Annals of Operations Research* 181:249–269.
- Muller, L.F., S. Spoorendonk, and David Pisinger. 2011. “A hybrid adaptive large neighborhood search heuristic for lot-sizing with setup times.” *European Journal of Operational Research* Volume 218 (Issue 3): 614–623.
- Pillay, N. 2010. “An Overview of School Timetabling Research.” *Proceedings of the International Conference on the Theory and Practice of Automated Timetabling*. Belfast, United Kingdom, 321–335.
- Pisinger, D., and S. Ropke. 2005. “A general heuristic for vehicle routing problems.” *Computers & Operations Research* 34 (August): 2403–2435.
- Post, Gerhard, Luca Di Gaspero, Jeffrey H. Kingston, Barry McCollum, and Andrea Schaerf. 2012, August. “The Third International Timetabling Competition.” *Proceedings of the Ninth International Conference on the Practice and Theory of Automated Timetabling (PATAT 2012)*. Son, Norway.
- Ropke, S., and D. Pisinger. 2006. “An Adaptive Large Neighborhood Search Heuristic for the Pickup and Delivery Problem with Time Windows.” *Transportation Science* 40 (November): 455–472.
- Shaw, P. 1997. A New Local Search Algorithm Providing High Quality Solutions to Vehicle Routing Problems.
- Sørensen, Matias, Simon Kristiansen, and Thomas R. Stidsen. 2012. “International Timetabling Competition 2011: An Adaptive Large Neighborhood Search algorithm.” *Proceedings of the Ninth International Conference on the Practice and Theory of Automated Timetabling (PATAT 2012)*.
- Stegg, Jörg, and Michael Schröder. 2008. “A Hybrid Approach to Solve the Periodic Home Health Care Problem.” In *Operations Research Proceedings 2007*, edited by Jörg Kalcsics and Stefan Nickel, Volume 2007 of *Operations Research Proceedings*, 297–302. Springer Berlin Heidelberg.

Simulating FTR Strategy in New Zealand Electricity Market

M. Leon, Dr G. Zakeri & Dr A. Downward
Department of Engineering Science
University of Auckland
New Zealand
mleo029@aucklanduni.ac.nz

Abstract

The New Zealand Electricity Market uses spot prices to determine electricity price at different nodes. These prices can be different at different nodes due to congestion. If a *gentailer* has a generator at a cheaper node and allocated demand at a more expensive node, they risk losing profit or even getting a loss. A Financial Transmission Right (FTR) manages this risk. It gives holders the right to a payment of the difference in spot price between two nodes.

In May 2013, the first FTR auction will be held in New Zealand between Benmore and Otahuhu nodes. We have analysed different factors that may affect *gentailers'* profits with FTRs, including demand, generation cost and consumer price. From these factors we created a model that simulates risk averse strategies for *gentailers* in an FTR auction. In modelling these strategies, several assumptions have been made to allow us to gain intuition into the auction. These are excluding line losses in the network, competitive spot market and having only two *gentailers*.

A dispatch model was used to determine the spot prices at the two nodes at various demand levels. Then optimal FTR bid curves of the two *gentailers* were constructed from an optimization model, assuming risk aversion. Finally an FTR auction is simulated to derive profits for each *gentailer*.

Key words: FTR, electricity market, risk aversion

1 Introduction

The New Zealand Electricity Market uses spot prices to determine electricity price at different nodes. These prices can be different at different nodes due to congestion. If a *gentailer* has a generator at a cheaper node and allocated demand at a more expensive node, they risk losing profit or even getting a loss. A Financial Transmission Right (FTR) mitigates this risk by giving holders the right to a payment of the difference in spot price between two nodes. However FTRs are only advantageous when the designated path is in the same direction as the congested flow, i.e. the nodal price at the extraction point is greater than the nodal price at the injection point (Shahidehpour, Yamin, and Li 2002).

In May 2013, the first FTR auction will be held in New Zealand between Benmore and Otahuhu nodes (Electricity Authority 2012). The spot price differences between the two nodes can be seen in Figure 1. From Figure 1, we see that the price difference between the two nodes can be as high as \$86/MWh.

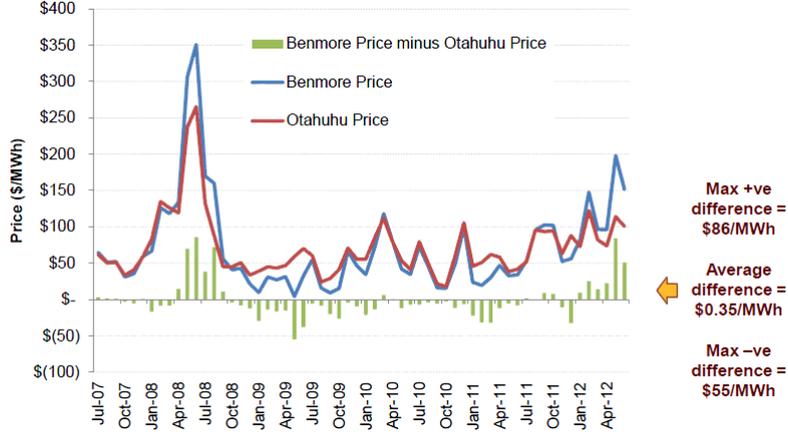


Figure 1: Comparison of spot prices in OTA and BEN (Electricity Authority 2012)

This paper highlights the effects of different factors to an FTR auction. These factors include demand, competitor behaviour, generation and retail prices. Before going into the auction, it is important to understand the electricity market dispatch model in New Zealand and how FTRs can affect *gentailers'* profits. We will show these effects using the following example. Figure 2 shows a two-node model with three generators. Suppose that *gentailer* 1 owns G_1 with an allocated demand (i.e.

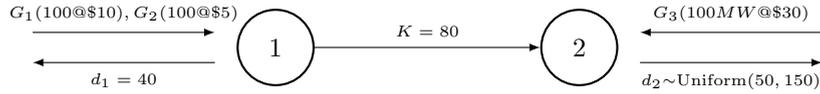


Figure 2: Two-node model example

they are obligated to purchase this amount from the spot market, to satisfy their retail load) at node 2, d_2 , with retail price p_{c1} . Demand at node 1, d_1 , is 40 MW and flow limit on the line, K , is 80 MW.

Without an FTR, our profit, $Z(d_2)$, can be calculated as

$$Z(d_2) = x_1(p_1 - c_1) - d_2(p_2 - p_{c1}) \quad (1)$$

where x_1 represents the amount generated by G_1 , p_1 represents the nodal price at node 1, c_1 represents generation cost of G_1 and p_2 represents the nodal price at node 2. Note that we assumed competitive spot market in this example, which means that the generator offer prices correspond to the generation costs only.

The term $x_1 p_1$ in the above equation represents gentailer 1's revenue for selling power at node 1, $x_1 c_1$ represents the generation cost of G_1 , $d_2 p_2$ represents the cost of purchasing power at node 2 and $d_2 p_{c1}$ represents the revenue from selling power to retail customers at node 2. Equation (1) shows that *gentailers* risk buying electricity for a higher price than they pay for when $p_2 > p_1$. If the *gentailer* has no generation

at node 2, there is no way of self supplying at node 2. An FTR can reduce this risk. With an FTR from node 1 to node 2, the profit equation becomes

$$Z(d_2, q) = x_1(p_1 - c_1) - d_2(p_2 - p_c) + q_F(p_2 - p_1) - q_F p_F \quad (2)$$

where q_F is the FTR quantity in MW and p_F is the price of FTR. The term $q_F(p_2 - p_1)$ represents the FTR payout while $q_F p_F$ represents the purchase cost of the FTR.

1.1 Dispatch Model

The nodal prices of electricity is set by the dispatch problem (Schweppe et al. 1988). The nodal prices can be defined as the cost of increasing demand at the node by another unit of electricity. The objective of the problem is to meet the energy demand at minimum cost. If demand is fixed, then this is equivalent to minimizing the cost of generation for a fixed demand. Line loss can also be included in the problem, making it a non-linear optimization problem. In this paper, we have assumed that there are no line losses to help us gain intuition into the risk averse strategies.

Indices

n = node: 1, 2; t = tranch: 1, ..., 5.

Parameters

M_{nt} = 1 if tranch t is located at node n , 0 otherwise;
 c_t = generation cost and offer price of tranch t ;
 q_t = quantity offered by tranch t ;
 d_n = demand at node n ;
 K = line capacity;
 A^T = $[-1, 1]$, node-arc incidence matrix.

Decision variables

x_t = amount of electricity cleared by tranch t ;
 f = electricity flow from node 1 to 2.

Model Dispatch Problem

$$\begin{aligned} \min \quad & c^T x \\ \text{subject to} \quad & Mx + Af = d & (3) \\ & -K \leq f \leq K & (4) \\ & 0 \leq x \leq q & (5) \end{aligned}$$

Explanation The objective is to minimise generation cost. Constraint (3) ensures demand at each node is satisfied. Constraint (4) ensures that flow does not exceed line capacity. Constraint (5) ensures that tranch capacity is not exceeded.

The shadow price of constraint (3) represents the nodal price at each node. The results of the dispatch problem for the scenarios from Figure 2 are shown in Figure 3. Figure 3a shows the generation amounts and flow while Figure 3b shows the nodal prices, as functions of the demand at node 2.

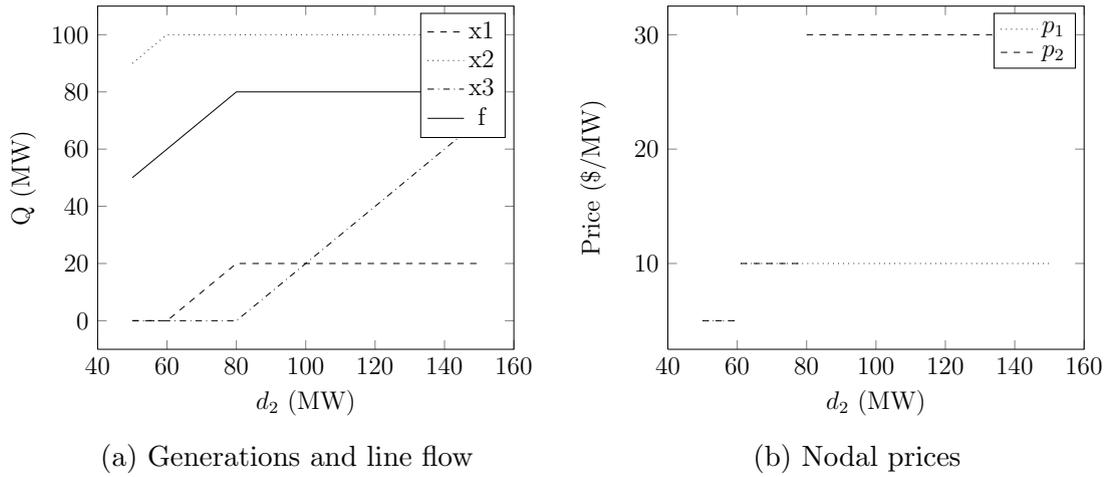


Figure 3: Dispatch Problem Results

When $d_2 \leq 60$ MW, the total demand in the network is less than 100 MW and this can be covered by G_2 , the cheapest generator. Thus nodal price at the two nodes are the same as the generation cost of G_2 , \$5. When d_2 exceeds 60 MW, G_1 starts getting utilized. The nodal prices at both nodes are the generation cost of G_1 , \$10.

When d_2 exceeds 80 MW, the line becomes congested. The cheaper generators at node 1 can no longer supply all the demand at node 2. While any increase of demand at node 1 can still be supplied by G_1 , demand increase at node 2 would need to be supplied by the more expensive G_3 . Thus the nodal price at node 2 increases to \$30.

1.2 FTR Payoffs

In this subsection, we look at how an FTR can affect a *gentailer's* profit. The profit equation with FTR was stated in equation (2) above. The equation uses the quantity of FTR, q_F , and the FTR price, p_F . These two values are determined from the FTR auction, which will be discussed in a later section of this paper. For now, we assume that $q_F = 30$ MW and $p_F = \$10/\text{MW}$. We also assume that allocated demand for *gentailer* 1 is half of all the demand at node 2, i.e. $d_2 = 0.5d_2$. The nodal prices are obtained from the dispatch problem model.

Figure 4 shows how an FTR affects a *gentailer's* profit. FTR holders only gain benefit from it when the nodal price between the two nodes are different. In this example, this happens when $d_2 > 80$ MW. When $d_2 \leq 80$ MW, *gentailer* with “FTR = 0” is better off as they are not obliged to pay the FTR price, p_F . There are two visible dips on the plot. The first occurs when d_2 is 60 MW while the second occurs when d_2 is 80 MW. The first dip occurred due to the increase in nodal price from \$5 to \$10. The second dip occurred due to the increase in nodal price from \$10 to \$30. In the “FTR = 30 MW” scenario, the second dip is not as significant because of the FTR payoff.

From this result, we found that FTR helps *gentailers* to manage risks. The maximum possible profit is reduced but the minimum possible profit is increased. This gives them a more stable return. Non FTR holders can expect to get profit as low as \$400 compared to \$600 for FTR holders.

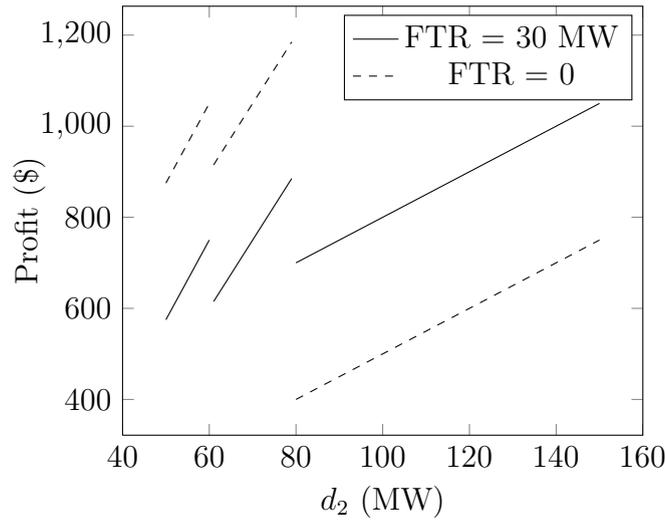


Figure 4: FTR Profit

2 Model

We have used this model based on an optimization on risk functions paper (Ruszczynski and Shapiro 2006) to simulate risk averse strategies for *gentailers* in an FTR auction. The optimization model has been used in simpler examples, such as the newsvendor problem (Choi and Ruszczynski 2008).

Indices

k = demand scenario: $1, \dots, N$.

Parameters

N = number of demand scenarios;

p_k = probability of scenario k happening;

$Z_k(q_F)$ = *gentailer's* profit at scenario k , function of variable q_F ;

β = worst case scenario level, a number between 0 and 1;

λ = risk aversion level, higher means more risk averse, $0 \leq \lambda \leq 1$.

Decision variables

q_F = FTR quantity to purchase;

η = Value at Risk;

v_k = deviation above the beta-quantile for scenario k ;

w_k = deviation below the beta-quantile for scenario k .

Model Risk Averse Genter

$$\begin{aligned} \max \quad & \sum_{k=1}^N p_k Z_k(q_F) - \frac{\lambda}{\beta} \sum_{k=1}^N p_k [(1 - \beta)w_k + \beta v_k] \\ \text{subject to} \quad & Z_k(q_F) = \eta + v_k - w_k, \quad k = 1, \dots, N & (6) \\ & w_k \geq 0, v_k \geq 0, \quad k = 1, \dots, N & (7) \\ & \sum_{k=1}^N p_k = 1 & (8) \end{aligned}$$

Explanation The objective is to maximise the risk-averse profit equation. Z_k is a linear function of FTR quantity, q , and FTR price, p_F .

By simulating the model at different values of p_F , we can find the volume of FTR a *gentailer* is willing to purchase at different prices. The result for the two node example is shown in Figure 5. Table 1 shows the quantities that the *gentailer* would bid in the auction, assuming competitive bids. It shows that when FTR price is \$11, the optimal FTR quantity the *gentailer* should purchase is 26.375 MW, at \$12, 25.5 MW, etc. Thus we can create a bid curve shown in column “bid q_F ”. If for example the FTR clearing price is \$12, the *gentailer* would obtain $0.5 + 1.125 + 1.125 + 22.75 = 25.5$ MW of FTR.

The FTR clearing price is set from the auction. The objective of the auction is to maximize FTR revenues which is the same as maximizing the aggregate benefit function of the buyers (Zakeri and Downward 2011). *Gentailers* commit to bids quantities and prices and the prices are set to the marginal clearing bids for each FTR.

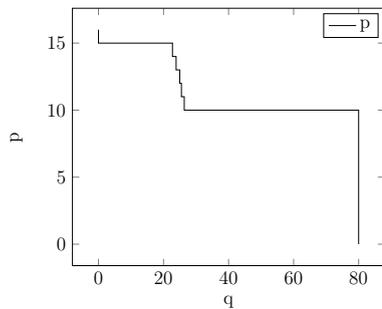


Figure 5: p_F vs q_F

p_F	q_F	bid q_F
16	0	0
15	22.75	22.75
14	23.875	1.125
13	25	1.125
12	25.5	0.5
11	26.375	0.875
10	80	53.625

Table 1: Bid Quantities

3 Results

The risk averse newsvendor model has been simulated over different line capacities, risk aversion levels and demand allocations. There are two *gentailers* involved. The first subsection shows results on two *gentailers* with allocated demand at the same node while the second subsection shows results on allocated demand at opposite nodes.

3.1 Demand at same node

In this scenario, we have set the following $K = 80$, $p^T = [10 \ 5 \ 30]$, $q^T = [100 \ 100 \ 100]$, $pc_1 = 40$, $pc_2 = 35$, $N = 101$, $d_1 = 20$, $d_2 \sim \text{Uniform}(50, 150)$, $\lambda_1 = \lambda_2 = 5$, $\beta_1 = \beta_2 = 0.05$.

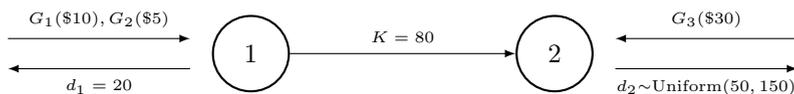


Figure 6: Model for demand at same node example

Gentailer 1 owns generator 1 at node 1, *gentailer* 2 owns generator 2 at node 1 and generator 3 is owned by a third party at node 2. Both *gentailers* have allocated demand split equally at node 2. We have varied λ and K to see how they affect the *gentailers* bids and hence the auction clearing prices.

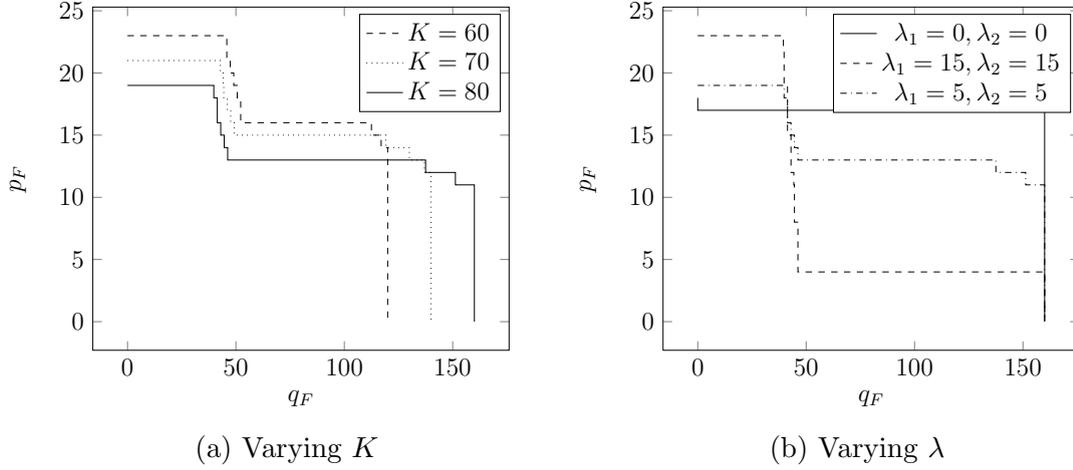


Figure 7: Combined bids of two gentailers

Figure 7a shows that as the line capacity is reduced, the FTR price increases. The price difference between the two nodes becomes more likely and FTR volume is also limited which makes FTR more desirable.

The fair price of FTR can be calculated as the average price difference between the two nodes. In this example it is found to be around \$18. At this price, the *optimal quantity* for different level of risk aversions would be the same. Figure 7b shows that when both *gentailers* are more risk neutral ($\lambda = 5$), they are willing to pay more for FTR quantities greater than the *optimal quantity*. When both *gentailers* are risk averse ($\lambda = 15$), they are willing to pay more for FTR quantities less than the *optimal quantity*.

3.2 Demand at opposite nodes

In this scenario, we created a different model from the previous subsection. We have set the following $K = 50$, $p^T = [5 \ 20 \ 30 \ 35]$, $q^T = [100 \ 100 \ 100 \ 100]$, $pc_1 = 30$, $pc_2 = 30$, $N = 101$, $\lambda_1 = \lambda_2 = 5$, $\beta_1 = \beta_2 = 0.05$. Nodal demands d_1 and d_2 are perfectly correlated with $d_1 = 10 + 20U_1$ and $d_2 = 40 + 80U_1$ where $U_1 \sim \text{Uniform}(0, 1)$.

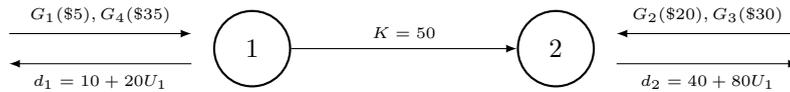


Figure 8: Model for demand at opposite node example

Gentailer 1 owns generator 1 at node 1 and allocated demand at node 2, *gentailer* 2 owns generator 2 at node 2 and allocated demand at node 1. Generators 3 and 4 are owned by a third party and located at nodes 2 and 1 respectively. We have varied λ and K to see how they affect the auction.

Figure 9a shows that reducing line capacity has the same effect as the previous example. It increases FTR price for the same quantity. Figure 9b shows that when

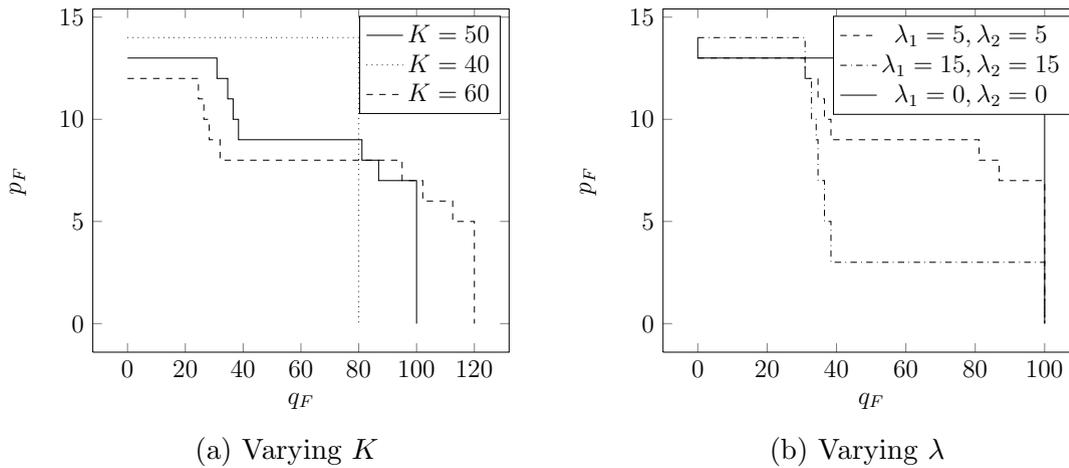


Figure 9: Combined bids of two gentailers

the quantity is greater than the quantity at fair price, a more risk neutral *gentailer* is more willing to pay more than a risk averse *gentailer* and vice versa.

The difference of this example to the previous one is that *gentailer* 1 needs FTR from node 1 to 2 more than *gentailer* 2. At maximum demand, $d_1 = 30$ and $d_2 = 120$, the nodal prices are \$5 and \$20 at nodes 1 and 2 respectively. *Gentailer* 2 does not need the FTR as the price at node 1 is already cheaper. However a more risk neutral *gentailer* 2 might still buy the FTR to gain profit when nodal price difference is greater than the FTR price. *Gentailers* offers for both examples are shown in Figure 10. It shows that when the *gentailers* have allocated demands at different nodes, one of the *gentailers* is willing to pay more than the other.

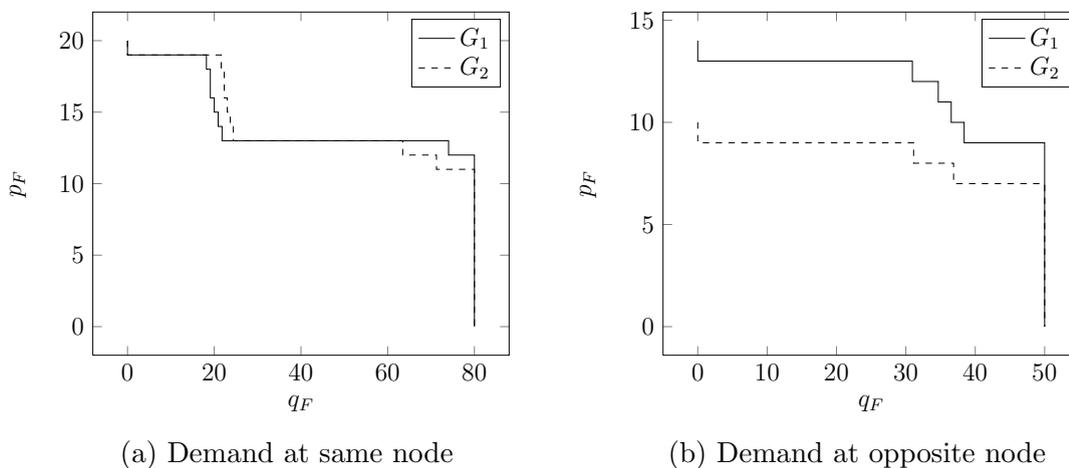


Figure 10: Bid curves for each example

4 Future Work

We will continue with this project by using historical electricity market data to test the model, which can be obtained online (WITS 2012).

We also noted that it is possible for an FTR auction to be revenue inadequate. This is when the money obtained from an FTR sale is not enough to cover the FTR payoffs. In this situation, the FTR payoff may need to be scaled down. We can then

create a new bid curve with the new payoff and simulate a new auction. This can then be done iteratively.

5 Conclusions

Motivated by a risk-averse newsvendor formulation (Ruszczynski and Shapiro 2006), we have constructed risk-averse strategies for *gentailers* in an FTR auction. We then simulated these strategies in a two node and two *gentailers* model, varying demand allocations, risk aversion level and line capacities. The results are as we expected, i.e.

- As line capacities decrease, the price of FTR increases.
- Risk averse *gentailers* are willing to pay more when the available FTR quantity is less than the *optimal quantity* at a fair price and pay less when the available FTR quantity is greater.
- When two *gentailers* have allocated demand at opposite nodes, one of the *gentailers* needs the FTR more than the other as shown in Figure 10.
- When both *gentailers* have allocated demand at the same node, FTR auction becomes more competitive, i.e. both *gentailers* need the FTR as much as the other.

Future work will be to test the model using real data from New Zealand electricity market. We will also look at how revenue inadequacy affects the *gentailer's* bid and the consecutive auction.

Acknowledgments

I would like to thank Dr. Golbon Zakeri and Dr. Anthony Downward for their supervisions and inputs throughout this work.

References

- Choi, S., and A. Ruszczyński. 2008. "A risk-averse newsvendor with law invariant coherent measures of risk." *Operations Research Letters* 36:77–82.
- Electricity Authority. 2012, August. Overview of FTRs.
- Ruszczynski, Andrzej, and Alexander Shapiro. 2006. "Optimization of Convex Risk Functions." *Mathematics of Operations Research* 31 (3): 433–452.
- Schweppe, F. C., M. C. Caramanis, R. D. Tabors, and R. E. Bohn. 1988. *Spot pricing of electricity*. Kluwer Academic Publishers.
- Shahidehpour, M., H. Yamin, and Z. Li. 2002. *Market Operations in Electric PowerSystems*. Wiley-Interscience.
- WITS. 2012, November. WITS Free to Air @ONLINE.
- Zakeri, Golbon, and Tony Downward. 2011. "A Literature Survey of Financial Transmission Rights Auctions."

An Application of Data Envelopment Analysis to External Radiotherapy Treatment Planning

Kuan-Min Lin¹, John Simpson^{1,2}, Giuseppe Sasso^{3,4}, Andrea Raith¹ and Matthias Ehrgott¹

¹ Department of Engineering Science, University of Auckland, Auckland, New Zealand

² Auckland Radiation Oncology, Auckland, New Zealand

³ Auckland City Hospital, Auckland, New Zealand

⁴ Discipline of Oncology, School of Medical Sciences, University of Auckland, Auckland, New Zealand

km.lin@auckland.ac.nz

Abstract

The iterative trial-and-error nature of radiotherapy treatment planning results in an inefficient planning process and in order to reduce such inefficiency, plans can be accepted without achieving the best attainable quality. We propose a quality assessment method based on Data Envelopment Analysis (DEA) to address this inefficiency. This method compares a plan of interest to a set of past delivered plans and searches for evidence of potential further improvement. With the assistance of DEA, planners will be able to make informed decisions on whether further planning is required and ensure that a plan is only accepted when the plan quality is close to the best attainable quality. We apply the DEA method to assess the quality of 37 prostate plans in terms of rectal sparing and target coverage. Five plans that are considered of lesser quality by DEA are re-optimized with the goal to further improve rectal sparing. After re-optimization, all five plans improve in rectal sparing without clinical significant deterioration in target coverage. The results demonstrate that DEA can correctly assess the potential for further improvement in terms of the chosen input and output parameters.

Key words: Radiotherapy treatment planning, Data envelopment analysis

1 Introduction

Recent advancements in radiotherapy technologies such as intensity modulated arc therapy (IMAT) (Yu & Tang, 2011), image-guided radiotherapy (Xing et al., 2006) and adaptive radiotherapy (Wu et al., 2008) have offered great therapeutic benefits in cancer treatment. However, despite this progress, maximizing the benefits afforded by them can be challenging due to the difficulties associated with radiotherapy treatment planning.

Radiotherapy treatment planning involves managing several conflicting objectives related either to the planning target volume (PTV) or healthy organs at risk (OARs). The major commercial treatment planning systems manage these conflicting objectives through scalarization. Radiotherapy treatment planning in this approach involves a planner entering a number of plan objectives into the treatment planning system. Note that the chosen plan

objectives may not be the same as those used to evaluate the plan acceptability, but rather selected based on the personal experience of the planner. Each of these objectives is associated with an importance score (weight). The treatment planning system derives treatment plans based on the plan objectives and the associated weights. During the planning process, the planner iteratively adjusts the plan parameters (i.e. the objective weights and/or the objectives) and generates new plans in order to find a plan with the best quality. However, because the exact effects of changing the plan parameters cannot be known a priori, it is hard for the planner to verify if there is further potential to improve a plan. If a plan is deemed to be of inadequate quality it would take further time to produce another plan, without knowing in advance whether the new plan is going to be superior to the previous plan. This trial-and-error aspect of the planning process is inefficient and in order to reduce such inefficiency, plans can be accepted without achieving the full potential of the available technology.

This planning dilemma can be addressed by comparing the plan quality against past plans. By doing so, the planners will have a better knowledge of what is achievable and thus can make informed decision on whether further improvement is possible. A number of plan assessment approaches that use past plans as references have been proposed in the literature (Moore et al., 2011, Zhu et al., 2011, Wu et al., 2009, Hunt et al., 2006). These quality assessment approaches predict the achievable OAR sparing based on the geometrical relationship of the PTV and the OAR. Thus, given a particular geometrical relationship, if a newly generated plan has a significantly higher dose to the OAR than the predicted dose, the plan is considered of inadequate quality and re-planning is required. However, as these approaches do not consider the dose to the PTV, one may unintentionally conclude that more OAR sparing is available without realizing that the improvements in OAR sparing would likely deteriorate the PTV dose coverage. This might create additional planning inefficiency since a high quality plan with good PTV coverage and acceptable OAR sparing may be considered of inadequate quality simply because it does not achieve the maximal OAR sparing.

In this study we propose using Data Envelopment Analysis (DEA) (Cooper, Seiford & Zhu, 2011, Charnes, Cooper & Rhodes, 1978) to assess the plan quality. DEA is a management science method for assessing the performance of a set of decision-making units (DMUs) that convert inputs into outputs. In a loose economic interpretation, the inputs represent the cost we pay for producing outputs. The concept of DEA is directly applicable to the problem of assessing treatment plan quality in radiotherapy in which the doses to OARs are considered as the cost we pay for delivering dose to the PTV. One of the most valuable strengths of DEA is its ability to handle multiple parameters. This strength makes DEA ideal for radiotherapy plan assessment in which several conflicting planning criteria need to be considered. DEA has been applied in performance assessment of healthcare systems (Chilingerian & Sherman, 2011), including formative evaluation of radiotherapy services (Santos & Amado, 2012) and even to compare prostate cancer treatment options (Ramer, Holder & Papanikolaou, 2008). However, to the best of our knowledge, it has never been used for case-based quality assessment for radiotherapy treatment planning.

The purpose of this work is to demonstrate how DEA can be applied to assess the quality of radiotherapy treatment plans. In section 2, we introduce the DEA model used in this study. In section 3 we present a case study where DEA is applied to assess the quality of prostate radiotherapy plans. The results and discussion are presented in section 4 and 5 respectively.

2 Introduction to DEA model

In this section we introduce the DEA model used in this study. For a detailed introduction to DEA we refer the interested readers to chapter 6 of Coelli et al. (2005) and to Cooper et al. (2011).

We use an input-oriented variable-return-to-scale DEA model in the envelopment form with additional non-discretionary output variables that take the environmental factors into account. Assume there are I DMUs each with N inputs, M outputs and L environmental variables. For the i th DMU the inputs and the outputs are represented by vectors $x^i \in \mathbb{R}^N$, $q^i \in \mathbb{R}^M$ and $z^i \in \mathbb{R}^L$, respectively. The data for all I DMUs can be represented by the input matrix $X \in \mathbb{R}^{N \times I}$, the output matrix $Q \in \mathbb{R}^{M \times I}$ and the environmental matrix $Z \in \mathbb{R}^{L \times I}$ in which the i th column contains the data for the i th DMU.

The efficiency score θ^i for the i th DMU is derived by solving the following DEA linear programming model:

$$\begin{aligned}
 & \min \theta^i \\
 & s. t. \quad -q^i + Q\lambda \geq 0 \\
 & \quad \theta^i x^i - X\lambda \geq 0 \\
 & \quad -z^i + Z\lambda \geq 0 \\
 & \quad e^T \lambda = 1 \\
 & \quad \lambda \geq 0,
 \end{aligned} \tag{1}$$

in which $\theta^i \in \mathbb{R}$ and $\lambda \in \mathbb{R}^I$ are the decision variables. This model is solved I times, once for each of the I DMUs. The optimal solution of (1) is θ^{i*} and λ^{i*} , where θ^{i*} is the efficiency score of the i th DMU and λ^{i*} is a vector of weights. The formulation attempts to maximally scale down the input vector x^i while satisfying all the constraints in (1). In the circumstance where no scaling is possible, i.e. when $\theta^{i*} = 1$, the i th DMU is identified as fully efficient.

Conceptually, scaling the input vector x^i shifts the DMU i to a projected point $(Q\lambda^{i*}, X\lambda^{i*})$ on the efficient frontier, which is a multidimensional surface formed by all of the points representing efficient DMUs. The projected point $(Q\lambda^{i*}, X\lambda^{i*})$ is also referred to as the target for the i th DMU. The target represents the inputs and outputs that the i th DMU should aim for to make itself efficient. The existing DMUs which contribute to the target are referred to as the peers of the i th DMU. For an efficient DMU, the DMU itself is its target as well as its only peer.

In this study we assume that higher values of the environmental variables are likely to impair the efficiency of the DMUs. The constraint $-z^i + Z\lambda \geq 0$ ensures that the target $(Q\lambda^{i*}, X\lambda^{i*})$ has higher or equal values of the environmental variables than those of the i th DMU. Since high values of the environmental variables are considered unfavourable to the efficiency, the target indicates what the i th DMU can potentially achieve even when influenced by the same or worse environmental situation.

3 Application of DEA to Prostate Radiotherapy Treatment Plans

One of the most difficult tasks in prostate radiotherapy treatment planning is managing the dose delivered to the PTV and the rectum. The rectum is usually the OAR that most influences the ability to achieve the optimal dose to the PTV. Therefore, in this preliminary study, we only consider the dose delivered to the PTV and the rectum. The goal is to generate a dose distribution that matches the prescription dose in the PTV as closely as possible while maximizing rectal sparing. In a loose economic interpretation, the dose delivered to the

rectum is considered as the cost for delivering dose to the PTV. Specifically, we use D95 (the minimum dose that is received by 95% of the volume of a structure) of the PTV as the output and generalized equivalent uniform dose (gEUD) (Niemierko, 1999) of the rectum as the input for the DEA model. In addition, we use the percentage volume of the rectum that overlaps the PTV as an environmental variable. The higher the overlap the more difficult it is to achieve good PTV coverage and low OAR dose simultaneously. While many other dose descriptors or anatomical descriptors can be alternatively used for the DEA model, it is out of the scope of this study to investigate the most preferable parameters. Instead, we empirically select these parameters and focus on investigating the validity of using DEA as a quality assessment method in radiotherapy treatment planning.

We use an input oriented model for the analysis since we are interested in maximal OAR sparing available for a given dose to the PTV. Mathematically, we may assume constant return to scale (CRS) for the model since we are able to obtain a constant return of PTV D95 for each unit change of rectum gEUD by scaling the entire dose distribution. However, because our DEA model only considers PTV D95 and rectal gEUD, scaling the entire dose distribution may result in undesirable dose, such as hot spots in other OARs or cold spots in the PTV. This implies that the target suggested by a CRS model may include plans that are not clinically acceptable. Therefore, instead of assuming CRS, we use the assumption of variable return to scale (VRS), in which the efficient points are approximated using linear combinations of DMUs that are most preferable for the corresponding input level. Since these DMUs with different input levels are all clinically acceptable plans, a VRS model offers a better approximation of what is clinically attainable for a given input level than a CRS model.

A series of 37 anonymized clinically intact prostate treatment plans were provided by Auckland Radiation Oncology, following approval and guidelines of New Zealand Health & Disabilities Ethics Committees for observational study. All plans were the actual plans used for the subsequent delivery of treatment and were generated using the same treatment planning system over a 1 year period utilising the same plan acceptability criteria. All plans were planned for Volumetric Modulated Arc Therapy (VMAT) delivery with Pinnacle v9 and the Smartarc module (Philips, Netherlands) using a single 360 degree arc. The plan parameter values were extracted using CERR (Deasy, Blanco & Clark, 2003). We used an in-house DEA software package, pyDEA (Raith et al., 2012), to assess the efficiency of these 37 prostate plans. After obtaining the results from the analysis, five plans were considered significantly inefficient and selected for re-optimization. The input/output parameter values of the selected plans before and after re-optimization are included in table 1. Re-optimized plans are indicated by the original plan ID with an asterisk. Each of the selected plans has a percentage overlap volume significantly different from the other selected plans. These plans are selected in order to test the ability of DEA in assessing plans with variations in anatomical structure relationships. A planner was instructed to further improve rectal sparing while maintaining overall clinical acceptance for the selected plans without access to the results of DEA. After re-planning, the re-optimized plans were included in the dataset and the DEA analysis repeated. The purpose of re-optimization is to investigate the usefulness of DEA in identifying potential improvements in terms of the chosen input and output parameters.

4 Results

Table 1. The efficiency scores and the input/output parameter values of the selected plans. Re-optimized plans are indicated by the original plan ID with an asterisk.

Plan ID	Efficiency (original)	Efficiency (re-optimized)	Output: D95 PTV (Gray)	Input: Rectal gEUD (Gray)	Fractional overlap
10	0.964	0.955	71.575	63.457	0.065
10*	N/A	0.987	71.525	61.314	0.065
19	0.975	0.963	71.725	62.562	0.052
19*	N/A	0.987	71.325	60.710	0.052
26	0.980	0.980	71.325	63.545	0.110
26*	N/A	0.991	71.025	62.181	0.110
31	0.978	0.968	71.425	61.522	0.038
31*	N/A	1.000	71.725	59.567	0.038
35	0.966	0.960	71.425	63.570	0.080
35*	N/A	0.993	71.525	61.658	0.080

The efficiency scores before and after including re-optimized plans as well as the input/output values used in DEA for the selected plans are provided in table 1. Plans 10, 19, 26, 31 and 35 were identified as substantially inefficient for their range of percentage overlap volume and were re-optimized. The re-optimization of plan 31 produced an additional efficient plan in the dataset. This plan extends the efficient frontier slightly and results in lower or equal efficiency scores of all other plans compared to those of the original dataset. The efficiency score of the re-optimized plans are higher than those of the original plans, with an average improvement of 0.026 units. Note that this improvement is quite significant since the standard deviation of the efficiency scores is only 0.012.

The original, re-optimized and the re-optimized target values for the selected plans are summarized in table 2. We do not include the target information of the original dataset since we consider the target information of the re-optimized dataset a better approximation for the true efficient frontier. The overlap fractions are not included since they are the same as the corresponding values in table 1. Although planners were not provided with the results of DEA, the values for the re-optimized plans are very close to the target of re-optimized plan, with a maximum difference of 0.806 Gy (table 2). Note that the target of re-optimized plan represents DEA's prediction of best achievable plan. This minor parameter difference between the prediction and the re-optimized plan verifies the ability of DEA in predicting potential improvement in terms of the chosen input and output parameters.

Re-optimization of the five plans resulted in an average reduction of 1.84 Gy in rectal gEUD with only an average reduction of 0.07 Gy in PTV D95 (table 2). Figure 1 shows the original and re-optimized DVHs for plan 10. For all five plans, the improvement in rectal sparing is considerable while there are no clinically significant differences between original PTV coverage and re-optimized PTV coverage. However, given the single input and output used here, it is not to say that the re-optimized plans are superior to the original plans in every clinical aspect. Despite this, the fact that planners could be instructed to achieve better rectal sparing without compromising the target coverage and subsequently were able to do so, is a positive finding.

Table 2. The original, re-optimized and the corresponding target parameter values for the selected plans. The measurement unit for the parameters is Gray.

Plan	Parameter	Original	Re-optimized	Target (re-optimized)	Dose reduction ^a	Prediction error ^b
10	Rectal gEUD	63.457	61.314	60.508	2.143	0.806
	D95 PTV	71.575	71.525	71.525	0.050	0.000
19	Rectal gEUD	62.562	60.710	59.930	1.852	0.780
	D95 PTV	71.725	71.325	71.571	0.400	0.246
26	Rectal gEUD	63.545	62.181	61.632	1.364	0.549
	D95 PTV	71.325	71.025	71.025	0.300	0.000
31	Rectal gEUD	61.522	59.567	59.567	1.955	0.000
	D95 PTV	71.425	71.725	71.725	-0.300	0.000
35	Rectal gEUD	63.570	61.658	61.211	1.912	0.447
	D95 PTV	71.425	71.525	71.525	-0.100	0.000
Avg	Rectal gEUD	62.931	61.086	60.570	1.845	0.516
	D95 PTV	71.495	71.425	71.474	0.070	0.049

^a Dose reduction is calculated as original parameter value minus re-optimized parameter value.

^b Prediction error is the absolute difference between the re-optimized target parameter value and the re-optimized parameter value.

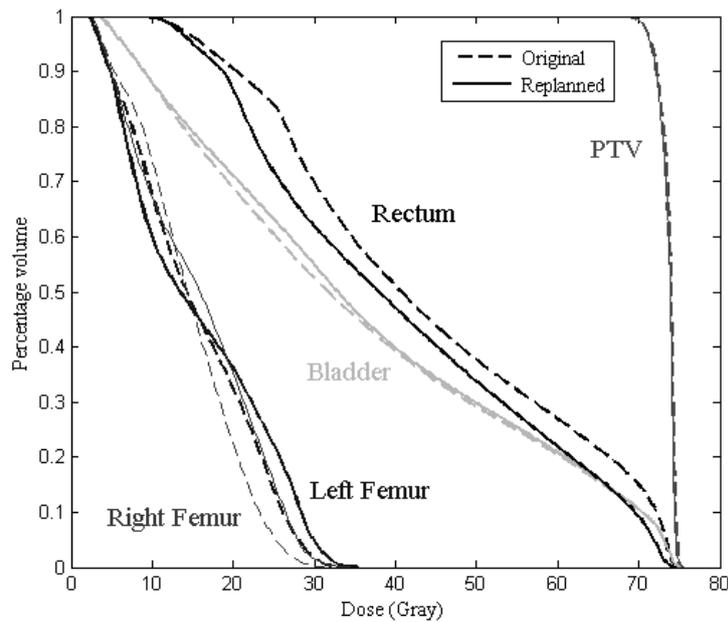


Figure 1. The original and re-optimized DVHs of plan 10.

5 Discussion

In this study we investigate the validity of using DEA as a quality assessment tool for radiotherapy planning. We use DEA to assess the quality of 37 prostate plans using an input-oriented VRS model with rectal gEUD as the input, PTV D95 as the output and the percentage volume of rectum overlapping PTV as a non-discretionary output variable. Five plans that are considered of low quality by DEA are re-optimized with the goal to further improve the dose to rectum while maintaining overall clinical acceptance. After re-planning, the dose to the rectum for all five plans improved considerably without clinically significant deterioration in PTV coverage. In addition, the input and output parameters for the re-optimized plans are very close to DEA's prediction of best attainable plan, with a maximum difference of 0.8 Gray. These results confirm that DEA is capable of identifying plan improvement potential and predicting the best attainable plan in terms of the chosen input and output parameters.

There are several advantages of using DEA method as a quality assessment tool for radiotherapy treatment planning. Firstly, the non-parametric nature of DEA does not require practitioners to assume a function form for the frontier. This allows practitioners to select the parameters that are considered most relevant in assessing the quality of radiotherapy treatment plans without too much concern on the underlying relationship among these parameters. Secondly, DEA can handle multiple inputs and multiple outputs easily whereas statistical methods cannot be easily applied to multiple inter-related parameters. This advantage allows plan assessment based on multiple planning criteria and therefore captures the conflicting nature of treatment objectives more adequately. Thirdly, DEA constructs an efficient frontier based on the best results in the dataset. This is distinctly different to regressions analysis that attempt to fit the regression function at the centre of the data spread and provide estimations for the "average" attainable results rather than the best attainable results. In this radiotherapy application, since we are interested in the best attainable results, we consider DEA a more preferable method than the regression methods. Fourthly, DEA not only provides the efficiency score for the plan being assessed, but also target information, including the peers and the corresponding weights. A treatment planner can compare the target with the treatment plan being assessed and decide if further planning is required. If the target is largely composed from a particular peer, a treatment planner can trace back to the peer, assess how the peer is derived and perhaps conduct the re-optimization using similar treatment objectives and/or objective weights. In addition, DEA's capability to accurately predict the best attainable dose for both the PTV and the OAR allows planners to set achievable plan objectives and thus reduce the trial-and-error attempts required to find a satisfactory plan, as suggested by Wu et al. (2011). Last but not least, DEA is readily available in many software packages (Barr, 2004) and can be conducted independently of the treatment planning system with negligible computational effort. This provides clinics with an approach to improve planning efficiency and plan quality without the need to change the treatment planning system.

While DEA offers many advantages, it is not without some potential limitations. One limitation is that the efficiency score for a plan is a relative measure compared to other plans in the dataset. Thus a plan rated fully efficient for a dataset might not be truly an optimal plan, but simply a superior plan compared to the plans in the dataset. However, as more efficient plans are generated and included in the dataset, DEA will be able to learn from the plans and will be able to approximate the true efficient frontier more accurately. As a result, this

limitation would become less significant over time. Another limitation is that, as more inputs and outputs are included in the formulation, DEA starts to lose discrimination power on the performance of the DMUs. Introducing more inputs and outputs imposes more constraints in the formulation. As a consequence, more DMUs will be deemed efficient or close to be efficient. To address this limitation, in this study we only include the most relevant objectives in the formulation, i.e. PTV coverage and rectal sparing, while ensuring other objectives are clinically acceptable. Other practical approaches that can improve the discrimination power of DEA are offered by Podinovski and Thanassoulis (2007). The last limitation is that a plan can be rated efficient simply because it has an optimal value in one of the DEA input/output parameters compared to all other plans. This limitation can be alleviated by checking if the plan is referred to as a peer for other plans. In general, given a database of reasonable size, an efficient plan with preferable output to input ratio is likely to be used as a peer for another plan. On the contrast, a plan considered efficient simply because it has an optimal value in one of the input/output values is usually not used as a peer for other plans. Thus by checking the peer counts, we can effectively identify efficient plans that may not be truly desirable.

Further investigation is required to extend the DEA model to include more and/or other plan assessment criteria. In this preliminary study we use PTV D95 and rectal gEUD to account for the dose to the PTV and the rectum, respectively. A proper assessment of prostate plans requires the assessment of several other OARs such as bladder and femur heads. However, these are generally unchallenging organs to spare and, in our opinion, would not require DEA analysis. It may however in the case of prostate planning be useful to include second PTV or rectal parameters such as PTV conformity index or rectal V70Gy (percentage volume of a structure that receives at least 70 Gy of radiation). As previously discussed, adding to the number of plan assessment criteria in the DEA model degrades plan quality discrimination. In future research, we will investigate the most effective plan assessment criteria that should be included in the DEA model, followed by an investigation on how these plan assessment criteria can be incorporated in the DEA model while maintaining sufficient discrimination on the quality of the plans. Other treatment sites likely to require a larger number of inputs and outputs such as head and neck will also be investigated in future work.

Acknowledgement

We thank Dayan Loria from Auckland Radiation Oncology for conducting the re-optimization.

6 Reference

- Barr, R., 2004. "Dea software tools and technology." *In*: ed Cooper, W.W., L.M. Seiford., and J. Zhu. *Handbook on Data Envelopment Analysis*, Springer US, New York.
- Charnes, A., W.W. Cooper., and E. Rhodes. 1978. "Measuring the efficiency of decision making units." *European Journal of Operational Research*, **2**:429-444.
- Chilingerian, J.A., H.D. Sherman., 2011. "Health-care applications: from hospitals to physicians, from productive efficiency to quality frontiers." *In*: ed Cooper, W.W., L.M. Seiford., and J. Zhu. *Handbook on Data Envelopment Analysis*. Springer US, New York.
- Coelli, T.J., D.S.P. Rao., C.J. O'donnell., and G.E. Battese. 2005. *An introduction to efficiency and productivity analysis*, Springer US, New York.

- Cooper, W.W., L.M. Seiford, and J. Zhu. 2011. "Data envelopment analysis: history, models, and interpretations." In: ed Cooper, W.W., L.M. Seiford, and J. Zhu. *Handbook on Data Envelopment Analysis*. Springer US, New York.
- Deasy, J.O., A.I. Blanco, and V.H. Clark. 2003. "CERR: a computational environment for radiotherapy research." *Medical Physics*, **30**:979-985.
- Hunt, M.A., A. Jackson, A. Narayana, and N. Lee. 2006. "Geometric factors influencing dosimetric sparing of the parotid glands using IMRT." *International Journal of Radiation Oncology Biology Physics*, **66**:296-304.
- Moore, K.L., R.S. Brame, D.A. Low, and S. Mutic. 2011. "Experience-based quality control of clinical intensity-modulated radiotherapy planning." *International Journal of Radiation Oncology Biology Physics*, **81**:545-551.
- Niemierko, A., 1999. "A generalized concept of equivalent uniform dose." *Medical Physics*, **26**:1100.
- Podinovski, V., E. Thanassoulis. 2007. "Improving discrimination in data envelopment analysis: some practical suggestions." *Journal of Productivity Analysis*, **28**: 117-126.
- Raith, A., K. Harton, A. Lee, H. Priddey, and M. Rouse. 2012. "pyDEA - A Software package and user interface for DEA." Department of Engineering Science, The University of Auckland, Auckland.
- Ramer, R., A. Holder, and N. Papanikolaou. 2008. "Utilization of data envelopment analysis (DEA) to compare prostate treatment options." *Medical Physics*, **35**: 2815.
- Santos, S., C.F. Amado. 2012. "Using data envelopment analysis for formative evaluation of radiotherapy services: an exploratory study." In: ed Tànfani, E., A. Testi. *Advanced Decision Making Methods Applied to Health Care*. Springer Milan, Verlag Italia.
- Wu, B., F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, M. Chuang, R. Taylor, R. Jacques, and T. McNutt. 2009. "Patient geometry-driven information retrieval for IMRT treatment plan quality control." *Medical Physics*, **36**: 5497-5505.
- Wu, B., F. Ricchetti, G. Sanguineti, M. Kazhdan, P. Simari, R. Jacques, R. Taylor, and T. McNutt. 2011. "Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning." *International Journal of Radiation Oncology Biology Physics*, **79**:1241-1247.
- Wu, Q.J., D. Thongphiew, Z. Wang, B. Mathayomchan, V. Chankong, S. Yoo, W.R. Lee, and F.-F. Yin. 2008. "On-line re-optimization of prostate IMRT plans for adaptive radiation therapy." *Physics in Medicine and Biology*, **53**:673-691.
- Xing, L., B. Thorndyke, E. Schreibmann, Y. Yang, T.-F. Li, G.-Y. Kim, G. Luxton, and A. Koong. 2006. "Overview of image-guided radiation therapy." *Medical Dosimetry*, **31**:91-112.
- Yu, C.X., G. Tang. 2011. "Intensity-modulated arc therapy: Principles, technologies and clinical implementation." *Physics in Medicine and Biology*, **56**:R31-R54.
- Zhu, X., Y. Ge, T. Li, D. Thongphiew, F.-F. Yin, and Q.J. Wu. 2011. "A planning quality evaluation tool for prostate adaptive IMRT based on machine learning." *Medical Physics*, **38**:719-726.

Linear Optimization over the Nondominated Set of a Multiobjective Linear Programming Problem

Zhengliang Liu, Matthias Ehrgott, Andrea Raith
Department of Engineering Science
University of Auckland
New Zealand
ethan.liu@auckland.ac.nz

Abstract

An algorithm is designed to obtain the optimal solution of a linear function $f(y)$ over the nondominated set Y_N of a multiobjective linear programming problem $\min\{Cx : x \in X\}$. The modified version of an outer approximation algorithm is employed to partially generate the nondominated set in the objective space, during which the values of the function f at new vertices are evaluated and regarded as a guide to determine which vertex is to be selected for the next iteration. A new type of cut is introduced not only to cope with local optima but also to speed up the outer approximation process. Several existing outcome based algorithms are briefly reviewed. Randomly generated problems are used to test and compare our algorithm and other algorithms including the brute force method, which enumerates all of the nondominated extreme points. The results show the merit of the new algorithm for large problems. Determining the nadir point, as one of the applications of the algorithm, is discussed in the end.

Key words: Multiobjective linear optimization, Nondominated set, Outer approximation, Nadir point.

1. Introduction

Optimization over the efficient set of a multiobjective linear programming problem, well-known as a hard global optimization problem, has been attracting the attention of researchers for decades. Since it was first considered by Philip (1972), various types of algorithms have been developed to solve this problem. Most of these algorithms are based in decision space, and are reviewed in Yoshitsugu (2002). This article categorizes existing algorithms into several types of methods, such as adjacent and Nonadjacent vertex search algorithms, face search algorithm, Lagrangian relaxation method, dual approach, bisection search algorithm, etc. However, expensive global optimization techniques used in those methods are problematic in applications. Especially when the dimension of the problem in decision space is large, computational efforts made to solve the problem are overwhelmingly burdensome. Therefore, solving the problem in the objective space has become a promising area. Benson and Lee (1996) consider an outcome-based algorithm for optimizing over the efficient set of a

bicriteria linear programming problem. Based on the initial work, Fülöp and Muu (2000) suggest a branch and bound variation (Algorithm 3) of Benson and Lee's algorithm. The revised algorithm, instead of constructing a sequence of consecutive nondominated edges in the outcome space, generates a refining sequence of partitions covering the whole or subsets of the nondominated set in the hope that a better lower bound will be detected during the search process. Thoai (2000a) is concerned with optimizing a nondecreasing quasiconvex function over the nondominated set. This method uses cutting planes to remove infeasible region in outcome space. Thoai (2000b) (Algorithm 4) uses the branch and bound technique to minimize some continuous function over the efficient set of a multiple objective programming problem. A bipartition technique is used in the branching process. Lower and upper bounds of each part are evaluated. Nguyen Thi, Thi, and Tran (2008) propose a polyblock outer approximation method (Algorithm 5) maximizing a continuous and increasing function over the nondominated set. It starts with a block containing the feasible region in outcome space. Subsets of the block are removed and the function values at vertices are evaluated iteratively. Benson (2011) proposes another branch and bound method (Algorithm 6) to globally minimize a finite, convex function over the weakly efficient set of a multiple objective nonlinear programming problem.

In this paper, we consider a minimization problem which incorporates one linear function as the objective function. The feasible set is the nondominated set in the objective space of a multiobjective problem. This algorithm is based on an outer approximation algorithm first introduced by Benson (1998) and further improved by Ehrgott, Löhne, and Shao (2011), which has been successfully used to find all the nondominated extreme points of a multiobjective linear programming problem. It works in the outcome or objective space instead of the decision space. This article is organized as follows: In Section 2, the problem linear optimization over the nondominated set is introduced and the revised outer approximation algorithm is summarized. Section 3 details out algorithm. In Section 4, we use a simple numerical example to further illustrate the algorithm. Section 5 is dedicated to computational tests of the algorithm using some randomly generated problems. Performance of our algorithm and other existing ones is compared. In the last section, we use this algorithm to find the nadir point as an application of Algorithm A.

2. Linear programming over the nondominated set and the revised outer approximation algorithm

2.1 Linear programming over the nondominated set

A multiobjective linear programming problem (*MOLP*) is

$$\text{Min}\{Cx : x \in X\}, \quad (\text{MOLP})$$

where $X = \{x \in \mathbb{R}^n : Ax \leq b\}$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. $C \in \mathbb{R}^{p \times n}$ is a $p \times n$ matrix. The rows c^k , $k = 1, \dots, p$ are the coefficients of p linear functions $c^k x$, $k = 1, \dots, p$.

The feasible set Y in objective space \mathbb{R}^p is defined by

$$Y = \{Cx : x \in X\}.$$

It is well known that the image Y of a nonempty, compact polyhedron X under a linear map C is also a nonempty, compact polyhedron of dimension $\dim Y \leq p$. The proof of the theorem can be found in Rockafellar (1970).

We use the following notation, $y^1 = y^2$ if $y_i^1 = y_i^2$ for $i = 1, \dots, p$; $y^1 < y^2$ if $y_i^1 < y_i^2$ for $i = 1, \dots, p$; $y^1 \leq y^2$ if $y_i^1 \leq y_i^2$ for $i = 1, \dots, p$; $y^1 \leq y^2$ if $y_i^1 \leq y_i^2$ for $i = 1, \dots, p$, and $y^1 \neq y^2$.

Definition 1 A feasible solution $\hat{x} \in X$ is an efficient solution of (MOLP) if there does not exist $x \in X$ such that $Cx \leq C\hat{x}$. The set of all efficient solutions of (MOLP) will be denoted by X_E and called the efficient set in decision space. Correspondingly, $\hat{y} = C\hat{x}$ is called a nondominated point and $Y_N = \{Cx : x \in X_E\}$ is the nondominated set in objective space of (MOLP).

Definition 2 A feasible solution $\hat{x} \in X$ is a weakly efficient solution of (MOLP) if there does not exist $x \in X$ such that $Cx < C\hat{x}$. The set of all weakly efficient solutions of (MOLP) will be denoted by X_{WE} and called the weakly efficient set in decision space. Correspondingly, $\hat{y} = C\hat{x}$ is called a weakly nondominated point and $Y_{WN} = \{Cx : x \in X_{WE}\}$ is the weakly nondominated set in objective space of (MOLP).

Optimizing over the efficient set of a multiobjective linear programming problem is written as

$$\min \varphi(x), \text{ s.t. } x \in X_E, \quad (\text{P1})$$

where X_E is the efficient set of problem MOLP.

In this paper, we focus on a special case of (P1), where function φ can be written as $\varphi(x) = f(Cx)$ and $y = Cx$, then P1 can be reformulated as

$$\min f(y), \text{ s.t. } y \in Y_N, \quad (\text{P})$$

where $Y_N := \{y \in \mathbb{R}^p \mid y = Cx, x \in X_E\}$. In the case of linear optimization, we let $\varphi(y) = u^T y$, where u is a column vector. It is obvious that if y^* is a solution to problem (P), then any $x^* \in X$ such that $Cx^* = y^*$ is an optimal solution to problem (P1). The solutions of this linear system are infinitely many in most cases. Decision makers can always solve another linear programming with preferred a preferred objective such as

$$\min e^T x, \text{ s.t. } Cx = y^*$$

to choose one solution.

However problem (P) is not a convex problem. Even though the problem has been simplified with linearity, the nondominated set Y_N is nonconvex in all but trivial cases. For that reason, there might be local optima making (P) a hard problem.

THEOREM 1

The optimal solution of problem (P) occurs at extreme points if the function $f(y)$ is linear.

Proof: Assume y is an optimal solution of problem (P) and $y \notin Y_{Ex}$, where Y_{Ex} is the set of extreme points of Y . Then y can be expressed as a convex combination of $\hat{y}_i \in Y_{Ex}$, i.e.

$$y = \sum_{\hat{y}_i \in Y_{Ex}} \alpha_i \hat{y}_i, \text{ where } \alpha_i \in (0, 1), \sum_i \alpha_i = 1. \text{ Then}$$

$$u^T y = u^T \sum_i \alpha_i \hat{y}_i = \sum_i \alpha_i u^T \hat{y}_i \geq \sum_i \alpha_i u^T \tilde{y} = u^T \tilde{y},$$

where $\tilde{y} = \arg \min \{u^T \hat{y}_i \mid \hat{y}_i \in Y_{Ex}\}$. If $u^T \hat{y}_i = K$, for every i and some K , the equality sign is obtained, otherwise we have found a feasible solution \tilde{y} which makes the function values no

worse than that at y , which contradicts the assumption that y is the optimal solution of problem (P).

2.2 The revised outer approximation algorithm

For problem (MOLP), Benson's "outer approximation algorithm" (1b) starts with a p dimensional simplex containing the polyhedron Y in the outcome space. And then through generating supporting hyperplanes to Y iteratively, it shapes the initial simplex to approximate the nondominated set of the polyhedron from outside. Finally, all of the extreme points of the nondominated set are found. More details can be found in Benson (1998). In Ehrgott, Löhne and Shao (2011), a revised version of Algorithm 1b is constructed. It starts with a set bounded from below instead of a bounded simplex. The revised version is showed below.

Revised Algorithm (1b)

Initialization:

- (i1). Compute ideal point y^l , where $y_i^l = \min\{c^i x \mid x \in X\}$, $i = 1 \dots p$. If $y^l \in Y_N$, the algorithm terminates with y^l the only nondominated extreme point, otherwise go to i2.
- (i2). Choose a point $\bar{p} \in \text{int } Y + \mathbb{R}_{\geq}^p$
- (i3). Set $S^0 := \{y^l\} + \mathbb{R}_{\geq}^p$. Store both the vertex set $V(S_0) = \{y^l\}$ and the inequality representation of S_0 .
- (i4). Initialize the candidate set $\text{CandiSet} = \{y^l\}$ and $Y_{NE} = \emptyset$, where Y_{NE} is the set of nondominated extreme points of Y . Set $k = 1$ and go to iteration k .

Iteration k :

- Step k1.** If for each $v \in V(S_k)$, $v \in Y$ is satisfied, then go to Step k5, Otherwise, choose any $v^k \in V(S_k)$ such that $v^k \notin Y$ and continue.
- Step k2.** Find the unique value λ_k of λ , $0 < \lambda < 1$, such that $\lambda_k v^k + (1 - \lambda_k) \bar{p}$ belongs to the boundary of Y , and set $q_k = \lambda_k v^k + (1 - \lambda_k) \bar{p}$.
- Step k3.** Set $S_{k+1} = S_k \cap \{y \in \mathbb{R}^p \mid w_k^T y \geq b_k^T u_k\}$, where (u_k^T, w_k^T) can be found by solving the LP

$$\max\{b^T u - v^{kT} w \mid A^T u - C^T w = 0, e^T w = 1, u, w \geq 0\}.$$
- Step k4.** Using $V(S_k)$ and the definition of S_{k+1} given in Step k3, determine $V(S_{k+1})$. Set $k = k + 1$ and go to iteration k .
- Step k5.** Let the total number of iterations be $K = k$. $V(S_k) = Y_{NE}$.

THEOREM 2

The modified version of Benson's "outer approximation algorithm" is finite.

Proof : See Ehrgott, Löhne and Shao (2011) Theorem 4.6.

3. The new algorithm to solve problem (P)

3.1 Algorithm 2

According to Theorem 1, we know that the optimal solution of problem P always occurs on extreme points, and Algorithm 1b enumerates all the extreme points, therefore, it provides a way to obtain optimality of problem (P) through evaluating the function value at each extreme point. This methodology involves two phases: in the first phase, we solve a MOLP problem by using Algorithm 1b. After the first phase, the set of nondominated extreme points is found. In the second phase, the value of the function f at each of extreme point is evaluated; the smallest one is selected as the optimal value of (P).

In order to solve problem (P), one additional step following algorithm 1b is added.

Step k6. In the set Y_{NE} , choose the point with the best function value.

This point is the optimal solution of problem (P).

We name this brute force algorithm with the additional step k6 Algorithm 1. However, if the two phases are combined together in a way such that the generation and evaluation of extreme points are done parallelly and iteratively, a large amount of computation efforts might be saved. More specifically, in the process of algorithm B1, once a new hyperplane is generated and added as a cut, a set of new extreme points is found. And then we evaluate the function values at these points. The one with the best function value is then selected to construct the cut for the next iteration. If the selected point $\hat{y} \in Y + \mathbb{R}_{\geq}^p$, another type of cut named threshold cut is added; otherwise a normal cut in Algorithm 1b is added. Here we call it an improvement cut.

Definition 3

An improvement cut is $\{y \in \mathbb{R}^p \mid w_k^T y \geq b_k^T u_k\}$ as described in Algorithm B1.

A threshold cut is $\{y \in \mathbb{R}^p \mid u^T y \leq u^T \hat{y}\}$, where \hat{y} is the incumbent solution, i.e. the best feasible nondominated point found so far.

The threshold cut removes the region where the function value is no better than the incumbent solution. Obviously, all points on the hyperplane have the same function value. Therefore, a threshold cut is not to be followed by another one of this type. In order to take control of this process, we introduce a threshold cut indicator, which is switched on if a cut of this type is added, and switched off when an improvement cut is added.

Details of the algorithm are as follows:

Algorithm 2

Initialization:

(i1). Compute ideal point y^l , where $y_i^l = \min\{c^i x \mid x \in X\}$, $i = 1 \dots p$. If $y^l \in Y_N$, the algorithm terminates with y^l the optimal solution and $f(y^l)$ the best objective function value. Otherwise go to (i2).

(i2). Choose a point $\bar{p} \in \text{int } Y + \mathbb{R}_{\geq}^p$

(i3). Set $S^0 := \{y^l\} + \mathbb{R}_{\geq}^p$. Store both the vertex set $V(S_0)$ and the inequality representation of S_0 .

(i4). Initialize the candidate set $\text{CandiSet} = \{y^l\}$. Set $k = 1$, threshold cut indicator $T = \text{False}$, and go to iteration k .

While $\text{CandiSet} \neq \emptyset$ do

Iteration k :

Step k1. Choose a point $\hat{v}^k = \arg \min\{u^T v^k \mid v^k \in V(S_k)\}$ if $\hat{y}_k \in Y + \mathbb{R}_{\geq}^p$ then go to Step k4. Otherwise, go to Step k2.

Step k2. Find the unique value λ_k of λ , $0 < \lambda < 1$, such that $\lambda_k \hat{v}^k + (1 - \lambda_k) \bar{p}$ belongs to the boundary of Y , and set $q_k = \lambda_k \hat{v}^k + (1 - \lambda_k) \bar{p}$.

Step k3. Set $S_{k+1} = S_k \cap \{y \in \mathbb{R}^p \mid w_k^T y \geq b_k^T u_k\}$, where (u_k^T, w_k^T) can be found by solving the linear programme

$$\max\{b^T u - \hat{v}^{kT} w \mid A^T u - C^T w = 0, e^T w = 1, u, w \geq 0\}. \quad T = \text{False}.$$

Step k4. If $T = \text{False}$, set $S_{k+1} = S_k \cap \{y \in \mathbb{R}^p \mid u^T y \leq u^T \hat{y}_k\}$ $T = \text{True}$.

Step k5. Using $V(S_k)$ and the definition of S_{k+1} given in Step k3, determine $V(S_{k+1})$. Set $k = k + 1$ and go to iteration k .

To calculate the ideal point, one needs to solve p single objective linear programming problems whose objective function is each row of matrix C . If the ideal point turns out to be feasible, it dominates all other points above, i.e. it is a feasible and therefore optimal solution to problem (P). To find the unique value of λ_k , solve the linear programme

$$\lambda_k = \max\{\lambda : x \in X, \lambda \hat{v}_k + (1 - \lambda) \bar{p} \geq Cx, \lambda \in (0, 1)\}.$$

Generating the improvement cut is the same as the cut generated in Algorithm 1b. the vertex enumeration algorithm Chen, Hansen, and Jaumard (1991) is used to find all vertices generated by intersecting the new hyperplane and S_k . This algorithm does not generate the whole set of the nondominated extreme points, therefore, Y_{NE} is not one of the results of this algorithm. It is mainly because in order to find the optimal extreme points, it is unnecessary to generate all of the extreme points.

THEOREM 3

Algorithm 2 is finite.

Proof: According to THEOREM 3, Algorithm 1b is finite. And it is obvious that the number of iterations of Algorithm 2 is no more than that of Algorithm 1b. furthermore, cuts generated are either supporting hyperplanes to Y or cuts on which $f(y)$ takes the same value. Hence, the optimal solution is not cut off during this process. Therefore, algorithm 2 is finite.

4. A numerical example

This section illustrates Algorithm A with a simple numerical example.

$$\begin{aligned} \min \quad & u^T y \\ \text{s.t.} \quad & y \in Y_N \\ Y_N = \min \{ & Cx \mid Ax \leq b, x \leq 0 \}, \end{aligned}$$

$$\text{where } A = \begin{pmatrix} -4 & -1 \\ -3 & -2 \\ -1 & -5 \\ 1 & 1 \end{pmatrix} \quad b = \begin{pmatrix} -4 \\ -6 \\ -5 \\ 6 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad u = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

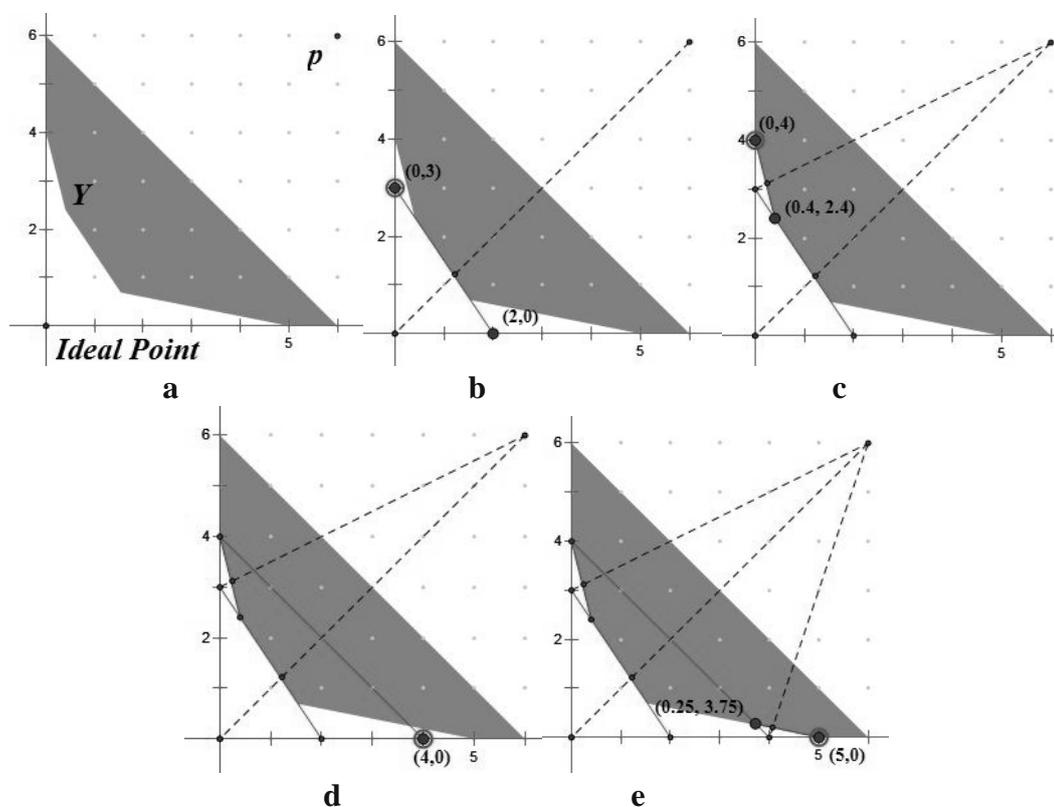


Figure 1.

Fig.1 illustrates how our algorithm 2 solves the numerical example. Plot a shows the feasible region Y , the ideal point and an interior point $p \in \text{int } Y + \mathbb{R}_{\geq}^2$. Plot b demonstrate the first iteration, in which the first improvement cut is constructed, and $(0,3)$ is selected. Plot c shows that in the next iteration, another improvement cut is added, and the minimum value of $f(y)$ found so far is -4 . On plot d, a threshold cut is added and point $(4,0)$ is found. Finally, the optimal solution $(5,0)$ is obtained after another improvement cut is generated.

Table.1

	Iteration	Point chosen	Cut type	Candidate	Y_{NE}	$f(y)$
b	1	(0,0)	Improvement	(2,0) (0,3)	\emptyset	-3
c	2	(0,3)	Improvement	(2,0)	(0,4) (0.4,2.4)	-4
d	3	(0,4)	Threshold	(4,0)	(0,4)	-4
e	4	(4,0)	Improvement	\emptyset	(0,4) (5,0) (0.25,3.75)	-5

5. Computational experiments

In order to generate random test problems quickly, the method by Charnes, Raike, Stutz, and Walters (1974) is used. In order to compare different algorithms, problems generated are all linear. All of the algorithms were implemented in Matlab R2010b, version 7.11.0.584, by using Cplex solver, on an Intel(R) Core(TM) i3 CPU computer.

Table 2 below shows the average runtimes (in second) of Algorithm A and B of solving twenty problems of the same size.

Table.2

m	n	p	1	2	3	4	5	6
2	2	2	0.0166	0.0147	0.0102	0.0104	0.9156	0.0211
5	5	2	0.0208	0.0158	0.0243	0.0156	1.3932	0.0973
10	18	2	0.0400	0.0252	0.0727	0.0407	0.6787	0.1067
50	90	2	0.1034	0.0765	0.0476	0.6062	0.4300	0.6703
200	200	2	3.7465	2.5074	0.1941	1.7474	2.3445	3.6720
10	18	3	0.1838	0.0765	-	0.1747	6.5392	0.2557
50	90	3	8.7173	0.4015	-	0.1572	15.334	1.5828
100	198	3	25.136	1.6284	-	1.5073	31.215	3.4237
5	10	4	1.6977	0.2002	-	0.0293	12.857	0.8088
10	18	4	9.4583	0.2218	-	0.0365	17.054	1.5352
10	15	6	30.466	4.7573	-	0.0372	36.321	4.9149
10	20	8	121.91	16.131	-	0.0608	59.249	29.127

Compared to Algorithm 1, the superiority of Algorithm 2 is obvious especially when solving relatively large problems. Since the number of faces of the polyhedron in objective space increases with the increase of the number of objective functions, it is expensive to generate all of them as Algorithm 1 does. The merit of partially generating the faces is demonstrated through Algorithm 2. Algorithm 3 is only capable of solving (P) when the

underlining (MOLP) has two objective functions. However, it is the fastest one for solving problems of this type. Algorithm 4 is the branch and bound Algorithm designed by Thoai (2000b), which turns out to be the fastest method to solve the all-linear problems. As it is shown in Table 1, time spent does not increase much as the size of the problem increase. It is mainly because that the relaxation problems solved during the each iteration provides promising lower bound. Especially for “all-linear” cases, it is quite likely that the upper bound and lower bound coincide in the initial steps. Algorithm 5 is only capable of solving problems with nondecreasing functions $f(y)$. However, when the problems are linear, it is trivial to use this method because in this case, problem (P) is the same as the following problem.

$$\min f(y), s.t. y \in Y,$$

where the feasible region is the whole polyhedron. This method compared to others seems not to be a fast one. However, the value of this algorithm might be revealed in solving problems with nonconvex function $f(y)$. Algorithm 6 also employs the branch and bound scheme. However, it is designed to solve more general problems with nonlinear constraints and objective function. It does not take the advantage of linearity, which is probably why it does not exceed our algorithm.

6. Nadir point

Ehrgott and Tenfelde-Podehl (2003) review some existing methods and propose a general method to compute nadir values and investigate how to use nadir points for compromise programming. Fig. 2 shows the difference between nadir Point and anti-ideal Point.

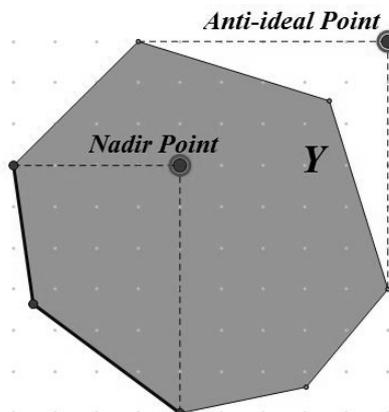


Fig.2

Algorithm A can also be used to find the nadir point by solving a series of problem (P) with maximizing along each coordinate direction.

$$\max\{e^i y \mid y \in Y_N\},$$

where $e^i \in \mathbb{R}^p$ is a unit vector with the i-th element being 1 and all the others to be zero.

7. Conclusion

In this study, an outer approximation algorithm is revised and implemented to solve a linear optimization problem which is subject to the nondominated set of a multiobjective linear problem. The reason that we choose to solve this problem in the objective space is because the image of the feasible region in the objective space is simpler than that in the decision space. Two types of cuts are used in our algorithm. After the algorithms are introduced, the

experimental results reveal the advantage of this new algorithm. Further study can be made to combine the branch and bound scheme with the cutting plane technique in the hope of taking advantage of both methods. And finally, this algorithm is shown how to find nadir points.

References

- Benson, H. P. (1998). An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem. *Journal of Global Optimization*, 13(1), 1-24.
- Benson, H. P. (2011). An outcome space algorithm for optimization over the weakly efficient set of a multiple objective nonlinear programming problem. *Journal of Global Optimization*, 52(3), 553-574. doi:10.1007/s10898-011-9786-y
- Benson, H. P., & Lee, D. (1996). Outcome-based algorithm for optimizing over the efficient set of a bicriteria linear programming problem. *Journal of Optimization Theory and Applications*, 88(1), 77-105. doi:10.1007/bf02192023
- Charnes, A., Raike, W. M., Stutz, J. D., & Walters, A. S. (1974). On generation of test problems for linear programming codes. *Journal of the ACM*, 17(10), 583-586. doi:10.1145/355620.361173
- Chen, P.-C., Hansen, P., & Jaumard, B. (1991). On-line and off-line vertex enumeration by adjacency lists. *Operations Research Letters*, 10(7), 403-409. doi:10.1016/0167-6377(91)90042-n
- Ehrgott, M., Löhne, A., & Shao, L. (2011). A dual variant of Benson's "outer approximation algorithm" for multiple objective linear programming. *Journal of Global Optimization*, 52(4), 757-778. doi:10.1007/s10898-011-9709-y
- Ehrgott, M., & Tenfelde-Podehl, D. (2003). Computation of ideal and Nadir values and implications for their use in MCDM methods. *European Journal of Operational Research*, 151(1), 119-139. doi:10.1016/s0377-2217(02)00595-7
- Fülöp, J., & Muu, L. D. (2000). Branch-and-Bound Variant of an Outcome-Based Algorithm for Optimizing over the Efficient Set of a Bicriteria Linear Programming Problem. *Journal of Optimization Theory and Applications*, 105(1), 37-54. doi:10.1023/a:1004657827134
- Nguyen Thi, B. K., Thi, H. A., & Tran, M. T. (2008). Outcome-Space Polyblock Approximation Algorithm for Optimizing over Efficient Sets. *Modelling, Computation and Optimization in Information Systems and Management Sciences*, 14, 234-243. doi:10.1007/978-3-540-87477-5_26
- Philip, J. (1972). Algorithms for the vector maximization problem. *Mathematical Programming*, 2(1), 207-229. doi:10.1007/bf01584543
- Rockafellar, R. T. (1970). *Convex analysis*. Princeton, N.J.: Princeton University Press.
- Thoai, N. V. (2000a). A class of optimization problems over the efficient set of a multiple criteria nonlinear programming problem. *European Journal of Operational Research*, 122(1), 58-68. doi:10.1016/s0377-2217(99)00068-5
- Thoai, N. V. (2000b). Conical Algorithm in Global Optimization for Optimizing over Efficient Sets. *Journal of Global Optimization*, 18(4), 321-336. doi:10.1023/a:1026544116333
- Yoshitsugu, Y. (2002). Optimization over the efficient set: overview. *Journal of Global Optimization*, 22(1-4), 285-317. doi:10.1023/a:1013875600711

The Art and Science of Matchmaking (as it Relates to Badminton)

Craig MacLeod

Orbit Systems

Wellington, New Zealand

craig.macleod@orbitssystem.co.nz

Abstract

The Wellington Badminton Club hosts around 60 players on club nights and creates a series of doubles matches for participants. Creating a series of games that are fun and challenging for all participants is quite a difficult task, and can be quite stressful for the people involved.

The manual process for creating the games was difficult, time consuming, and prone to errors. It seemed logical that a computer program be created to perform this task.

This paper describes the program created to design these matches, the types of problems that were solved, and those that still remain.

Key words: OR, optimisation, heuristic

Introduction

The Wellington Badminton Club hosts club nights each week for members and visitors alike. On a typical week, approximately 60 or more players will be present, and the club arranges what it hopes will be an enjoyable series of doubles games for everyone.

Historically, each player was assigned a grade from 1 (good) to 5 (would like to be good), and each person's name was written on a coloured tag that indicated both gender and skill level. On arrival at the hall, each person placed their tag on a nail board to indicate their presence. A club member would be assigned to design matches by placing the tags onto a "map" of the courts. At the start of each round, players could find their tag on the map to see where, and with who, they were playing.

Creating games for each round required the person making up the games to be reasonably familiar with all members. It was close to being a full-time job. Depending on the skill of the person involved, games could be generally quite good, or sometimes quite poor. Mistakes were reasonably frequent – for instance, swaps could alter the playing sequence and result in players missing multiple games.

Generally speaking, the task was restricted to a few senior members of the club, and these people typically could not actually play at all if they were assigned to making rosters.

A member of the club who wrote accounting software decided to automate the system. This program essentially mirrored the existing manual system, and required manual swapping of players to create an acceptable roster. Sadly, he was killed in an accident at home before the program could be improved beyond a semi-manual tool.

At this point, Orbit Systems became involved. With so many factors involved in creating a solution to the problem, and with a program already somewhat in place, it seemed logical to improve the basic heuristic already in place by stages. In theory, the users of the program should notice nothing different as the heuristic improves, other than the playing roster displayed should be a better one that requires fewer manual swaps to make it acceptable to the user.

What is the Problem that Needs Solving?

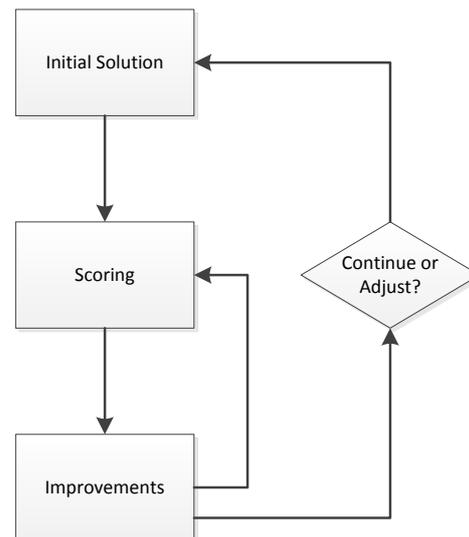
We identified a number of significant issues that needed to be resolved before a solution could be created. In essence, there needs to be some agreement about what the “objective function” is – otherwise, how would we know if the solution we found is any good?

- Who should be playing and who should be watching? This is an extremely sensitive issue for members – the method used to choose who should play needed to be extremely fair. This is complicated by the fact that players arrive and leave according to their own schedules.
- What actually makes a good game? Should all the players be the same ability? How often should there be mixed games? There is often very little consensus about what a good game is, but everyone has their own view about the games they participate in.
- Player rankings. For a computer to create good games for people, it needs to know how skilled they are. There is no generally accepted standard for this, and rankings are more difficult when doubles are concerned. There is also the fact that many people are not as good as they like to think they are. A game that such a person would like would annoy the 3 others on court as their perception is that they are playing with a player of lesser standard.
- Does familiarity really breed contempt? In other words, would a good solution early in the evening still be considered good if was used over and over again? The answer is likely not, as part of what makes a club night successful is the ability to play against different people with different styles of play. However, many players – especially the better ones – will draw the line at variety if the quality of the player falls too much.
- What about gender equality? Do we need to have even numbers of men and women, or can everyone just get along?

A Generic Solution Technique

If a problem is likely to be too difficult to optimise directly, some sort of solution heuristic will be required. The basic template for such a solution is as follows.

1. Find an initial solution of some sort. It need not be feasible, but feasible is better if possible. Different types of starting solutions are a very good idea also. Better starting solutions normally lead to better finishing solutions.
2. Design a scoring algorithm for a solution. This is probably the most important part of the solver. In theory at least, you could try random solutions and pick the one with the best score. If the scoring system scales well and reflects your view about what makes a good solution, you likely will find something acceptable in a reasonably short period of time.
3. Attempt to improve the initial solution, referring back to the scoring algorithm for confirmation. There can be swaps, re-arrangements and the like. When no more incremental improvements can be found, this part of the cycle terminates.
4. Decide if you've finished or not. If there are no other initial solution techniques to try, the solution is likely optimal, or time is up, you can report the current best solution.



Initial Solution: The Playing List

The question of who should be playing and who should be watching is quite a complicated one. In theory, we should be able to pick players at random from the available pool. This is fair in theory, and has the advantage that players have the best possible chance of playing against all other people present. In practice, however, members dislike random selection quite strongly.

The final rules for player selection have more to do with the perception of fairness than the reality of it, but this proved to be a necessary step for general happiness.

- Anyone waiting to play should get to play before anyone currently playing.
- If you must miss two games in a row, then everyone else should miss two games in a row before you miss two in a row again.
- The arrival sequence at the hall should determine the initial set of people playing.
- Once all of these rules have been implemented, a small random element is acceptable to break ties. The weighting should favour people who have missed a greater number of games.

In practice, we allow people to be “forced” into the playing roster manually. There are various reasons for that, but the person doing the scheduling does this with some risk of major unhappiness.

When there is flexibility in terms of who will be in the playing list (the final player in the list requires a tie-breaker), it makes sense that we could improve games by including “compatible” groups of players. While this makes sense in theory, it is very difficult to implement in practice. For example:

- Forcing an even number of men and women. Each “good” game (see later) features four of one gender or two of each, so it makes sense that having even numbers of men and women would result in better games being generated. This is not the case, however. Imagine that only two men are picked; one is very good, while the other is a beginner. There are no “good” games that will have both these men in them. It turns out that it’s better to pair people by skill levels where possible than to pair people by gender.
- Sometimes people are “incompatible” because their skill level is different to most of the others playing at that time. In this case, the program will consistently exclude them from the roster in any tie-breaking situation. This was felt to be too unfair to be implemented.

Scoring A: What Makes a Good Game?

There are some types of game that most people agree will be good ones – such as when all four players are quite close to the same skill level. We can’t just schedule the obvious matches each time, however.

General guidelines about what might make a good game include:

- All players are close to the same skill level.
- An even number of men and women is normally better.
- Two pairs of matched players can be quite good, but the skill-level gap can’t be too great.
- Some level of variety is good; the same four players are OK for a few games, but after that the enjoyment will decrease.

There are exceptions to every rule, of course:

- Some people simply don’t like each other. Regardless of skill level, pairing certain people is not a good idea and will never result in a good game.
- Beginners actually have better games when paired with better players. The rallies will be longer and more enjoyable as the better players can feed the shuttle in a way that will be returnable for the beginner. Of course, the better players don’t particularly like those games, but may volunteer if offered a better game later.
- We can’t simply use absolute skill level when assessing mixed doubles games. At the Wellington Badminton Club, the top men are a lot stronger players than the top women. However, the top men paired with the top women in a mixed doubles game is still considered a good game. Players must be compared on an absolute basis unless playing mixed doubles, when relative strength may also be used – but only when the men are better than the women in an absolute sense.

Ultimately, the roster program awards a score to all games. Points are awarded for the matching skill levels and certain types of pairings that people agree are good games. Points are deducted based on the number of times people have played together already, for large skill differences, and sometimes for personality clashes.

The score for any given roster is the sum of the scores for each match.

Scoring B: Assessing Player Skill Levels

All the scoring concepts mentioned to this point rely on having an accurate assessment of a player's skill level. Assessing each player's skill level can be quite difficult, but since it is so important, a separate program was created for just this task.

It can be difficult to assess skill in absolute terms, but for our purposes, a relative ranking is sufficient.

The program works as follows:

- Any number of ranked player lists can be created, and the list may cover just a subset of players if desired. Many people know only those members they play regularly, but can normally rank at least some people in order.
- Each ranked list is given a weighting – some people are deemed better at making these ranking lists than others.
- The program performs what amounts to being a bubble sort. For the comparison between any two players, all ranking lists containing those two players is considered. The weighted “is better than” score determines the swap value for the bubble sort.
- The final ranking list can be converted into a skill level for the program.

Given that there are people involved, these rankings are not without some controversy. Some players feel that they are better than they really are (better than everyone else thinks they are, anyway), and are not satisfied unless they play with better players. Of course, the better players will not be happy with such a match-up. There is no easy way around this issue, and several players have left the club because their ranking means they will generally have “bad” games.

Creating the Roster

Creating the roster of games for each round utilises the results of the previous three sections: the playing list; the skill level of the players; and a scoring system for any given game.

An outline of the program is as follows:

- Assign players to courts in a number of different ways. Players can be assigned in skill sequence, randomly, or with a greedy algorithm that attempts to maximise partially-assigned game scores as each player is added to the roster.

- Once players are assigned to an initial court, look for individual swaps that will improve the overall score for the roster. The program normally starts looking for swaps on the court that has the worst score.
- When no more swaps can be found, the roster is compared to the current best one, with a replacement being made after any improvement.
- Each match is assigned to a random physical court to ensure that skill level does not affect location in the hall.

Improvements could be made in a number of areas, but key areas include:

- The program currently only attempts single swaps. Multi-player swaps would likely lead to better solutions.
- The initial playing list uses a small random factor to break ties. The swapping algorithm could potentially swap with players who only missed playing because of the tie-breaker. This would result in better games when mismatched players end up in the playing list.

Manual Adjustments

No matter how good the initial roster is, there will be times when changes must be made. There a number of reasons for this, including:

- The operator sees a much better combination of players that the swapping algorithm didn't see, or didn't make for some reason.
- A particular game is desired for personal reasons, and the computer would never come up with that game unaided.
- An earlier mistake requires "payback" in some way.

The program allows for changes like this in a few different ways:

- Firstly, an entire court or courts that are physically available can be left off the roster. This allows complete discretion for the operator to make up some games. This option is used only rarely.
- Next, "requests" can be set up. Two or more players can be added to a request list. The computer will find a suitable player for any missing slot. The request can be "forced" in the sense that this game will definitely be created in the next round.
- Finally, the program allows manual swaps to be undertaken. The swaps can be made between courts – a modestly safe kind of swap. The swaps can also take someone out of the playing roster and put someone new in – these swaps need to be carefully considered as longer non-playing waiting periods can be created.

Because People Are People

Because people are people, the roster program needs to have available quite a number of statistics:

- A complete list of each match and who was playing in what sequence. It is quite common for people to check the playing roster poorly, then complain that they have missed multiple games. It is also quite common for people to play with the wrong partner and have a poor game as a result.
- For each player, the program lists all their games, whether the computer thinks it was a “good” game or not, how many games have been missed and so on.
- Some players just don’t mix well even though mathematically it looks like they should. We allow these players to be separated when possible.

Conclusion

The main questions to be answered about such a program are:

- Does it work?; and
- Are the games better than they used to be?

The program certainly works in a physical sense – it does make rosters. Benefits include:

- A roster can now be made in a few minutes – even when some consideration and swapping is undertaken.
- There are now twenty or more people at the club capable of making the rosters rather than two or three.
- The club has better records about how many people are present at different times during each evening, and how many times each member plays.

As for whether the games are better than before, the answer is not as clear. The general answer is that it still depends on who the person making the rosters is and was. One person in particular was reputed to be excellent at creating games, while others were not so well regarded.

At the very least we can say:

- Games are much more consistently created now, and the quality of games is not highly related to the person on duty that night. That said, manual swapping still allows personal preference to come through.
- The ease of use of the system means that no-one has to give up playing for the evening when they are rostered to create the matches. Two members can alternate games quite easily.
- The computer system is very “fair” in terms of playing time, and the number of games that will be missed. This was very difficult to achieve with the manual system, and impossible to check when a disagreement occurred.

Intra-period Market Clearing for a Multi-Use Catchment via CDDP

Indra Mahakalanda, Shane Dye, E. Grant Read, John F. Raffensperger
Department of Management,
University of Canterbury,
New Zealand.

indra.mahakalanda@pg.canterbury.ac.nz

Abstract

In this paper, we show how to efficiently optimise consumptive and non-consumptive water use in any river catchment that can be represented as a tree radiating from a single reservoir. We describe a deterministic Constructive Dual Dynamic Programming (CDDP) algorithm which implicitly clears a market, across all nodes in the catchment, to construct a net demand curve for water released from the reservoir, in each period. Given these intra-period demand curves, a stochastic CDDP algorithm can construct the demand curve for water stored in the reservoir, thus clearing an inter-temporal market to optimise both consumptive and non-consumptive water use across time.

Key words: Reservoir optimization, water markets, CDDP

1 Introduction

A reservoir operator faces the difficult decision as to how much water to release to users in the current period, and how much to store or release later. Her decision depends on the amount of water available, on uncertain inflows into the catchment, and on present and future consumptive and non-consumptive demands. This paper addresses this decision problem, with a focus on market clearing. Read and Hindsberger (2010) survey applications of Constructive Dual Dynamic Programming (CDDP) for inter-temporal reservoir optimization, some of which have been in market contexts. (Scott & Read, 1996).

Dye, Read, Read, and Starkey (2012) describe experiments using CDDP to determine the value that could be added by using stochastic optimisation in an inter-temporal market clearing context (Starkey, Dye, Read, & Read, 2012). However, those previous studies do not explicitly address the structure of the river network, and interactions between consumptive and non-consumptive water users at various points within it, whether by a market or otherwise. The stochastic CDDP algorithm decomposes the *inter*-temporal optimisation into a sequence of single period trade-offs between the benefits of immediate release and the expected benefits of storage for later release. The benefit from immediate release is modelled by a “Demand Curve for Release” (DCR), in each period. This paper describes an efficient *intra*-period CDDP algorithm to determine the DCRs from a river network. The intra-period optimisations are deterministic, forming sub-problems to an overall stochastic model.

River networks have a number of features affecting the demand for reservoir release. A stream or inflow that flows into the catchment is known as a tributary flow. These tributary flows are often subject to temporal variations. The river flow is often diverted

to cater for ecological demands and consumptive uses; these diversions typically have only a remote chance to return to the catchment directly. Non-consumptive uses, on the other hand, gain benefit (or in the case of pumps, incur costs) based on water flow. The water itself is still available for further use downstream. Both contribute to defining a DCR.

The DCR defines the amount of water the reservoir manager should economically release, given the benefit from supplying water to consumptive and non-consumptive users, as a function of the marginal value (or price) of stored water. The DCR could be determined using benefit curves estimated by the reservoir manager, but in this application we envisage the benefits being defined by clearing a market within which each consumptive user bids for water to be delivered to their node, and each non-consumptive user bids for water flow on their arc. To clear the full inter-temporal market requires the determination of a DCR for every inflow state in every period over the planning horizon.

2 Intra-period multi nodal market model

To construct a DCR solely as a function of long term reservoir storage, we assume that the costs and benefits of consumptive and non-consumptive use at different locations within the river network are independent. We also assume that those costs and benefits are independent of supplies and demands met in previous periods, and can be described by functions of water use which are convex for costs and concave for benefits.

The resulting DCR forms input for the CDDP market clearing module. We let nb be the net demand curve for water at a particular node, representing the net quantity of water demanded at that node as a function of price. We construct nodal net demand curves for water by combining the consumptive demands, distributary requirements and tributary flows at each node with the nodal net demand curves from neighbouring nodes, in a way which respects all constraints.

2.1 Tree representation of a catchment

We assume that the river network has a tree structure. Each node will have a parent (“next”) node, being the neighbouring node on the unique path towards the reservoir. Figure 1 illustrates the various features of a catchment. The CDDP algorithm produces a DCR by aggregating the water demand curves of all nodes towards the reservoir. Trading, and consumptive use, can occur at any node, except node 0 which is the reservoir. We generalise the DCR concept by defining a Demand Curve for Water, whether “released” or not, at each node. Non-consumptive uses occur on arcs between nodes, and affect marginal benefits in the demand curve for water for the parent node.

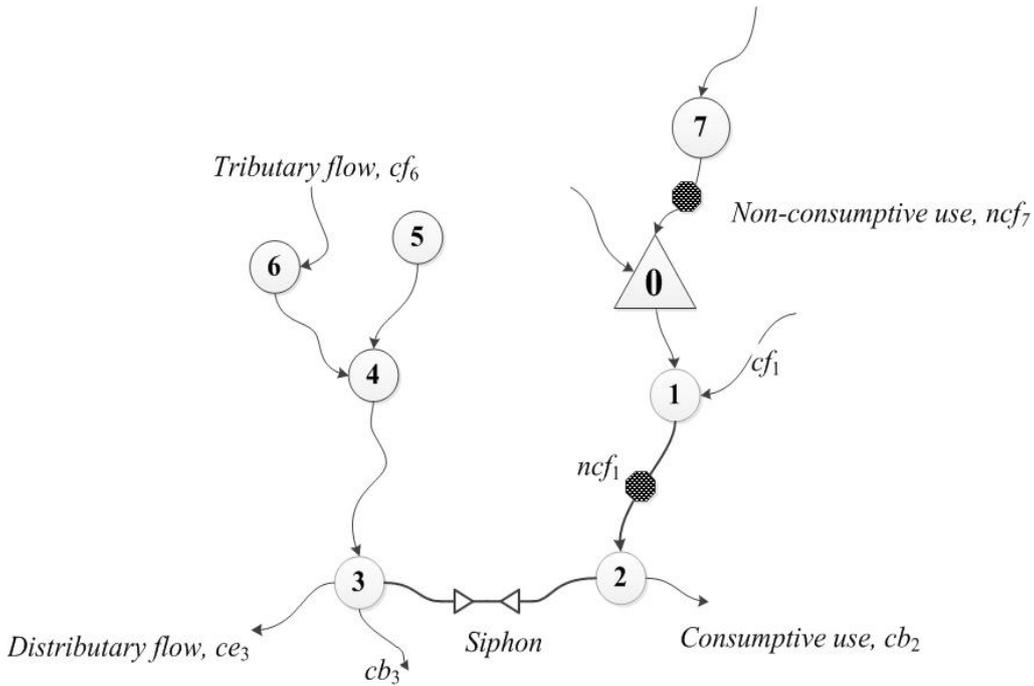


Figure 1: A single reservoir multi node catchment.

Our multi-nodal CDDP algorithm works inwards from the “leaves” of this tree, back to the reservoir (node 0), irrespective of flow direction. We first describe the model in terms of the demands, supplies and water flows as they would usually be modelled. We then describe the algorithm, which generates net demand curves. For this purpose nodal data is re-cast as net demands and arcs are reoriented towards the reservoir.

2.2 The catchment optimisation sub- problem

In this section we simply set up a mathematical description of the catchment optimisation sub-model which would need to be solved in order to determine the optimal benefit, $NB(r)$, from any particular level of reservoir release, r . A DCR could then be defined by differencing these $NB(r)$ values, but the CDDP algorithm that follows actually computes the DCR directly.

Let $T = (N=0..n,A)$ be a tree with the reservoir at the root node 0, and N nodes indexed so that child nodes (further from the reservoir) have a larger index than their parents. Arcs A are denoted j with $j \sim (i,k)$ indicating that the usual water flow direction for arc j is from node i to node k .

Each non-root node $i \in N$ has $F_i^{\max} \geq 0$ units of tributary inflow. (Flows directly into the reservoir are dealt with in the stochastic optimisation described later.) Variable f_i represents the amount of this inflow captured at node i , $0 \leq f_i \leq F_i^{\max}$. Inflow not captured is lost to the system. For full generality, $CF_i(f)$ denotes the increasing convex cost function associated with the capture of this flow. This generality allows for supply participants who incur costs or who charge for the flow injection, but for a typical tributary flow $CF_i(f)=0$. Each node i has $C_i \geq 0$ units of (potential) consumptive demand, with the associated decreasing convex benefit function $CB_i(c_i)$. Variable $0 \leq c_i \leq C_i$ represents the consumptive demand met at node i .

Variable $0 \leq d_i \leq E_i$ represents the distributary flow from node i . The minimum flow into node i is modelled as a target flow $E_i (\geq 0)$, with a convex increasing cost function $CE_i(E_i - d_i)$ associated with missing the target by $(E_i - d_i)$ units. $E_i = 0$ and $CE_i=0$ for nodes with no distributary flow, but distributary flows could be required to meet minimum environmental flow requirements.

Each arc $j \in A$ has an associated variable x_j representing downstream flow, with a lower bound of L_j and upper bound of U_j . The lower bound could be positive to meet a strict environmental flow requirement, or negative if pumping or some other upstream transfer were available. The variable $r \in \mathbb{R}$ represents the release from the reservoir. For a given release r , we solve the following model:

$$[1] \quad NB(r) = \max \sum_{i=1}^n \{CB_i(c_i) - CE_i(E_i - d_i) - CF_i(f_i)\} + \sum_{j \in A} NCB_j(x_j)$$

Subject to:

$$[2] \quad f_i + \sum_{j \sim (k,i)} x_j = c_i + d_i + \sum_{j \sim (i,k)} x_j \quad \forall i = 1, 2, \dots, n$$

$$[3] \quad \sum_{j \sim (k,0)} x_j - \sum_{j \sim (0,k)} x_j \leq r$$

$$[4] \quad -L_j \leq x_j \leq U_j; 0 \leq c_i \leq C_i; 0 \leq d_i \leq E_i, 0 \leq f_i \leq F_i^{max}$$

2.3 The multi-nodal CDDP algorithm

Rather than solve a whole series of sub-problems explicitly, the CDDP algorithm for clearing the intra-period nodal market implicitly defines (marginal) benefit as a function of release, across the whole range of possible releases in this period, by directly constructing marginal water value curves to compute the reservoir DCR. Note that this is equivalent to optimising model [1]-[4] for all values of r simultaneously, in a single pass. Arc bounds restrict the opportunities which may be exploited at connected nodes. The cumulative effect of these restrictions is maintained by a domain associated with each marginal value curve. We use the notation $f:S$ to denote function f restricted to domain S .

The model data is pre-processed to simplify the algorithm description. All cost and benefit functions at nodes are converted to marginal net demand functions. Here, these are described as defining marginal prices as a function of water supplied. But they could, equivalently, be expressed as defining net demand as a function of price. Specifically, for each node i , the pre-processing calculates the following net demand functions.

- Tributary flows form net demand $cf_i: [-F_i^{max}, 0]$ where $cf_i(f) = \frac{d}{dq} CF_i(-f)$.
- Consumptive demands form net demand $cb_i: [0, C_i]$ where $cb_i(c) = \frac{d}{dq} CB_i(c)$.
- Distributary flows form net demand $ce_i: [0, E_i]$ where $ce_i(d) = \frac{d}{dq} CE_i(E_i - d)$.

By convention, arcs are directed away from the reservoir. The flow bounds for arc j are adjusted to account for this, setting $L'_j = L_j$ and $U'_j = U_j$ if water flow is away from the reservoir, and $L'_j = -U_j$ and $U'_j = -L_j$ if water flow is directed towards the reservoir. Non-consumptive uses are similarly affected forming net demand function $ncf_j: [L'_j, U'_j]$ where $ncf_j(x) = \frac{d}{dx} NCB_i(x)$ for arcs with water flow away from the reservoir and $ncf_j(x) = -\frac{d}{dx} NCB_i(-x)$ otherwise.

The algorithm starts from the node with the largest index. We first form a nodal net demand curve ($nb:S$) by “horizontally adding” all tributary flows, distributary and consumptive demands at the node. This net demand curve, ($nb:S$), is transferred to the next node by first truncating using arc flow bounds; then “vertically adding” any non-consumptive flow demand components to form $(\tilde{nb}_i; \tilde{S}_i)$. This defines the net demand curve at the parent node, and the process is repeated until we reach the reservoir (node 0), through the tree.

Description of horizontal $[+^h]$ addition and vertical $[+^v]$ addition.

When the net demand functions are piecewise constant over a regular domain, we can efficiently implement both horizontal and vertical addition. Vertical addition amounts to adding corresponding function values, while horizontal addition amounts to sorting a collection of bid sections.

Vertical addition is point-wise addition with respect to the horizontal axis over the coincident *domain*. Formally:

$$h: [s, t] = f: [a, b] +^v g: [c, d] \text{ is defined by } s = \min(a, c), t = \max(b, d), \text{ and}$$

$$h(x) = \begin{cases} f(x) + g(x) & x \in [a, b] \cap [c, d] \\ f(x) & x \in [a, b] \setminus [c, d] \\ g(x) & x \in [c, d] \setminus [a, b] \end{cases}$$

Horizontal addition is point-wise addition with respect to the vertical axis over the coincident *range*. It corresponds to adding the horizontal axis quantities for each fixed value on the vertical axis. We define horizontal addition for decreasing functions only. Formally:

$$h: [s, t] = f: [a, b] +^h g: [c, d] \text{ is defined by } s = a + c, t = b + d, \text{ and}$$

$$h(x) = \begin{cases} f(x - c) & f(x - c) \geq g(c) \\ g(x - a) & g(x - a) \geq f(a) \\ \min\{\lambda | \exists y \in [a, b], f(y) \geq \lambda, g(x - y) \geq \lambda\} & \text{otherwise} \end{cases}$$

(We can alternatively write $h(x)$ as $h(x) = (f^{-1} + g^{-1})^{-1}(x)$ where this is well defined.)

To incorporate non-consumptive use, the algorithm adds the non-consumptive benefit (cost) to the incoming truncated $nb(\cdot)$ from the previous node. A non-consumptive “user” does not actually use water. She transfers water from one node to another, at some cost (e.g. for a pump) or to gain some benefit (e.g. for generation). The trade-off between the upstream/downstream difference in marginal benefit, and the cost or benefit of non-consumptive use, determines the net value from each flow. So we account for non-consumptive use by adding demand curves vertically, thus affecting the prices transferred to the parent node to each increment of flow between the two nodes.

It is straightforward to show that the following combined procedure constructs the DCR corresponding to $NB(r)$ from the previous model. We omit the proof for brevity.

Procedure Demand Curve for release (DCR) for single tree reservoir network

$nb_i: S_i$ is the nodal demand curve for flow , $nb_i = 0, S_i = \emptyset, \forall i = 0, \dots, n$

For $i = n$ to 1 Step – 1 do

$$nb_i: S_i \leftarrow [nb_i: S_i +^h cb_i: [0, C_i] +^h ce_i: [0, E_i] +^h cf_i: [-F_i^{min}, -F_i^{max}]]$$

set $j \sim (i, par(i))$ or $j \sim (par(i), i)$, whichever exists

$$\overline{nb}_i: \overline{S}_i = truncate [nb_i: S_i, [L'_j, U'_j]] \text{ // Truncated to arc limits}$$

$$\widetilde{nb}_i: \widetilde{S}_i = [\overline{nb}_i: \overline{S}_i +^v ncf_i: [L'_j, U'_j]]$$

$$nb_{par(i)}: S_{par(i)} \leftarrow [nb_{par(i)}: S_{par(i)} +^h \widetilde{nb}_i: \widetilde{S}_i] \text{ // } par(i) \text{ is the parent to node } i$$

Next i // Repeat process until reach node 0.

Output: DCR $\leftarrow nb_0: S_0$

3 Illustration of multi-nodal CDDP algorithm

We next illustrate the CDDP algorithm to construct the DCR in a single catchment, as shown in Figure 1. For simplicity, we assume that the catchment has no suppliers, but only consumptive users submitting demand bids, and non-consumptive users submitting bids for flow.

3.1 Constructing demand curves for the nodes above the reservoir

Consider a single node located upstream from the reservoir, as in Figure 2 (which shows detail from Figure 1). As shown in Figure 3, we first stack consumptive demands in descending price order, to form the consumptive demand curve for water, $cb_7(c)$. We also form a notional bid stack representing a relatively high marginal value for distributary flow demands to meet local environmental constraints. We represent tributary flow ‘supplies’ as negative quantities in the “net demand curve”. The algorithm then constructs the nodal demand curve for water $nb(\cdot)$ incrementally, by horizontal addition: $nb_7: S_7 \leftarrow [cb_7: [0,4]^h + ce_7: [0,2]^h + cf_7: [-4,0]^h]$. This is equivalent to sorting all increments of consumptive, distributary and (net) tributary demand by decreasing incremental value.

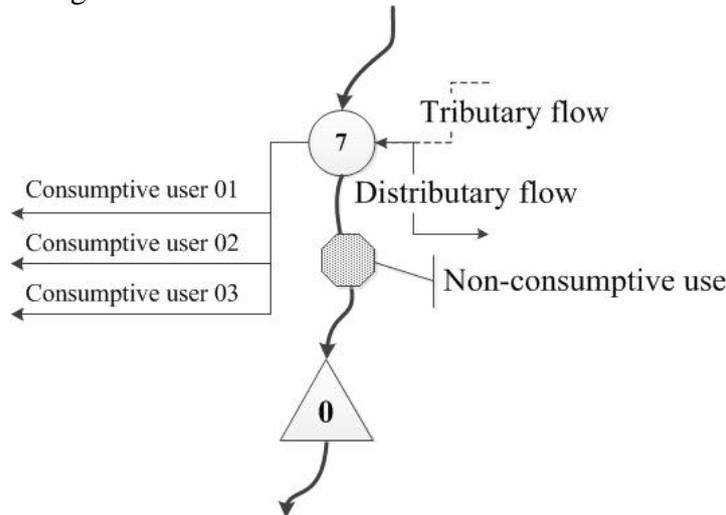


Figure 2: A node located above the reservoir

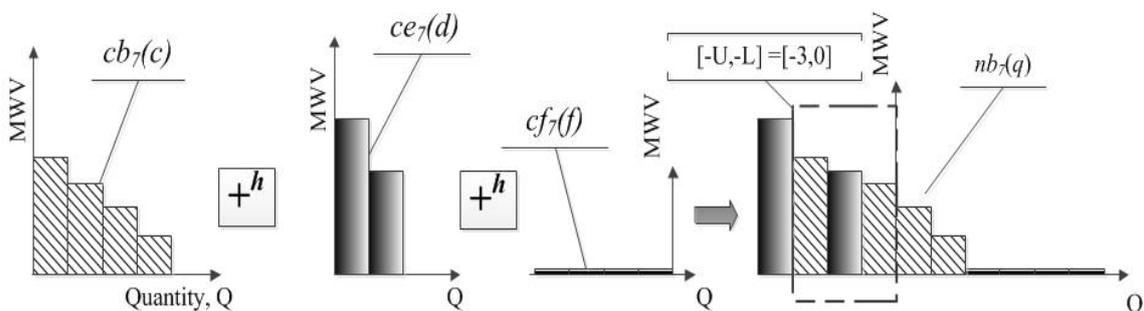


Figure 3: Horizontal addition ($+^h$) of distributive, consumptive and tributary flows to form nodal demand curve, $nb(q)$

Earlier, we defined flow away from the reservoir as positive, and flow towards reservoir (or reverse flow) as negative. Therefore, when we work towards the reservoir from upstream, the flow is negative. In line with the flow direction, we set the upper and lower limits for flow as $[-U_7, -L_7]$. The resulting net demand curve represents the willingness of upstream participants to pay in order to limit flows to the next

downstream node. The arc bounds truncate this nodal demand curve, as shown in Figure 3, reflecting limits on the ability of the downstream node to take advantage of the opportunity this represents. But here there is also a non-consumptive use to account for.

The benefit of non-consumptive flow is accounted for by the vertical addition $\widetilde{nb}_7: [-3,0] = [\overline{nb}_7: [-3,0] + {}^v ncf_7: [-3,0]]$ as in Figure 4. Note that, since flow from node 7 to the reservoir gains additional non-consumptive benefit, the net benefit gained by using water at node 7, rather than letting it flow through to the reservoir is reduced, and the prices in the net demand curve passed through to the reservoir are also reduced. This effect is achieved by the pre-processing, which will make $ncf_7 \leq 0$, implying vertical ‘subtraction’.

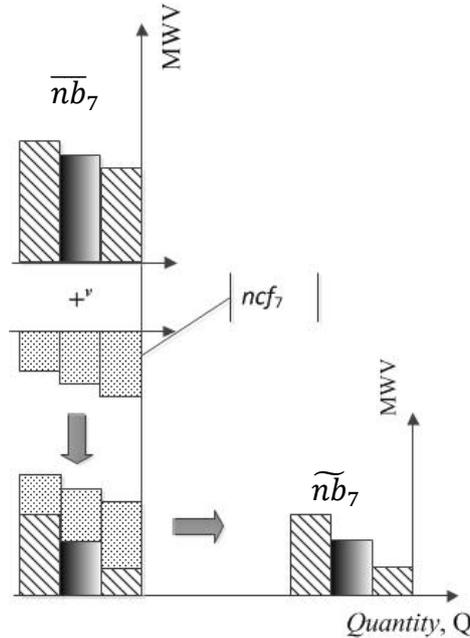


Figure 4: Vertical addition with non-consumptive flow demands (ncf)

3.2 Constructing demand curves for node 4

The next nodes processed are nodes 6 and 5. Both might be considered “downstream” from the reservoir. But the processing of nb_6 and nb_5 , the net demand curves for these nodes, follows the process for node 7, because the flow direction from both is “towards” the reservoir. In this case, arcs (5,4) and (6,4) have no non-consumptive uses, so the net demand curves constructed at nodes 6 and 5 can be transferred directly to node 4 after truncation for flow limits on the corresponding arcs. In this case, $(\widetilde{nb}_6: \widetilde{S}_6)$ and $(\widetilde{nb}_5: \widetilde{S}_5)$ represent the willingness of participants at nodes 6 and 5, respectively, to pay to limit flows to node 4. These are treated as further consumptive net demand curves at node 4, in addition to the local demands there, when determining nb_4 .

3.3 Truncating for siphon flow limits

Processing node 4 and forming $(\widetilde{nb}_4: \widetilde{S}_4)$ for input into the formation of nb_3 , follows an identical process. Figure 5 illustrates processing node 3 and, in particular, accounting for arc (3,2). This arc represents a syphon which allows flow in either direction. To account for the unknown flow direction, the processing truncates nb_3 by using a domain interval with a negative lower bound and a positive upper bound.

The resulting $(\widetilde{nb}_3: \widetilde{S}_3)$ represents the willingness of participants to provide flow from node 3 to node 2 for quantities in the negative part of the interval, and the desire of participants for flow from node 2 to node 3 for quantities in the positive part of the

interval. To put this another way, if the price of water were low enough, then node 3 would like to ‘buy’ water from node 2 (at prices to the right of the vertical axis). However, if the price were high enough, node 3 would be willing to ‘sell’ water to node 2 (at the prices to the left of the vertical axis).

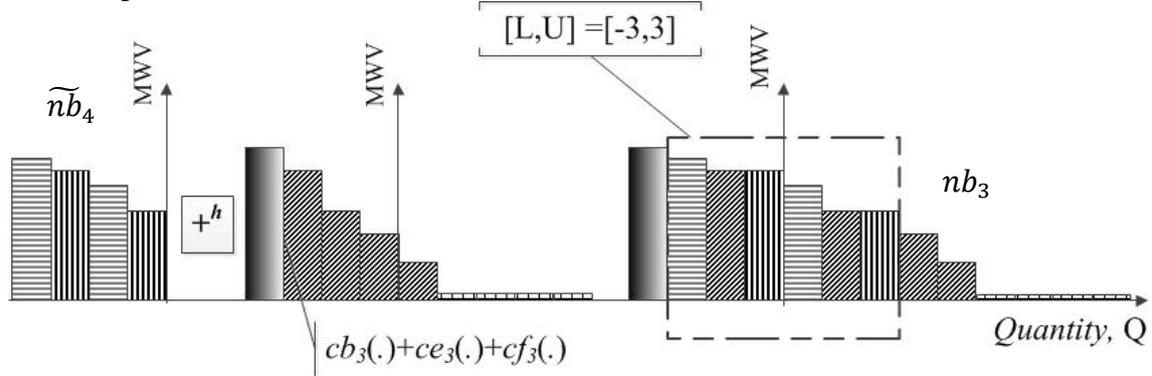


Figure 5: Processing node 3 and truncating across arc (3,2)

3.4 Processing upstream

The flow direction on the syphon was indeterminate, while the other arcs processed so far have had water flow ‘towards’ the reservoir. But the water flow direction on arc (1,2) is away from the reservoir. The flow limits on such arcs will imply non-negative upper/lower limits on the net demand curve passed back through the tree, towards the reservoir. And non-consumptive uses on such arcs will add positive increments to the prices in those demand curves, representing the additional non-consumptive benefit from flow away from the reservoir, in this case. Figure 6 illustrates the formation of $\widetilde{nb}_2: [0,4]$.

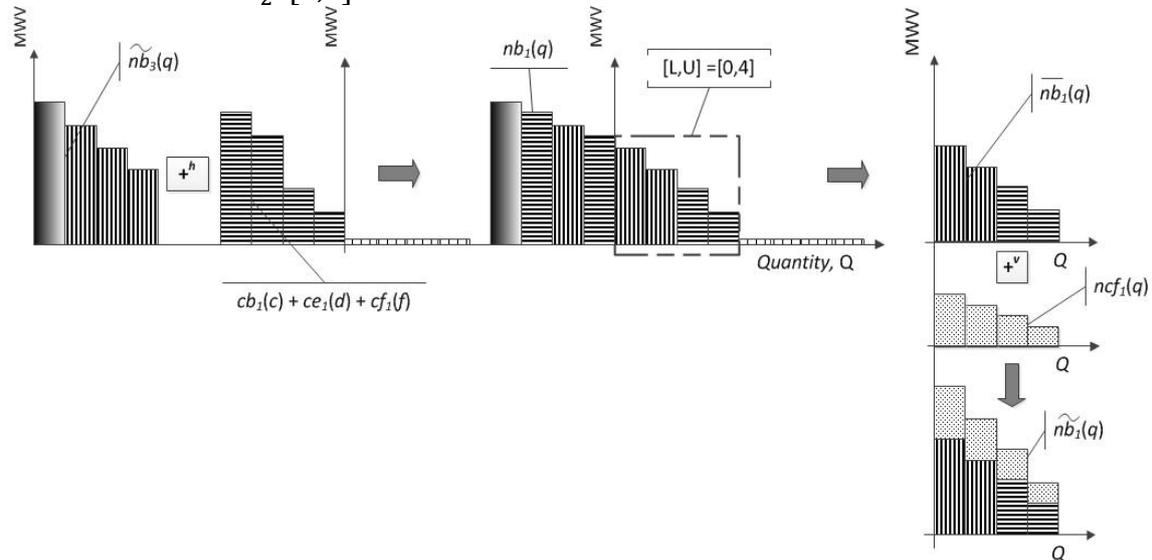


Figure 6: Left: horizontal addition of DCF of previous nodes. Right: local demands at node 1 and vertical addition of non-consumptive flows

Node 1 is then processed in a similar manner. Finally, we reach node 0, where we obtain the DCR for the reservoir itself by horizontal addition of the net demand curves for node 1 (\widetilde{nb}_1) and node 7 (\widetilde{nb}_7).

4 Using the multi-nodal DCR in a long term Stochastic CDDP model

The multi-nodal CDDP constructs a monotone decreasing DCR for each period. But that only defines the optimal release, and intra-period market-clearing, as a function of the reservoir's marginal water value (MWV). We do not know what the release should actually be, and have not "cleared the multi-period market", until we know the reservoir MWV. So the final step in the process must be to identify a particular solution (at least for the first market period), starting from a particular storage level.

Before we do that, though, we must determine the optimal release policy, over time. And we do that by applying the stochastic CDDP model of Starkey et al. (2012) to determine a DCS, defining the MWV of stored water, as a function of storage level, in each period. That algorithm assumes a DCS for the last period, and works backwards to determine the DCS for the beginning of each period from the DCS at the end of that period, and the intra-period DCR for that period, as determined by the algorithm above.

5 Conclusions and further research

In this paper, we illustrated a nodal catchment model that can be applied to optimise usage over time, or to clear an inter-temporal market for any catchment with one reservoir and a tree configuration. To solve the model, we proposed a simple and efficient two-level application of CDDP. First, a multi-nodal deterministic CDDP, presented here, constructs aggregate demand curves for release in each period. Then a stochastic CDDP constructs aggregate demand curves for storage in the single reservoir for each period. The complexity of the intra-period network can readily be increased by adding upstream and downstream nodes, and physical links, while still maintaining efficient computation.

Increasing the number of long term storage reservoirs will increase the complexity of the algorithm, because the "curse of dimensionality" will eventually apply to any DP-based technique for multiple reservoirs. But Read and Hindsberger (2010) describe several two reservoir implementations of stochastic CDDP, while Read, Dye and Read (2012) describe development work on a multi-reservoir generalisation of the stochastic inter-period algorithm.

6 References

- Dye, S., Read, E. G., Read, R. A., & Starkey, S. R. (2012). *Easy Implementations of Generalised Stochastic CDDP Models for Market Simulation Studies* Paper presented at the 4th IEEE and Cigré International Workshop on Hydro Scheduling in Competitive Markets, Bergen, Norway.
- Read, E., & Hindsberger, M. (2010). Constructive dual DP for reservoir optimization. *Handbook of Power Systems I*, 3-32.
- Read, R., Dye, S., & Read, E. G. (2012). *Generalized CDDP for Reservoir Management*. Working paper, . Department of Management, University of Canterbury.
- Scott, T. J., & Read, E. G. (1996). Modelling hydro reservoir operation in a deregulated electricity market. *International Transactions in Operational Research*, 3(3), 243-253.
- Starkey, S. R., Dye, S., Read, E. G., & Read, R. A. (2012). *Stochastic vs. Deterministic Water Market Design: Some Experimental Results*. Paper presented at the 4th IEEE and Cigré International Workshop on Hydro Scheduling in Competitive Markets, Bergen, Norway.

A Simulation Model of Military Pilot Training

Jason Markham & Nebojsa Djorovic
New Zealand Defence Force
Wellington
New Zealand

jason.markham@nzdf.mil.nz or nebojsa.djorovic@nzdf.mil.nz

Abstract

The Royal New Zealand Air Force trains and employs military pilots on six different aircraft types, requiring sufficient flow of personnel through multiple training stages. Discrete event simulation has been used to assess training throughput optimisation policies because of the discrete, variable and probabilistic nature of pilot training flows. This paper describes the use of a sub-model for training vacancy allocation and streaming decisions. Some preliminary results are also discussed and future work directions are also discussed.

Key words: Discrete event simulation, pilot, military, training, Arena, sub-model

1 Introduction

The Royal New Zealand Air force (RNZAF) operates six fleets of aircraft, in addition to two fleets of training aircraft. After recruitment and officer training, pilot trainees complete the wings course and are then streamed to helicopter or fixed wing aircraft. Conversion is the final training stage before employment on an operational aircraft, occurring on average about 5 years after first contact with recruiters, as illustrated in Table 1 below. There are three training points where pilot throughput is most at risk due to low selection/pass rates and high variability: (1) recruitment, (2) wings course and (3) co-pilot.

Table 1. Typical pilot training throughput parameters

	Training duration (years)	Selection or pass rate (%)	Selection/pass variability
Recruitment	1.0	2%	High
Officer training	0.5	95%	Low
Wings course	1.2	65%	High
Stream training	0.5 - 1.5	95%	Low
Conversion course	0.5 - 1.0	95%	Low
Co-pilot	3 - 4	80%	Moderate
Captain	3 - 4	99%	Low

2 Problem

A model of pilot training has been developed to help RNZAF leaders answer a number of current research questions:

- What is the minimum recruitment rate to sustain aircrew requirements?
- How many new training aircraft are required?
- What is the optimum length of pilot training return of service obligation?
- How will the two new helicopter fleets affect pilot throughput?

This model addresses the problem: *What is the optimum combination of recruitment, streaming and posting rules to achieve the required number of pilots?*

3 Model Description

There is relatively little flexibility around the timing or capacity of pilot training, so throughput variability is instead addressed by just-in-time streaming and flexible posting durations for co-pilots and captains. A discrete event simulation (DES) model has been developed in the proprietary software Arena, building on the generic workforce modelling approach presented in Djorovic, Gosse, Markham and Ta'ala (2009). Pilot entities advance through successive training stages with streaming decisions at the branch points in Figure 1 (below). The model uses a monthly time-step.

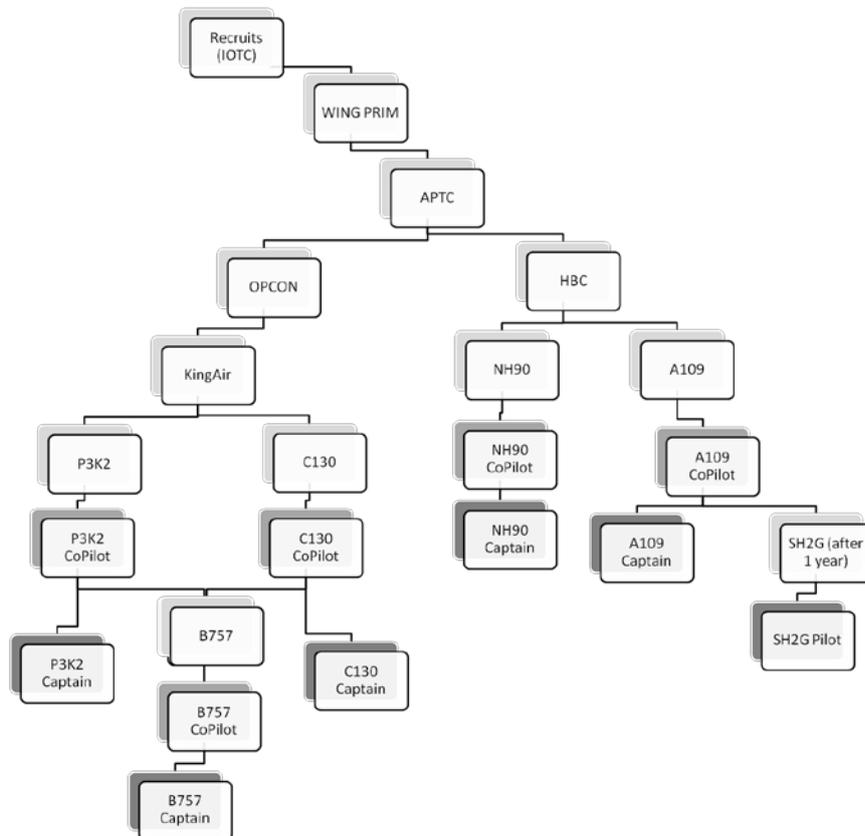


Figure 1. Pilot training flows (flows from top to bottom)

A sub-model has been developed to allocate training vacancies and make streaming decisions (Figure 2). This approach simplifies the model layout and avoids structural replication. Generic sub-models can be used across families of workforce models to minimise testing overheads and improve confidence in their use.

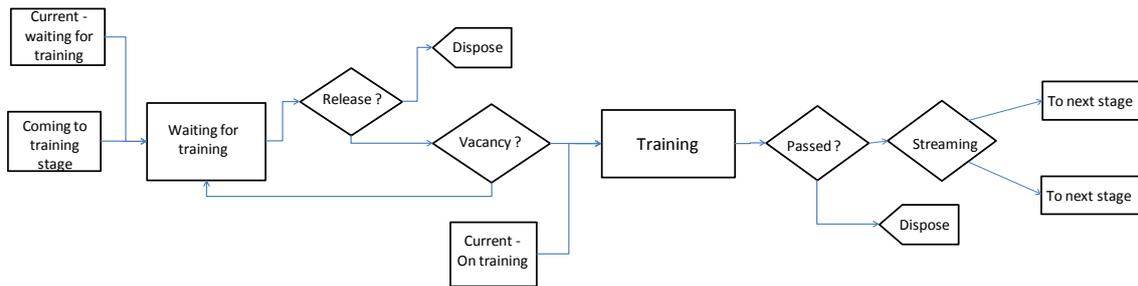


Figure 2. Training sub-model

4 Preliminary Results

Some preliminary results are shown for the P3K2 aircraft (Figure 3) which has historically experienced pilot shortages. These results illustrate how a current shortage of trainee pilots will cause a dip in co-pilots in 2014 followed by very low captain numbers in 2016. The model streaming rules and recruitment flows allow suitable responses to this problem to be explored.

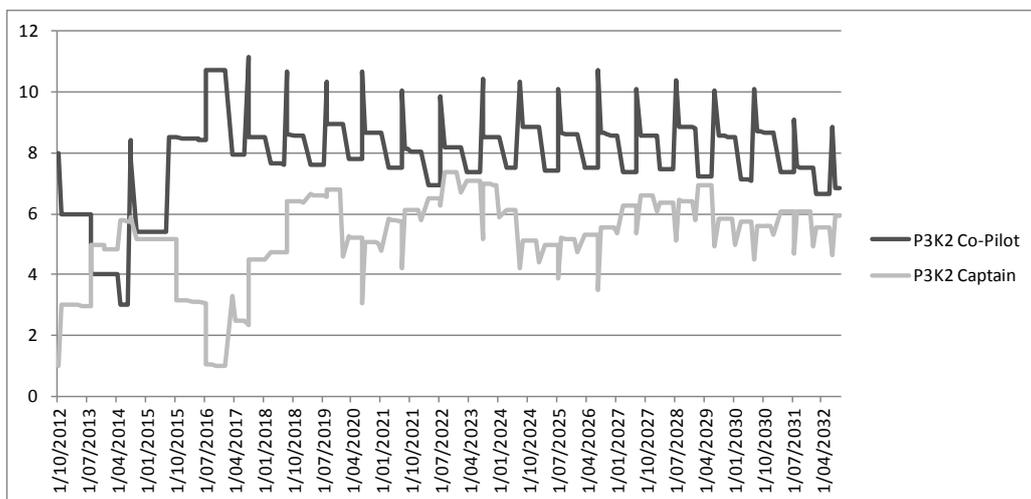


Figure 3. Simulation results for P3K2 crews (20 replications)

5 Conclusion

Discrete event simulation continues to provide the New Zealand Defence Force with a useful workforce analytical capability. The model of pilot training demonstrates the use of a sub-model and illustrates how the model will be used to answer topical research questions. Future work will be directed towards seeking optimum solutions and socialising the model with decision makers.

6 References

Djorovic, N., Gosse, M., Markham, J., Ta'ala, J. 2009. "Initial use of discrete event simulation for New Zealand military workforce analysis." *Proceedings of the ORSNZ 2009 Conference*, pp. 120-128.

SolverStudio for Excel

Andrew J Mason
Department of Engineering Science
University of Auckland
New Zealand
a.mason@auckland.ac.nz

Abstract

We present SolverStudio, <http://solverstudio.org>, a new free Excel add-in that allows users to easily build and run advanced optimisation models in Excel. Unlike the standard Solver provided with Excel, SolverStudio supports models developed using modern modelling languages including AMPL, GMPL, PuLP and Gurobi's Python environment. SolverStudio also supports solving AMPL models in the cloud using the NEOS server. SolverStudio allows the user to define data items as cell ranges on the spreadsheet, and then seamlessly manages data transfers between these and the modelling environment.

Key words: Excel, SolverStudio, AMPL, Gurobi, PuLP, Modelling Language, Optimization, Optimisation.

1 Introduction

For many years, we have introduced our students to optimisation using Excel with its built-in Solver optimiser and its more powerful counterpart OpenSolver (OpenSolver 2012). However, we also want our students to have access to modelling language such as AMPL (AMPL 2012) or GAMS (GAMS 2012). These modelling languages provide a formal modelling environment that emphasises the mathematical structure of the model and clearly distinguishes this from the model data. However, the change from a graphical interface such as Excel's to the command-line interfaces used by these tools presents an unnecessary barrier for our students. They are also often puzzled by the somewhat arcane text-based files required to define the data required for their models. We have developed SolverStudio, a new Excel add-in, to help address these concerns by allowing optimization modelling languages, such as AMPL to be used, within the familiar Excel environment.

Our experience developing OpenSolver has shown that users are building surprisingly large optimisation models in Excel. (One such model we were sent had 70,000 variables and a similar number of constraints.) Excel spreadsheets (and thus models) are notoriously difficult to debug and verify, making the validation of these models very difficult. Furthermore, if OpenSolver is being used to solve these model, a large amount of time is be spent by OpenSolver simply extracting the model from the spreadsheet. By clearly separating the model and its data, SolverStudio can provide a more robust formal modelling environment. In addition, SolverStudio is able to avoid the slow model extraction process required by OpenSolver, and thus can deliver much faster solution times for large models. This papers introduces SolverStudio and provides a brief overview of its operation.

SolverStudio is not the first tool for integrating modelling languages with Excel. Systems available commercially include the AIMMS Excel add-in (AIMMS 2012) that is part of the larger AIMMS modelling and interface development system, and the Microsoft Solver Foundation add-in (Solver Foundation 2012). Both of these have their own proprietary modelling languages that, we suggest, are not as commonly used as the languages supported by SolverStudio. They also have licensing requirements that make widespread usage more difficult. Furthermore, the future of Microsoft Solver Foundation is somewhat uncertain; see DevLabs (2012). We believe SolverStudio avoids these difficulties.

2 SolverStudio Operation

The SolverStudio add-in is available as a free download from <http://solverstudio.org>. After downloading, a setup programme is run that installs the SolverStudio add-in into Excel. SolverStudio requires the Microsoft “.Net 4” software, which usually requires administrator privileges to be installed. However, if this is already available (as is the case with most current versions of Windows), then SolverStudio can typically be installed by users without any special administrator privileges.

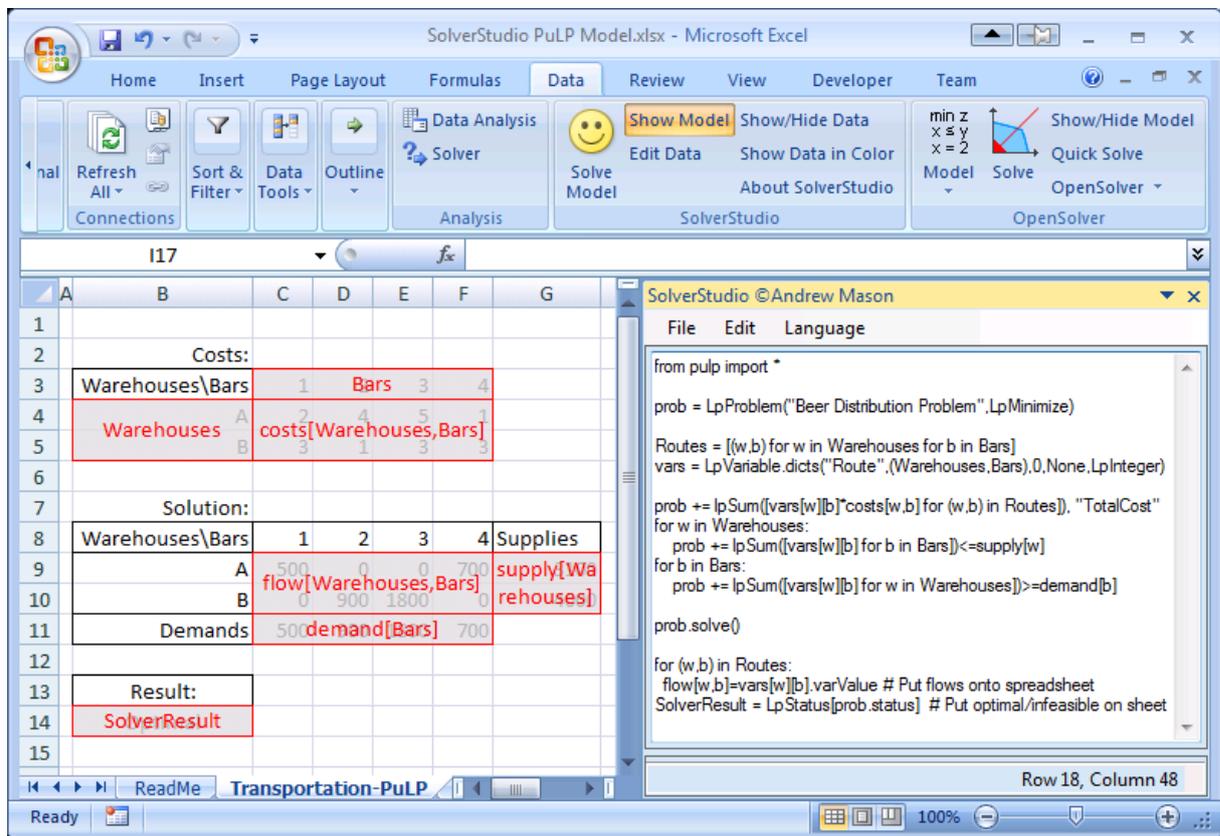


Figure 1: A SolverStudio transportation model in PuLP with the data items highlighted.

Once installed, SolverStudio presents the user with a new set of commands in the Data ribbon, as shown in Figure 1. This figure shows, on the right, the model editing pane that SolverStudio adds to Excel; this pane is used to create and edit the optimisation model. Figure 1 also shows what we term the ‘data items’ for this model that have been defined as named (and indexed) ranges on the spreadsheet. (In this figure, these data items have been highlighted by SolverStudio on the sheet.) In this model, we

have two sets (named Bars and Warehouses respectively), three indexed parameters ('supply,' 'demand' and 'flow') and a single-cell data item named 'SolverResult'. The optimisation model, which in this example is written in Python using PuLP (PuLP 2012), simply references these data items by name. (These set and indexed parameter data items are available as Python lists and dictionaries respectively in the code.) When the model is solved, SolverStudio automatically creates these Python variables, loads them with data from the spreadsheet, and then makes them available to the model. Any changes made to these by the model code are then written back to the spreadsheet when the optimisation finishes.

As Figure 1 shows, SolverStudio allows users to simply enter their data into a spreadsheet, avoiding the complex text-based data representation required by modelling languages such as PuLP and AMPL. Once the data items have been entered on the sheet, their cell ranges are then named for use in the model by using SolverStudio's Data Items editor, shown in Figure 2. In this example, the data items editor shows the entries used to define the data items shown in Figure 1.

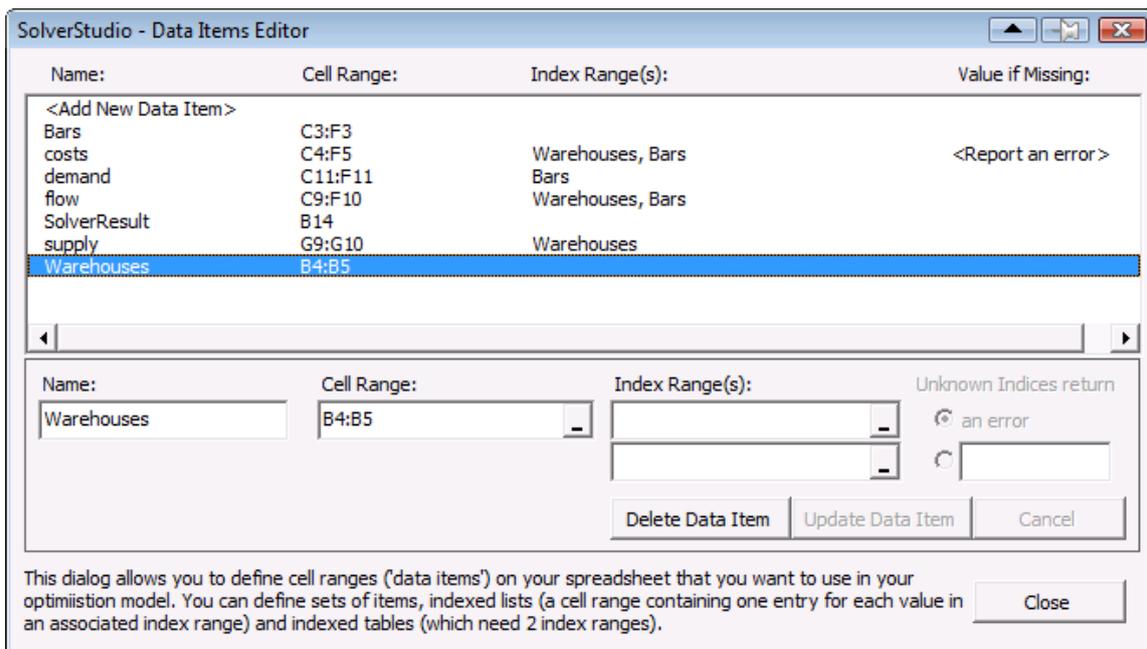


Figure 2: SolverStudio's data items editor showing the data items in Figure 1.

Once the model and its data have been created, the SolverStudio "Solve Model" button then executes the model while managing the required data transfers with the spreadsheet data items. In the example above, code is included in the model to write the final optimised results into the 'flow' data item, thus making the solution visible on the spreadsheet. The 'SolverResult' data item is similarly used to display the optimisation result (typically 'optimal' or 'infeasible') on the sheet.

The example above uses the Python-based open-source PuLP modelling language developed by Stuart Mitchell. However, SolverStudio supports a wide range of modelling languages. As well as PuLP, these currently include the commercial modelling systems AMPL and GAMS, the open source AMPL look-alike GMPL (GLPK 2012), and the commercial Python-based Gurobi modelling library (Gurobi 2012). The open-source languages PuLP and GMPL are included in the SolverStudio download, while the commercial tools need to be purchased and installed by the user. However, both AMPL and GAMS provide free restricted versions of their systems,

while Gurobi make their code available for academic use; SolverStudio can work with all of these versions when installed by the user. Indeed, to encourage student use of AMPL, SolverStudio provides a menu item that automatically downloads and installs the free restricted version of AMPL. SolverStudio is designed to allow easy integration of additional languages, and so we hope the list of languages will grow as time permits.

3 SolverStudio Usage and Cloud-based NEOS Optimisation

We recently used SolverStudio with the free restricted version of AMPL to introduce our 3rd year Engineering Science students to the AMPL modelling language. As part of this course, the students needed to build and solve a model that exceeded the 300 variable/constraint limit of the free AMPL version. One alternative was to use the AMPL look-alike GMPL. GMPL lacks the ordered sets found in AMPL, and so we needed to change our model to work around this. However, we then found that the GMPL solver (GLPK) was much slower than the Gurobi solver included with AMPL. Another option was to use the online NEOS system (NEOS 2012) which allows AMPL models to be submitted via a web interface for solving using a wide range of optimisers provided by NEOS. NEOS also provides an XML-RPC interface which allows optimisation jobs to be submitted programmatically. The latest version of SolverStudio now includes support for solving AMPL models using NEOS. The NEOS solvers do not have the limitations found in the free version of AMPL, and so adding this NEOS support allowed our students to easily solve their large AMPL problems. SolverStudio automatically manages the model changes and solver selection process required by NEOS, resulting in a seamless integration with NEOS that closely mimics the experience of having AMPL installed locally.

As well as using SolverStudio in our teaching, we are also working with Stuart Mitchell (the author of PuLP) to support a SolverStudio implementation of a very large model for determining optimal interventions to reduce water usage in Florida, USA. We hope this is the first of many commercial uses of SolverStudio.

4 Conclusions

We have developed and publicly released SolverStudio, a new tool for Excel that makes modern optimisation modelling languages available within the familiar environment of Excel. We believe that SolverStudio provides a natural upgrade path for users of Solver and OpenSolver wishing to build and solve larger optimisation models. SolverStudio is unique in providing access to a wide range of commercial and open-source modelling languages from one integrated platform. We look forward to seeing SolverStudio become a commonly used tool for both teaching and implementing optimisation modelling languages.

Acknowledgments

SolverStudio would not have been possible without the assistance of Stuart Mitchell who created a customised version of his PuLP modelling system for SolverStudio. We gratefully acknowledge the contributions from Cameron Walker and Michael O'Sullivan and their students in testing SolverStudio and providing the motivation for the integration with NEOS. We also acknowledge the useful input received from AMPL and Gurobi during the development of SolverStudio.

5 References

- AIMMS (2012), AIMMS Excel Add-In User's Guide - Introduction to the AIMMS Excel Add-In. Retrieved 20 July 2012.
http://www.aimms.com/aimms/download/manuals/aimms3exc_introduction.pdf
- AMPL (2012), AMPL®: A Modeling Language for Mathematical Programming. Retrieved 1 November 2012, <http://ampl.com/>
- DevLabs (2012), The Future of Microsoft Solver Foundation. Retrieved 20 July 2012.
<http://social.msdn.microsoft.com/Forums/en-US/solverfoundation/threads>
- GAMS (2012), GAMS Home Page. Retrieved 20 July 2012.
<http://http://www.gams.com/>
- GLPK (2012), GLPK (GNU Linear Programming Kit). Retrieved 20 July 2012.
<http://www.gnu.org/software/glpk/glpk.html>
- Gurobi (2012), Gurobi Optimization. Retrieved 20 July 2012, <http://www.gurobi.com/>
- NEOS (2012), NEOS Server for Optimization. Retrieved November 2012.
<http://www.neos-server.org/neos/>
- OpenSolver (2012), OpenSolver for Excel. Retrieved 10 November 2012.
<http://opensolver.org>
- PuLP (2012), Pulp: An LP modelling system written in Python. Retrieved 20 July 2012.
<https://projects.coin-or.org/PuLP>
- Solver Foundation (2012). Microsoft Solver Foundation. Retrieved 20 July 2012.
<http://msdn.microsoft.com/en-us/devlabs/hh145003.aspx>

Project Management: A Comparison of Three Popular Approaches

Maryam Mirzaei, Victoria J. Mabin
School of Management
Victoria University of Wellington
Maryam.Mirzaei@vuw.ac.nz, Vicky.Mabin@vuw.ac.nz

Abstract

We all undertake activities which produce something unique and therefore could be categorized as projects. There are several approaches to project management, some specific to project management and some tailored to solve project management problems. This paper provides an overview of the Project Management Body of Knowledge (PMBOK), Lean Project Management (LPM) and Critical Chain Project Management (CCPM). There are characteristics which are particular to each of these and there are similarities among them. For example the core idea in PMBOK is that a project can be isolated into sub-processes and successful completion of sub processes will lead to overall success of the project. CCPM and LPM are both developed within manufacturing and share the pull mechanism and flow. However, CCPM provides focus, emphasis on schedule and accommodates uncertainty; while LPM concerns eliminating waste and reducing cost. Furthermore PMBOK and LPM attempt to improve efficiency throughout the system while CCPM focuses attention on the 'Critical Chain'. The aim of this paper is to highlight the conceptual assumptions behind these approaches and by doing so demonstrate their suitability for different contexts.

Key words: Critical Chain; Lean; PMBOK; Project Management Approach

1. Introduction

One of the most important organisational developments in recent years has been the significant growth in project work across different sectors and industries. However the conceptual base of project management continues to attract criticism for its lack of relevance to practice. More recently new approaches have emerged that challenges fundamental assumptions in the existing practice. Lean project Management (LPM) and Critical Chain Project Management (CCPM) are two examples of redefining projects and underlying assumptions. What differentiates these approaches from other innovations and improvements in project management tools and techniques is the new perspective in the way projects are perceived and symbolized.

There is a tendency in most project management approaches to treat projects as fundamentally similar to each other. Consequently, project management has been conceptualized as a universal phenomenon. Most project management textbooks introduce a set of functions and activities considered common to all projects, and overlook the fundamental differences that exist across projects.

However projects demonstrate fundamental differences in their nature of scope, the relevance of uncertainty, urgency and complexity. Dvir, Lipovetsky, Shenhar, and

Tishler, (1998) confirm that projects have fewer characteristics in common than previously considered. Research on project success and classification tends to recognize and address this issue. However, the focus is on relevance of techniques rather than approaches. On the other hand new approaches in project management are based on a different set of assumptions which may fit many projects, but still not all types of projects considering the differences that projects demonstrate. The focus of this paper is to describe the theories and underlying assumptions between three approaches, namely Project Management Body of Knowledge (PMBOK) as the representative of traditional Project Management; Lean Project Management (LPM) and Critical Chain Project Management (CCPM) as two more recent approaches. Understanding those assumptions as well as understanding projects will provide a good guideline to assess how a new approach can contribute to project success. A brief description of each approach will be followed by a comparison that highlights their similarity and differences in order to examine their suitability for various types of projects.

2. Overview of Project Management Body of Knowledge (PMBOK)

One of the most influential and referenced sources in project management is The PMBOK (Kendall, Pitagorsky, & Hulett, 2001; Reich, 2006). The first edition of PMBOK was published in 1996 as a collection of processes and knowledge areas which were generally accepted as best practice within the project management discipline. As an internationally recognized standard, PMBOK provides the fundamentals of project management, irrespective of the type of project, be it construction, software, engineering, automotive or many more. PMBOK defines projects as a “temporary endeavour undertaken to create a unique product, service or result” and project management as the application of knowledge, skills, tools, techniques and processes to effectively manage a team towards this final deliverable (PMI, 2008, pp. 5 &6).

PMBOK recognizes 5 basic process groups and 9 knowledge areas which are claimed to be typical of almost all projects. The basic concepts are applicable to projects, programs and operations. The five basic process groups are:

1. **Initiating:** defines and authorizes the project (i.e. create project charter).
2. **Planning:** shapes the outcomes/goals for the project by knowledge area.
3. **Executing:** carries out project plans.
4. **Controlling and Monitoring:** assesses actual project outcomes to planned targets and makes corrective actions when necessary.
5. **Closing:** obtains a formal acceptance of the product/service by stakeholders and tapers out project activities in a planned, organized fashion.

Each process is described in term of Inputs (documents, plans, designs, etc.), Tools and Techniques (mechanisms applied to inputs), and Outputs (documents, products, etc.), in a way that provide comprehensive guidance for someone who is going to apply it. The nine knowledge areas are: Integration, Scope, Time, Cost, Quality, Human Resource, Communications, Risk, and Procurement.

The guide is very comprehensive; each knowledge area contains some or all of the project management processes. Throughout the book there is an emphasis on processes and procedures necessary for project management competencies. Tasks are considered as discrete entities and the durations, costs, start dates and finish dates are precisely

defined. The approach in planning is deterministic and reductionist (Fitsilis, 2008). A brief introduction to the theories behind each process group will be outlined next:

A large part of the book is assigned to planning (Koskela & Howell, 2002). The emphasis is on detailed planning and following the original plan throughout the project (Cicmil, Williams, Thomas, & Hodgson, 2006). The structure of the book demonstrates the influence of transformation theory of production in which every process is viewed as an input-output system. Transformation theory also promotes the idea that in order to manage the whole project successfully we need to manage its parts successfully, which results in decomposition of activities.

The execution processes are based on job despatching theory in manufacturing, which consists of deciding and communicating the assignment to the job station. However in project management the first part falls within planning and therefore the execution is reduced to mere communication.

Controlling processes are based conceptually on a cybernetic model of management control (thermostat model) with the assumption that there is a standard of performance which can be measured. The variance can be used for corrective actions in order to reach the desired standard. This is also the same concept as a feedback control model (Koskela & Howell, 2002).

The idea in PMBOK is that the project / process can be isolated into sub-processes and successful completion of sub-processes will lead to overall success of the project. The project manager is required to oversee every task regardless of its importance. Measures such as Earned Value usually are the ratio of output to input and treat all tasks as per their estimated cost; therefore no task is considered superior to another.

3. Overview of Lean Project Management (LMP)

It is widely agreed that LPM is based on the Toyota product system (TPS) methodology (Ballard & Howell, 2003; Leach, 2006). Krafcik (1988) used the term “lean” when he compared American, European and Japanese motor vehicle firms. ‘Lean’ was chosen because the Japanese used less of everything; time, resources and money and produced vehicles with fewer defects and greater variety than their competitors. Some of the features and fundamental principles that characterize the concept of lean production are: *waste reduction, continuous improvement, zero defects, multifunctional teams and decentralized responsibilities.*

The first fundamental principle in lean production is *waste reduction* (Chen, Lindeke, & Wyrick, 2010; Hines, Francis, & Found, 2006; Karlsson & Ahlstrom, 1996). Ono (1988) identified seven forms of waste in manufacturing which are, Defects in products, Overproduction of goods not needed, Inventories of goods awaiting processing or consumption, Unnecessary processing, Unnecessary movement of people, Unnecessary transport of goods, Waiting by employees for process equipment to finish work or for an upstream activity to complete.

Continuous improvement directs eliminating waste to reducing cost, as well as improving products and processes in order to increase customer satisfaction (Chen et al., 2010). This is a never ending process. In fact it is a direction rather than a state (Ballard & Tommelein, 2012; Karlsson & Ahlstrom, 1996).

Zero defects is important because in order to maintain a continuous flow of a production process, all products have to be defect free throughout the process at the production line. The intention is that errors should be prevented before they occur, the underlying objective with such inspection is not to find defects but to prevent them (Karlsson & Ahlstrom, 1996).

Multifunctional team and decentralized responsibilities means employees can perform several different tasks. This increases flexibility and decreases dependency on individual employees which reduces vulnerability of production system state. Furthermore there is no supervisory level in the hierarchy. In practice this means that employees at every level of the organization are looking and experimenting to improve their own work. Team leaders take supervisory roles in the form of coaches. As a result, the number of hierarchical levels can be reduced (Karlsson & Ahlstrom, 1996).

Since the introduction of lean production, the understanding of the above concepts has evolved. Lean thinking emerged as a result of narrowing lean production concepts to a set of five principles proposed by Womack & Jones (1997) that focus on elimination of waste. The principles of lean thinking are as follows:

- Precisely Specify *Value*
- Identify the *Value Stream*
- Make *Value Flow*
- Let the Customer *Pull Value*
- Pursue *Perfection*.

Lean thinking then became applied to project management. It was first linked to construction when Koskela (1992) argued that construction industry needs to learn from the advancements in manufacturing. Construction is a type of manufacturing; however buildings are too large to move through fixed workstations. Instead workstations become mobile and move through the buildings. The sequence and timing of these workstation movements are driven by planning rather than by a fixed structure. This highlights the project management aspect of construction. Within the LPM, projects are defined as temporary production systems. While production is defined as designing and making things, and when this is done for the first time it is called a project. Therefore lean project is defined as “temporary production system structured to deliver the product while maximizing value and minimizing waste” (Ballard & Howell, 2003). However the nature of projects is considered to be a more fundamental form of production system than factory productions (Ballard, 2005).

Adaptation of lean thinking into project management is marked by the last planner method. It promotes a clear set of objectives for the delivery process, concurrent design of product and process. The scheduling in the last planner method is backwards with a focus on the work that adds value for the client. It does not include details of the whole plan because such details in early stages do not add any value. Instead last planner uses ‘lookahead window’, which is usually 6 weeks ahead, and detail of tasks and their prerequisites are analysed only for that duration ahead. The aim is to have tasks completed as promised. This measure of predictability is Percentage Promised Completed on time (PPC). Last planner promotes a shift from controlling and motivating to engage planning, preparing, and navigating with those people performing the work (Ballard & Howell, 2003; Ballard & Tommelein, 2012; Howell, 1999). The

uncertainty is addressed by ranking tasks according to the degree of uncertainty and allocating available time to fragile activities in rank order (Ballard & Howell, 2003). For costing purposes, Activity-Based Costing (ABC) which is based on “flow view” is used. It is claimed that ABC is not only preventing cost misrepresentation but also provides information that can help to eliminate waste (Kim & Ballard, 2001).

Mossman (2004) describes four main elements in last planner, which are as follows:

Programming Workshop: collaboratively creating and agreeing the production sequence this element requires early meeting of all suppliers and contractors and obtaining agreed goals. With an agreed program it is possible in this stage to look for ways to compress duration together with suppliers and contractors.

Make-ready: this process is in fact a constraint analysis which checks systematically that everything is in place for each of the tasks in the Look-Ahead window.

Production Planning: collaboratively agreeing production tasks for the next day or week through daily or weekly production planning meetings (PPM). The last planner also updates completion of the precedence and the availability of material, information, or any other prerequisites for future tasks.

Continuous Improvement: learning and improving PPC, the project, planning, and production processes.

There are examples of successful application of lean to construction and other types of projects (Ballard & Tommelein, 2012; Gabriel, 1997) and also debates on its applicability in various contexts, cultural barriers in implementing it, and its negative impacts on human resources (Green, 1999). In response to the latter claim, Howell and Ballard (1999) argued that participating in decision making and multi-skilling in fact enrich jobs.

4. Overview of Critical Chain Project Management (CCPM)

The Critical Chain concept was introduced as an application of Theory of Constraints (TOC) to project management (Goldratt, 1997). TOC is based on the assumption that “any system must have a constraint. Otherwise its output would increase without bound, or go to zero.” (Noreen, Smith, & Mackey, 1995). Since the introduction of Critical Chain the concept has been described in books either as a standalone method (Leach, 2000; Newbold, 1998; Wiefeling, 2007) or in combination with other project management approaches such as agile and lean project management (Anderson, 2003; Leach, 2006). Leach (1999) provides a description of the concepts and assumptions underlying CCPM. He addresses six undesirable effects in project management practice, namely: Excessive activity duration, Lack of positive variation, Failure to pass on positive variation, Delays caused by path merging, Multi-tasking, Loss of focus.

The term “effect” places an emphasis on the fact that there are underlying causes for them. He further argues that the core cause is “failure to manage uncertainty” which was claimed by Goldratt’s TOC analysis to be the leading problem of project failure (Leach, 1999). Patrick (2001) argued that a major contribution of CCPM is bridging between scope, time and risk. Leach (1999) also mentions the three theories used in CCPM: Theory of Constraints, Common cause variation, Statistical law of aggregation

4.1 Theory of Constraints (TOC): Within TOC, a project is symbolized as a system. The basic assumption is that, the project goal is what will be delivered to the larger

organization. Because the project is temporary, the ultimate measure of performance is considered to be the duration of the project (Yang, 2007). Furthermore, Steyn (2002) counts three reasons for such emphasis on scheduling: obtaining positive flow faster, contingency cost of delay, and preventing changes to stakeholders' needs. Leach (2000) also justifies defining the project constraint in terms of schedule by reminding readers the effect of scheduling on project cost and scope.

After defining the project as a system and the schedule as its measure of success, the five focusing steps are applied to it as follows:

1. Identify the system constraint: The longest chain of activities after including resource dependency determines the duration of the project, therefore marks the system constraint with the above measure (Goldratt, 1997; Leach, 2000).
2. Exploit the constraint: this is done by shortening the critical chain. It is assumed that the estimated durations are generally inflated in several layers of management to prevent delay (Goldratt, 1997). Therefore using 50% probable activity times will shorten the Critical Chain.
3. Subordinate everything else to the above decision: this means we need to make sure that every activity on the Critical Chain will commence as soon as the previous activity on the Critical Chain is completed. To achieve this, feeding buffers are employed (Goldratt, 1997; Leach, 2000; Newbold, 1998). "Feeding buffer is a time cushion placed between non Critical Chain work and the Critical Chain to protect the Critical Chain from variation on a non-Critical Chain path of work. It helps determine when to start non critical chain work." (Cox, Boyd, Sullivan, Reid, & Cartier, 2012, p. 76)
4. Elevate the system's constraint: After confirming that we have shortened the Critical Chain with the above two steps, we now look at other possibilities to further shorten the project duration, such as extra employees, new software or machines.
5. If in the previous steps constraints has been broken, go back to step one.

4.2 Common cause variation: Deming (1986) emphasized the importance of differentiating variation causes inherent in the system from other types of variations. Leach (2000) suggests CCPM is sufficient for all common cause variations which he calls internal risk, and leaves project risk management to handle external risks only.

4.3 Statistical law of aggregation: According to this law the project variance is the sum of the individual activity variance (PMI, 1996, p. 116). Therefore in CCPM instead of adding safety allowances for each task, aggregated allowances for uncertainty of estimates and activity performance is added as a buffer at the end of the chain of activities. As variance is the root of squared deviation from each of the expected values it is less than the sum of deviation which is generally used in calculating project duration (Budd & Cerveny, 2010). It is important to differentiate buffers from slack, because unlike slack, buffers are supposed to be consumed, gradually. While high buffer consumption demonstrates an issue to be resolved, no buffer consumption could mean there was a planning problem (durations are set too long). Even fluctuating buffer consumption could mean the project is not under control (Newbold, 2008).

However the application of the above theories will differ in a multi project environment. When there are several projects, CCPM treats each project individually with the above method and then staggers projects according to the bottleneck resource (the most constrained resource) and releases the projects in a way such that there is no idle time for the bottleneck resource. The buffer in such an environment is called the capacity

buffer which ensures bottleneck availability. Many authors have discussed the above process. However they have not offered a standard way to determine the size of capacity buffer. Once again the projects' performance is monitored via buffer consumption with the first priority given to the critical activities over non-critical, and second priority is given to projects with the highest level of project buffer consumption (Cohen, Mandelbaum, & Shtub, 2004; Herroelen & Leus, 2001; Huang, Chen, Li, & Tsai, 2012; Leach, 2000; Newbold, 1998).

CCPM also seeks to change project team behaviour. It encourages reporting early completion of activities and elimination of multitasking and by doing so claims to overcome problems caused by commonly observed human behaviours such as deliberate padding task times, Murphy's Law, Parkinson's Law, and so called Student Syndrome (Goldratt, 1997; Huang et al., 2012; Leach, 1999; Newbold, 1998; Woeppel, 2006).

CCPM has received much praise such as being the direction for project management in the 21st century (Newbold, 1998; Steyn, 2002; Vrincut, 2009). Many authors consider CCPM to be simple and workable with a stable schedule throughout the project (because CCPM does not change during project execution) (Herroelen, Leus, & Demeulemeester, 2002; PMI, 2008; Woeppel, 2006). It minimizes work in progress and focuses on key tasks and resources and does not split attention among numerous tasks. CCPM explicitly recognizes the fact. that the interaction between activity duration, precedence relations, resource requirement, and resource availability is what determines the project duration, not the sum of activity durations at the critical path (Herroelen & Leus, 2001; Herroelen et al., 2002; Raz, Barnes, & Dvir, 2003).

Furthermore Realization Technologies, Inc (2010) cited a list of organizations which have used CCPM and compared their results with their past records. Despite its success, CCPM has received criticism on its innovativeness and applicability to all kind of projects (Mckay & Morton, 1998; Raz et al., 2003). However it has been applies successfully to construction (Deac & Vrincut, 2010; Vrincut, 2009; Yang, 2007) and software development (Groves, Nickson, Reeve, Reeves, & Utting, 2000) many other types of projects (Bevilacqua, Ciarapica, & Giacchetta, 2009; Leach, 1999; Newbold, 2008; Srinivasan, Best, & Chandrasekaran, 2007; Stratton, 1998; Umble & Umble, 2000).

5. Comparison and Conclusion

The followings are some of the basic differences of the discussed models:

- Although all of these models have some links with production, only lean directly compares a project to a temporary production system. CCPM also symbolised a project as system. However PMBOK precisely differentiates projects from all non-project activities. Another similarity between CCPM and LPM is that both models are based on system thinking and holistic approach as opposed to PMBOK which is based on decomposition.
- With regard to human resources, LPM promotes multi-functional teams in order to reduce dependency on critical resources. However, CCPM is most likely to accept such dependency and plan the project around it. PMBOK is very comprehensive and broad with human resource management, although there is no particular suggestion for addressing scarce resources.

- With regard to stakeholders, PMBOK suggests overseeing all stakeholders according to their influence on the project, while LPM specifically emphasis meeting the precise needs of the final customer as the most important stakeholder. CCPM considers the system owner as the one who decides what the goal of the system is and therefore where the focus will be.
- In scheduling PMBOK suggests starting all activities as soon as possible and promotes the push mechanism. In contrast LPM promotes pull mechanism using the last planner to systematically reduce supply chain leads times and work in progress. CCPM starts the critical activities as soon as possible and non-critical activities as late as feeding buffers allow and by doing so reduces the work in progress and the total duration time.
- While PMBOK is more deterministic and deals with uncertainty as a separate knowledge area, both LPM and CCPM emphasise uncertainty as a leading concern throughout planning and execution. LPM assumes there is uncertainty of scope. Therefore in LPM planning is suggested to be an on-going process that does not provide many details on the latter stages of the project. CCPM manages the project around an uncertain schedule and uses the buffer management for keeping the project on schedule.
- LPM focuses on value and product rather than process, in contrast to process measures such as buffer consumption measures in CCPM, and the thermostat model in PMBOK. In LPM product and process is designed together. The focus is to eliminate all waste from the production process and not allocate any resources for an activity that does not add value for the customer. The delivery of the final product and its quality is the measure of success.

Despite the fundamental differences that these models demonstrate some authors try to incorporate these models together. For example Leach (2006) has used CCPM concepts in his book “Lean Project Management: Eight Principles for Success”. PMBOK has also included Critical Chain as a scheduling method without taking into account the fundamental roles of buffer management. Furthermore (Kendall et al., 2001) suggested incorporating CCPM and PMBOK and has shown how to organize the CCPM concepts around the knowledge areas of PMBOK. Horman and Keneley (1996) also suggested that LPM is not necessarily different from traditional project management, but rather is a complement to or maturation of PMBOK principles. Many of the above combined models are based on one model and use the other model inefficiently. Combining these models requires further research and proper recognition of both models with their associated assumptions and metaphors in order to avoid misinterpretation.

The above discussion clearly demonstrates that each approach to project management is based on different fundamental assumptions and theories. The ways projects are perceived, defined and symbolized in each model create different mind sets. This paper has highlighted some of the key differences between these three approaches. Considering the differences that projects demonstrate, future research is required to propose the suitability of each to a particular type of project.

References:

Anderson, D. J. (2003). *Agile Management for Software Engineering: Applying the Theory of Constraints for Business Results*. New Jersey: Pearson Professional Education.

- Ballard, G. (2005). Construction: One Type of Project Production System. *13th International Group for Lean Construction Conference: Proceedings* (pp. 29–35). International Group on Lean Construction.
- Ballard, G., & Howell, G. (2003). Lean project management. *Building Research & Information*, 31(2), 119–133. doi:10.1080/09613210301997
- Ballard, G., & Tommelein, I. (2012). Lean management methods for complex projects. *Engineering Project Organization Journal*, 2(1-2), 85–96.
- Bevilacqua, M., Ciarapica, F. E., & Giacchetta, G. (2009). Critical chain and risk analysis applied to high-risk industry maintenance: A case study. *International Journal of Project Management*, 27(4), 419–432. doi:10.1016/j.ijproman.2008.06.006
- Chen, H., Lindeke, R. R., & Wyrick, D. A. (2010). Lean automated manufacturing: avoiding the pitfalls to embrace the opportunities. *Assembly Automation*, 30(2), 117–123.
- Cicmil, S., Williams, T., Thomas, J., & Hodgson, D. (2006). Rethinking Project Management: Researching the actuality of projects. *International Journal of Project Management*, 24(8), 675–686. doi:10.1016/j.ijproman.2006.08.006
- Cohen, I., Mandelbaum, A., & Shtub, A. (2004). Multi-project scheduling and control: a process-based comparative study of the critical chain methodology and some alternatives. *Project Management Journal*, 35(2), 39–50.
- Cox, J. F., Boyd, L. H., Sullivan, T. T., Reid, R. A., & Cartier, B. (Eds.). (2012). *The TOCICO dictionary* (second.). New York: McGraw-Hill.
- Deac, V., & Vrincut, M. (2010). The advantages of using critical chain project management in planning construction projects. *Metalurgia International*, 54–58.
- Deming, W. E. (1986). *Out of the crisis*. Cambridge, Mass.: Massachusetts Institute of Technology, Center for Advanced Engineering Study.
- Dvir, D., Lipovetsky, S., Shenhar, A., & Tishler, A. (1998). In search of project classification: a non-universal approach to project success factors. *Research Policy*, 27(9), 915–935.
- Fitsilis, P. (2008). Comparing PMBOK and Agile Project Management software development processes. In T. Sobh (Ed.), *Advances in Computer and Information Sciences and Engineering* (pp. 378–383). Dordrecht: Springer Netherlands.
- Gabriel, E. (1997). The lean approach to project management. *International Journal of Project Management*, 15(4), 205–209. doi:10.1016/S0263-7863(96)00066-X
- Goldratt, E. M. (1997). *Critical Chain*. Great Barrington: The North River Press.
- Green, S. D. (1999). The dark side of lean construction: exploitation and ideology. *Proceedings IGLC* (Vol. 7, p. 21).
- Groves, L., Nickson, R., Reeve, G., Reeves, S., & Utting, M. (2000). A survey of software development practices in the New Zealand software industry. *Software Engineering Conference, 2000. Proceedings. 2000 Australian* (pp. 189–201). Presented at the Software Engineering Conference, 2000. Proceedings. 2000 Australian.
- Herroelen, W., & Leus, R. (2001). On the merits and pitfalls of critical chain scheduling. *Journal of Operations Management*, 19(5), 559–577.
- Herroelen, W., Leus, R., & Demeulemeester, E. (2002). Critical chain project scheduling-Do not oversimplify. *Project Management Journal*, 33(4), 46–60.
- Hines, P., Francis, M., & Found, P. (2006). Towards lean product lifecycle management: A framework for new product development. *Journal of Manufacturing Technology Management*, 17(7), 866–887.
- Horman, M., & Keneley, R. (1996). The application of lean production to project management. Presented at the Fourth International Workshop on Lean Construction.
- Howell, G. A. (1999). What is lean construction-1999. *Proceedings IGLC* (Vol. 7, p. 1).
- Howell, G., & Ballard, G. (1999). Bringing light to the dark side of lean construction: a response to Stuart Green. *Proc. 7th Ann. Conf. Intl. Group for Lean Construction* (pp. 33–37).
- Huang, C.-L., Chen, H.-C., Li, R.-K., & Tsai, C.-H. (2012). A comparative study of the critical chain and PERT planning methods: no bad human behaviours involved. *International Journal of Academic Research in Business and Social Sciences*, 2(8).
- Karlsson, C., & Ahlstrom, P. (1996). Assessing changes towards lean production. *International Journal of Operations & Production Management*, 16(2), 24–41.
- Kendall, G. I., Pitagorsky, G., & Hulett, D. (2001). Integrating Critical Chain and the PMBOK® Guide. International Institute for Learning, Inc.

- Kim, Y., & Ballard, G. (2001). Activity-based costing and its application to lean construction. *Proceedings of the 9th Annual Conference of the International Group for Lean Construction*, Singapore.
- Koskela, L. (1992). *Application of the new production philosophy to construction* (Vol. 72). Stanford University (Technical Report No. 72, Center for Integrated Facility Engineering, Department of Civil Engineering). Stanford, CA.
- Koskela, L. J., & Howell, G. (2002, June). The underlying theory of project management is obsolete. *Proceedings of the PMI Research Conference*.
- Krafcik, J. F. (1988). Triumph of the lean production system. *Sloan Management Review*, 30(1), 41–52.
- Leach, L. P. (1999). Critical Chain Project Management Improves Project Performance. *Project Management Journal*, 30(2), 39.
- Leach, L. P. (2000). *Critical chain project management*. Boston, MA: Artech House.
- Leach, L. P. (2006). *Lean project management : eight principles for success*. Boise, ID: Booksurge Publishing.
- Mckay, K. N., & Morton, T. E. (1998). Review of: “Critical Chain” Eliyahu M. Goldratt The North River Press Publishing Corporation, Great Barrington, MA, 1997. ISBN 0-88427-153-6. *IIE Transactions*, 30(8), 759–762. doi:10.1080/07408179808966521
- Mossman, A. (2004). Last Planner, Collaborative production planning, Collaborative programme coordination. *Rubicon Associates, Contract Journal Website*.
- Newbold, R. C. (1998). *Project management in the fast lane : applying the theory of constraints*. Boca Raton, FL.: St. Lucie Press.
- Newbold, R. C. (2008). *The billion dollar solution : secrets of prochain project management*. Lake Ridge, VA: ProChain Press.
- Noreen, E. W., Smith, D. A., & Mackey, J. T. (Cor). (1995). *Theory of Constraints and Its Implications for Management Accounting*. Great Barrington, MA: North River Press.
- Ono, T. (1988). *Toyota production system : beyond large-scale production*. Portland, OR.: Productivity Press.
- Patrick, F. S. (2001). Buffering against risk—Critical chain and risk management. *Proceedings of the Project Management Institute Annual Seminars & Symposium*.
- PMI. (2008). *A Guide to the Project Management Body of Knowledge: (4th edition.)*. Project Management Institute.
- Raz, T., Barnes, R., & Dvir, D. (2003). A Critical Look at Critical Chain Project Management. *Project Management Journal*, 34(4), 24.
- Realization Technologies, Inc. (2010). *Getting Durable Results with Critical Chain -- A Field Report (Chapter 4 of Theory of Constraints Handbook)*. New York: McGraw-Hill.
- Reich, B. H. Y. W. (2006). Searching for knowledge in the pmbok guide. *Project Management Journal*, 37(2), 11–26.
- Srinivasan, M. M., Best, W. D., & Chandrasekaran, S. (2007). Warner Robins Air Logistics Center Streamlines Aircraft Repair and Overhaul. *Interfaces*, 37(1), 7–21.
- Steyn, H. (2002). Project management applications of the theory of constraints beyond critical chain scheduling. *International Journal of Project Management*, 20(1), 75–80.
- Stratton, R. (1998). Critical chain project management theory and practice. *Project Management and Systems Engineering*, 4, 149.
- Umble, M., & Umble, E. (2000). Manage your projects for success: An application of the theory of constraints. *Production and Inventory Management Journal*, 41(2), 27–32.
- Vrincut, M. (2009). critical chain project management and the construction industry in Romania. *Review of International Comparative Management*, 10(5).
- Wiefling, K. (2007). *Scrappy project management : the 12 predictable and avoidable pitfalls every project faces*. Cupertino, CA.: Scrappy About.
- Woeppel, M. J. (2006). *Projects in less time : a synopsis of critical chain*. Plano, TX: Pinnacle Strategies.
- Womack, J. P., & Jones, D. T. (1997). Lean Thinking—Banish Waste and Create Wealth in your Corporation. *Journal of the Operational Research Society*, 48(11), 1148–1148.
- Yang, J.-B. (2007). How the Critical Chain Scheduling Method is Working for Construction. *Cost Engineering*, 49(4), 25–32.

The Linear Bi-Objective Multi-Commodity Minimum Cost Flow Problem

Siamak Moradi, Matthias Ehrgott, Andrea Raith.
Department of Engineering Science
The University of Auckland
New Zealand
s.moradi@auckland.ac.nz

Abstract

In this paper, we study minimum cost flow problems with two objectives and multiple commodities. We model the problem as a linear program and we solve it with two methods, the parametric simplex method applied to a bi-objective linear program, as well as the dichotomic method. We also apply a change of variables method to a bi-objective undirected two-commodity version of the problem. We test all methods on different types of networks and report the run times.

Key words: Network flows, bi-objective multi-commodity minimum cost flow problem, multi-objective optimization.

1 Introduction

Network flow models are used to model a variety of real-world decision making problems in a wide range of areas such as transportation, telecommunications, biology, medicine, economics, finance, etc (Ahuja, Magnanti, and Orlin 1993). In many application contexts, there are several objectives as well as several different commodities that have to be taken into account. Thus, multi-objective multi-commodity flow models are appropriate for modelling real-world decision making situations in the field of network optimization. Although there are several algorithms for solving bi-objective single-commodity flow problems (see (Raith and Ehrgott 2009) and references therein), there is only a single paper on the bi-objective flow problem with two commodities (Sedeño-Noda, González-Martín, and Alonso-Rodríguez 2005). In this paper, we model the bi-objective multi-commodity minimum cost flow (*BMCMCF*) problem as a linear program. We solve the problem with a bi-objective version of the simplex method, which is derived from the parametric single-objective simplex algorithm, as well as a dichotomic method. We also implement a change of variables methods, presented by Sedeño-Noda, González-Martín, and Alonso-Rodríguez (2005) for solving the bi-objective undirected two-commodity minimum cost flow (*BU2CMCF*) problem, with the parametric simplex method and the dichotomic

method. We investigate the performance of the methods on different sets of bi-objective network instances with several commodities.

The paper is organised as follows: In Section 2, the *BMCMCF* problem is introduced. In Section 3, we present two methods to solve the linear model of the *BMCMCF* problem. Finally, numerical results are illustrated in Section 4.

2 The *BMCMCF* problem

In the remainder of this paper we use the following orders on \mathbb{R}^2 :

$$\begin{aligned} y^1 \leq y^2 &\iff y_k^1 \leq y_k^2, \quad k = 1, 2, \\ y^1 \leq y^2 &\iff y_k^1 \leq y_k^2, \quad k = 1, 2; \quad y^1 \neq y^2, \\ y^1 < y^2 &\iff y_k^1 < y_k^2, \quad k = 1, 2, \end{aligned}$$

and the sets $\mathbb{R}_>^2 := \{y \in \mathbb{R}^2 : y \succ 0\}$, $\succ \in \{\leq, \geq, >\}$.

Consider a bi-objective optimization problem (*BOP*)

$$\begin{aligned} &\min(y_1(x), y_2(x)) \\ &\text{subject to } x \in \mathcal{X}. \end{aligned} \tag{1}$$

Let \mathcal{X} denote the feasible set of the BOP (1) and let $\mathcal{Y} = \{(y_1(x), y_2(x)) : x \in \mathcal{X}\}$ be the image of \mathcal{X} under the objective functions. A feasible solution $\hat{x} \in \mathcal{X}$ of the BOP (1) is efficient if and only if there does not exist any $x' \in \mathcal{X}$ with $(y_1(x'), y_2(x')) \leq (y_1(\hat{x}), y_2(\hat{x}))$. The image $y(\hat{x}) = (y_1(\hat{x}), y_2(\hat{x}))$ is called non-dominated. We will denote by \mathcal{X}_E the set of efficient solutions of (1) and by \mathcal{Y}_N its non-dominated image.

Let $G = (V, A)$ be a directed graph with a set of nodes or vertices $V = \{1, 2, \dots, n\}$ and a set of arcs $A \subseteq V \times V$ with $|A| = m$. Let (c_{ij}^k, d_{ij}^k) be the pair of unit flow costs on arc $(i, j) \in A$ for commodity k and let x_{ij}^k represent the amount of flow of commodity k going through arc $(i, j) \in A$. Furthermore, each arc has lower and upper bound capacities l_{ij} and u_{ij} with $l_{ij} \leq u_{ij}$, which are shared between the q commodities. The *BMCMCF* is defined by the following linear program

$$\begin{aligned} \min \quad & y(x) = \begin{cases} y_1(x) = \sum_{k=1, \dots, q} \sum_{(i,j) \in A} c_{ij}^k x_{ij}^k \\ y_2(x) = \sum_{k=1, \dots, q} \sum_{(i,j) \in A} d_{ij}^k x_{ij}^k \end{cases} \\ \text{s.t.} \quad & \sum_{\{j: (i,j) \in A\}} x_{ij}^k - \sum_{\{j: (j,i) \in A\}} x_{ji}^k = b_i^k, \quad k = 1, 2, \dots, q, \quad i = 1, 2, \dots, n \\ & l_{ij} \leq \sum_{k=1, 2, \dots, q} x_{ij}^k \leq u_{ij}, \quad \text{for all } (i, j) \in A \\ & x_{ij}^k \geq 0, \quad \text{for all } (i, j) \in A, \quad k = 1, 2, \dots, q. \end{aligned} \tag{2}$$

The first set of constraints represents flow conservation at the different nodes for all commodities. We assume that $\sum_{i=1, 2, \dots, n} b_i^k = 0$, $k = 1, 2, \dots, q$, otherwise the problem is infeasible. A value $b_i^k > 0$, $b_i^k < 0$, or $b_i^k = 0$, respectively, indicates that node i is a supply node, a demand node, or a transshipment node for commodity k . The second set of constraints ensures that for each arc total flow remains between

lower bound l_{ij} and upper bound u_{ij} . Because model (2) is a linear model, \mathcal{X} is a compact polyhedron and consequently \mathcal{Y} is also a compact polyhedron. Therefore, the images of all the efficient solutions lie on the boundary of \mathcal{Y} (Isermann 1974). These solutions are called supported efficient solutions and can be obtained by solving (single objective) weighted sum problems

$$\min_{x \in \mathcal{X}} \lambda_1 y_1(x) + \lambda_2 y_2(x)$$

for some $\lambda_1 > 0$, $\lambda_2 > 0$. The supported efficient solutions which define an extreme point of \mathcal{Y} are called extreme efficient solutions. By \mathcal{X}_{ex} and \mathcal{Y}_{ex} we denote the set of all extreme efficient solutions and non-dominated extreme points.

Sedeño-Noda, González-Martín, and Alonso-Rodríguez (2005) present a change of variables method for the *BU2CMCF* problem. Their method splits the problem into two bi-objective minimum cost flow problems with a single-commodity and they use the parametric network simplex method to solve these problems. Their method cannot be extended to more than two commodities. The literature on multi-objective minimum cost flow problems has been surveyed by Hamacher, Pedersen, and Ruzika (2007).

3 The *BMCMCF* problem and linear programs

The *BMCMCF* problem can be modeled as a linear program illustrated in model (2), where the objective functions as well as the feasible set are described by linear functions. This bi-objective linear program (*BOLP*) can be solved by several existing bi-objective linear programming algorithms. We use two methods to solve the *BMCMCF* problem, a bi-objective version of the parametric simplex method and the dichotomic method.

3.1 *BOLPs* and parametric simplex

We repeat the formulation used by Ehrgott (2005) for a *BOLP*

$$\begin{aligned} \min \quad & y = ((c^1)^T x, (c^2)^T x) \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{3}$$

where $c^1, c^2 \in \mathbb{R}^n$ are objective vectors. The feasible set in decision space is $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ defined by the $m \times n$ constraints matrix A and the right hand side vector $b \in \mathbb{R}^m$.

Theorem 1. (Isermann 1974) *A feasible solution $x^0 \in \mathcal{X}$ is an efficient solution of the *BOLP* (3) if and only if there exists some $\lambda \in \mathbb{R}_{>}^2$ such that*

$$\lambda^T \begin{pmatrix} (c^1)^T \\ (c^2)^T \end{pmatrix} x^0 \leq \lambda^T \begin{pmatrix} (c^1)^T \\ (c^2)^T \end{pmatrix} x \tag{4}$$

for all $x \in \mathcal{X}$.

From Theorem 1 we know that finding the efficient solutions of (3) is equivalent to solving the LP

$$\min y(\lambda) = \min\{\lambda_1(c^1)^T x + \lambda_2(c^2)^T x : Ax = b, x \geq 0\} \quad (5)$$

for all $\lambda \in \mathbb{R}_{>}^2$. Without loss of generality (dividing the objective function by $\lambda_1 + \lambda_2$) we can assume that $(\lambda_1, \lambda_2) = (\lambda, 1 - \lambda)$. We define the parametric objective function

$$c(\lambda) := \lambda c^1 + (1 - \lambda)c^2. \quad (6)$$

Thus we need to solve

$$\min y(\lambda) = \min\{c(\lambda)^T x : Ax = b, x \geq 0\}, \quad (7)$$

for all $0 \leq \lambda \leq 1$ which is a parametric linear program.

We can solve (7) using the simplex method by first solving it with $\lambda = 1$ and then iteratively finding entering variables with minimum ratio of deterioration of the first objective and improvement of the second objective. Once no more entering variables can be found the algorithm stops. Note that an optimal solution for $\lambda = 1$ may be weakly efficient. The algorithm may also find some solutions that do not define non-dominated extreme points. These solutions can be easily discarded at the end of the algorithm. This procedure is stated as Algorithm 1.

Algorithm 1 (Parametric simplex method for bi-objective LP's)

Input: Data A, b, c^1 and c^2 for a bi-objective LP.

Phase I: Let \mathcal{B} be an initial basis and $\mathcal{N} := \{1, \dots, n\} \setminus \mathcal{B}$.

Phase II: Solve the LP (7) for $\lambda = 1$ starting from basis \mathcal{B} found in Phase I yielding an optimal basis $\hat{\mathcal{B}}$.

Compute $\tilde{A} = A_{\hat{\mathcal{B}}}^{-1}A$, $\tilde{b} = A_{\hat{\mathcal{B}}}^{-1}b$, $(\tilde{c}^1)^T = (c^1)^T - (c_{\hat{\mathcal{B}}}^1)^T \tilde{A}$ and $(\tilde{c}^2)^T = (c^2)^T - (c_{\hat{\mathcal{B}}}^2)^T \tilde{A}$.

Phase III: While $I = \{i \in \mathcal{N} : \tilde{c}_i^2 < 0, \tilde{c}_i^1 \geq 0\} \neq \emptyset$.

$$\lambda := \max_{i \in I} \frac{-\tilde{c}_i^2}{\tilde{c}_i^1 - \tilde{c}_i^2}, \quad s \in \operatorname{argmax}\left\{\frac{-\tilde{c}_i^2}{\tilde{c}_i^1 - \tilde{c}_i^2} : i \in I\right\}, \quad r \in \operatorname{argmin}\left\{\frac{\tilde{b}_j}{A_{sj}}, \tilde{A}_{sj} > 0 : j \in \hat{\mathcal{B}}\right\}.$$

Let $\hat{\mathcal{B}} := (\hat{\mathcal{B}} \setminus \{r\}) \cup \{s\}$ and update \tilde{A} and \tilde{b} .

Discard the solutions which are not extreme supported efficient solutions.

Output: Sequence of λ -values and sequence of optimal basic feasible solutions.

3.2 Dichotomic approach

An alternative to the parametric approach is the dichotomic approach, first proposed by Cohon (1978). The method involves solving several single objective problems in weighted sum formulation (7). In this method, initial points $y(x^l)$ and $y(x^r)$ are obtained by solving LP (7) for $\lambda = 1$ and $\lambda = 0$, respectively. Next, weights are chosen to obtain a supported non-dominated point that has the maximal distance to the straight line connecting the two initial points $y(x^l)$ and $y(x^r)$. Whenever the image of the obtained efficient solution of such a problem does not lie on the line connecting the images of the two supported solutions $y(x^l)y(x^r)$, two new sub-problems can be formulated. Otherwise, there are no more extreme supported points between them, so the current sub-problem does not have to be split up further. We use the basis obtained in the previous iteration as starting solution for solving the new weighted sum problems. The dichotomic method stops if no new weighted

sum problems have to be solved, which means a complete set of extreme supported efficient solutions is obtained. Note that the initial solutions x^l and x^r for $\lambda = 1$ and $\lambda = 0$ may be weakly efficient and that the procedure may find some solutions that do not define non-dominated extreme points. These solutions can be easily discarded at the end of the algorithm. The dichotomic method is stated in Algorithm 2 and Algorithm 3.

Algorithm 2 (Dichotomic approach)

Input: Data A, b, c^1 and c^2 for a bi-objective LP.

Phase I: Let \mathcal{B} be an initial basis and $\mathcal{N} := \{1, \dots, n\} \setminus \mathcal{B}$.

Phase II: Solve the LP (7) for $\lambda = 1$ and $\lambda = 0$ from basis \mathcal{B} found in Phase I to find x^l and x^r and compute $y(x^l)$ and $y(x^r)$.

Phase III:

$$S = \{x^l, x^r\}, \text{SolveRecursion}(x^l, x^r, S), \mathcal{X}_{SE} = S.$$

Discard the solutions which are not extreme supported efficient solutions.

Output: Sequence of λ -values and sequence of optimal basic feasible solutions.

Algorithm 3 (Procedure SolveRecursion)

Input: x^l, x^r and S .

Phase I: Obtain \tilde{x} as an optimal solution of the LP (5) for $\lambda_1 = |y_2(x^l) - y_2(x^r)|$ and $\lambda_2 = |y_1(x^l) - y_1(x^r)|$ and compute $y(\hat{x})$.

Phase II: If $(y(\hat{x}) \cap \overline{y(x^l)y(x^r)}) = \emptyset$ then

$$S = S \cup \hat{x}, \text{SolveRecursion}(x^l, \hat{x}, S), \text{SolveRecursion}(\hat{x}, x^r, S).$$

Output: Sequence of λ -values and S .

4 Test networks

In this section, we discuss the performance of the parametric simplex method and the dichotomic method. We provide computational results obtained from several types of problems. All numerical tests are performed on a Microsoft Windows XP Professional Version 2002 Service Pack 3 computer with Intel (R) Xeon(R) CPU and 2.67 GHz, 3GB RAM. We use the OsiClpSolverInterface and open source optimisation solver COIN-OR (2012). The methods are implemented in C++. When measuring run-time, we disregard the time it takes to read the problem from a file and to write the solutions. In Sections 4.1, we use several types of directed bi-objective network instances with a single, two, three or five commodities. The application of the change of variables method to an undirected version of our bi-objective two-commodity instances is discussed in Section 4.2.

4.1 Directed bi-objective multi-commodity test instances

We investigate the performance of the methods on four sets of directed bi-objective multi-commodity test instances. For each set of instances we generate networks with one, two, three and five commodities. The first three groups of instances are the small and moderate size network instances with the same structure used by Raith and Ehrgott (2009). The first two groups are directed network instances generated by the NETGEN generator (Klingman, Napier, and Stutz 1974) which is

modified to include a second objective function and multiple commodities. Table 1 shows the number of nodes, arcs, sources and sinks, etc. which are set as NETGEN parameters to generate each set of networks. Problems N01 – N12 have varying sum of supply for each commodity ($\sum_{i \in \mathcal{V}: b_i^k > 0} b_i^k$) and problems F01 – F12 have fixed sum of supply for each commodity. There are 30 problems for each set of parameters. The third group of network instances consists of networks with a grid structure. In these networks, nodes are arranged in a rectangular grid with given parameters height h , width w , maximum cost c_{max} , maximum capacity u_{max} and sum of supply ($\sum_{i \in \mathcal{V}: b_i^k > 0} b_i^k$). All grid instances are listed in Table 2. Again there are 30 problems for each set of parameters. All of these examples are small and moderate size network instances. For investigating the performance of the methods with high density networks, we generate directed network examples with the same structure and parameters as the bi-objective undirected two-commodity examples used by Sedeño-Noda, González-Martín, and Alonso-Rodríguez (2005). We used the modified version of the NETGEN generator to include several commodities in these instances. Table 3 shows the number of nodes (n), the number of arcs (m) and the ratio (m/n). Using each combination of n and m we obtain 12 sets of network instances denoted L01 – L12. There are 30 examples for each set of parameters. In all of these examples, for all $k = 1, 2, \dots, q$ the node balances are set to $b_i^k = 10$ if $i = k$, $b_i^k = -10$ for $i = m - k$ and $b_i^k = 0$ otherwise. We set the maximum capacity of arcs (U) to 10, 20, 30 and 50 for all instances with one, two, three and five commodities respectively. We use positive integer costs generated by NETGEN for all objectives.

Name	Nodes	Arcs	Sources	Sinks	$\sum_{i \in \mathcal{V}: b_i^k > 0} b_i^k$	Transshipment sources	Transshipment sinks
N01/F01	20	60	9	7	90/100	4	3
N02/F02	20	80	9	7	90/100	4	3
N03/F03	20	100	9	7	90/100	4	3
N04/F04	40	120	18	14	180/100	9	7
N05/F05	40	160	18	14	180/100	9	7
N06/F06	40	200	18	14	180/100	9	7
N07/F07	60	180	27	21	270/100	14	10
N08/F08	60	240	27	21	270/100	14	10
N09/F09	60	300	27	21	270/100	14	10
N10/F10	80	240	35	38	350/100	17	14
N11/F11	80	320	35	38	350/100	17	14
N12/F12	80	400	35	38	350/100	17	14

Table 1: NETGEN test instances.

Name	h	w	Nodes	Arcs	c_{max}	u_{max}	$\sum_{i \in \mathcal{V}: b_i^k > 0} b_i^k$
G01	4	5	20	62	100	50	100
G02	5	8	40	134	100	50	100
G03	6	10	60	208	100	50	100
G04	8	10	80	284	100	50	100
G05	6	10	60	208	100	75	100
G06	6	10	60	208	100	100	100
G07	6	10	60	208	25	50	100
G08	6	10	60	208	50	50	100
G09	8	10	80	284	100	75	100
G10	8	10	80	284	100	100	100
G11	8	10	80	284	25	50	100
G12	8	10	80	284	50	50	100

Table 2: Grid test instances.

Nodes (n)	Arcs (m)	m/n
25	250, 375, 500	10, 15, 20
50	500, 750, 1000	10, 15, 20
75	750, 1125, 1500	10, 15, 20
100	1000, 1500, 2000	10, 15, 20

Table 3: High density test instances.

In Tables 4 – 6, the average number of non-dominated extreme points $|\mathcal{Y}_{ex}|$ as well as the average CPU time for the parametric (Pa) and dichotomic (Di) methods and for different numbers of commodities are presented. From these tables the following observations can be made:

- Both of the methods solve all of these instances in reasonable time, between 0.02 to 20 seconds.
- It can be seen that by increasing the number of commodities the CPU running-time increases significantly for both of the methods. This happens because the size of problem and the number of variables are proportional to the number of commodities.
- The number of non-dominated extreme points $|\mathcal{Y}_{ex}|$ and consequently the average CPU time increases by increasing the number of nodes (n) or by increasing the number of arcs (m).
- For the NETGEN and grid networks (Tables 4 – 5), the parametric method solves the problems in less CPU time than the dichotomic method. In the parametric method, in each iteration a new solution is obtained by entering a variable with minimum ratio of deterioration of the first objective and improvement of second objective into the basis. On the other hand to find a new solution in the dichotomic method, in each iteration a single objective weighted sum problem is resolved from the basis obtained from the pervious iteration. So in each iteration the number of operations required for the dichotomic method is significantly more than the number of operations needed for the parametric method.
- For the L01 – L12 instances, from Table 6 we can see that the number of non-dominated extreme points $|\mathcal{Y}_{ex}|$ for all of the instances is small and it stays almost constant (between 9 to 21) when increasing the number of nodes, arcs or commodities. This happens due to the fact that in all of these examples the amount of supply shipped through the networks is small.
- From Table 6 we can see that the dichotomic method solves the L01 – L12 problems in less CPU time than the parametric method. In these instances the number of arcs (variables) is large versus the number of non-dominated extreme points $|\mathcal{Y}_{ex}|$. Consequently in each iteration of the parametric method, obtaining the entering variable from the large number of variables becomes a time consuming process. So the dichotomic method performs better than the parametric method in these instances.

Name	1-commodity			2-commodity			3-commodity			5-commodity		
	\mathcal{Y}_{ex}	CPU time										
		Mean	Pa		Di	Mean		Pa	Di		Mean	Pa
N01	13.23	0.02	0.06	25.17	0.04	0.13	35.90	0.06	0.23	56.43	0.11	0.51
N02	16.10	0.03	0.07	32.47	0.05	0.19	45.43	0.08	0.34	72.97	0.15	0.79
N03	18.40	0.04	0.09	36.23	0.06	0.25	52.20	0.10	0.45	85.60	0.20	1.04
N04	27.97	0.05	0.16	53.30	0.11	0.49	71.90	0.18	0.88	119.80	0.39	2.17
N05	34.83	0.07	0.24	68.27	0.16	0.73	100.97	0.28	1.49	158.80	0.59	3.23
N06	43.40	0.08	0.34	80.40	0.20	0.99	116.87	0.37	1.91	189.27	0.83	4.42
N07	39.43	0.09	0.31	75.80	0.23	1.11	111.38	0.41	2.05	179.03	0.85	4.67
N08	54.17	0.14	0.58	104.37	0.35	1.74	155.00	0.64	3.31	246.40	1.42	7.47
N09	65.67	0.19	0.83	132.47	0.48	2.31	182.30	0.83	4.21	299.00	1.85	10.30
N10	55.17	0.15	0.63	106.63	0.42	1.96	152.87	0.75	3.68	242.03	1.55	8.72
N11	72.90	0.23	0.98	140.50	0.63	2.83	206.20	1.14	5.44	331.27	2.44	13.43
N12	85.60	0.31	1.31	169.37	0.83	3.88	248.33	1.60	7.77	398.03	3.31	18.12
F01	12.63	0.02	0.08	25.43	0.04	0.13	36.40	0.06	0.23	58.63	0.11	0.54
F02	18.67	0.03	0.12	34.10	0.05	0.20	46.97	0.08	0.34	74.47	0.15	0.81
F03	18.73	0.03	0.13	38.73	0.06	0.28	55.27	0.10	0.46	89.43	0.20	1.11
F04	22.43	0.05	0.19	43.20	0.10	0.40	62.69	0.17	0.75	102.38	0.38	1.87
F05	26.60	0.06	0.26	58.57	0.15	0.61	84.87	0.26	1.21	139.83	0.58	2.93
F06	34.53	0.08	0.39	65.57	0.18	0.82	97.50	0.34	1.60	160.87	0.80	3.94
F07	24.93	0.08	0.30	46.60	0.20	0.71	71.55	0.39	1.40	121.97	0.80	3.49
F08	36.83	0.13	0.57	71.67	0.33	1.24	105.73	0.60	2.32	177.60	1.39	5.61
F09	40.93	0.17	0.73	83.20	0.43	1.56	124.50	0.77	2.94	204.43	1.73	7.43
F10	22.70	0.14	0.39	46.07	0.33	0.92	71.87	0.61	1.90	120.39	1.32	4.76
F11	31.47	0.20	0.62	67.47	0.52	1.56	103.07	0.98	2.92	181.27	2.21	7.59
F12	39.60	0.29	0.90	86.27	0.75	2.19	132.70	1.41	4.30	220.03	3.07	10.69

Table 4: Results for directed NETGEN instances.

Name	1-commodity			2-commodity			3-commodity			5-commodity		
	\mathcal{Y}_{ex}	CPU time										
		Mean	Pa		Di	Mean		Pa	Di		Mean	Pa
G01	6.23	0.01	0.03	13.23	0.03	0.07	19.73	0.04	0.13	30.47	0.08	0.28
G02	13.00	0.03	0.08	26.87	0.08	0.25	37.40	0.13	0.46	59.77	0.29	1.10
G03	20.47	0.06	0.18	36.73	0.16	0.53	52.83	0.29	0.97	86.63	0.67	2.38
G04	23.73	0.10	0.29	45.90	0.28	0.88	67.10	0.51	1.65	113.17	1.22	4.19
G05	20.17	0.06	0.18	36.90	0.16	0.54	52.67	0.28	0.98	86.87	0.65	2.39
G06	20.20	0.06	0.17	36.03	0.16	0.51	52.47	0.28	0.97	87.57	0.67	2.40
G07	17.23	0.05	0.16	33.27	0.15	0.48	46.47	0.28	0.89	72.93	0.63	2.04
G08	19.00	0.05	0.17	35.63	0.15	0.51	52.67	0.28	0.98	83.63	0.65	2.31
G09	22.83	0.09	0.28	45.40	0.27	0.86	67.00	0.50	1.65	110.50	1.21	4.02
G10	25.77	0.10	0.32	44.70	0.27	0.84	67.23	0.53	1.64	110.53	1.23	4.05
G11	23.60	0.09	0.31	41.93	0.29	0.82	57.40	0.49	1.44	91.53	1.17	3.39
G12	22.83	0.10	0.29	44.07	0.27	0.85	63.43	0.50	1.57	104.23	1.21	3.82

Table 5: Results for directed grid instances.

Name	1-commodity			2-commodity			3-commodity			5-commodity		
	\mathcal{Y}_{ex}	CPU time										
		Mean	Pa		Di	Mean		Pa	Di		Mean	Pa
L01	11.50	0.06	0.11	9.07	0.11	0.14	8.27	0.18	0.18	10.80	0.39	0.33
L02	14.47	0.09	0.20	11.37	0.19	0.24	10.43	0.31	0.30	13.17	0.64	0.55
L03	17.53	0.14	0.30	12.03	0.26	0.33	11.97	0.42	0.44	14.83	0.92	0.81
L04	14.93	0.17	0.27	12.97	0.38	0.39	9.83	0.62	0.42	12.60	1.52	0.80
L05	17.07	0.30	0.44	14.80	0.63	0.64	14.03	1.18	0.83	13.90	1.98	1.23
L06	19.07	0.46	0.63	15.37	0.87	0.85	14.13	1.44	1.08	16.43	3.69	1.98
L07	15.77	0.36	0.43	12.60	0.78	0.57	12.07	1.37	0.80	13.23	2.60	1.29
L08	17.83	0.65	0.70	15.43	1.28	1.01	12.13	2.04	1.16	15.10	4.41	2.10
L09	18.53	0.93	0.95	18.17	2.04	1.53	14.60	3.33	1.76	16.70	9.33	3.14
L10	16.67	0.65	0.61	15.03	1.30	0.93	12.47	2.30	1.12	14.60	4.79	1.96
L11	18.73	1.08	0.98	15.97	2.29	1.47	11.70	3.60	1.53	16.07	10.85	3.24
L12	20.90	1.65	1.46	17.67	3.67	2.12	15.07	5.69	2.61	16.97	16.34	4.47

Table 6: Results for directed high density instances.

4.2 Un-directed bi-objective multi-commodity test instances

In this section, we use the change of variables method, presented by Sedeño-Noda, González-Martín, and Alonso-Rodríguez (2005), to solve the bi-objective undirected two-commodity network instances. In this method, each of the problems is split into two bi-objective undirected single-commodity subproblems which are solved with the parametric or the dichotomic method. By modifying the directed *B2CMCF* instances from Section 4.1, we obtain four groups of *BU2CMCF* examples. Network instances U-N01 to U-N12, U-F01 to U-F12 and U-G01 to U-G12 are *BU2CMCF* instances with the same structure as the networks explained in Tables 1 and 2. *BU2CMCF* instances U-L01 to U-L12 have the same structure as the networks explained in Table 3. In Tables 7 – 9, the average number of non-dominated extreme points $|\mathcal{Y}_{ex}|$ for the first sub-problems, second sub-problems and the main

problems are represented. The average CPU time for the parametric, dichotomic, change of variables parametric and change of variables dichotomic methods is shown in Tables 7 – 9.

From the tables and figures the following observations can be made:

- The number of non-dominated extreme points $|\mathcal{Y}_{ex}|$ and consequently the CPU time increases when using the undirected networks compared to their directed counterparts in Section 4.1.
- The parametric method performs better than the dichotomic method for these sets of instances.
- By using the change of variables method the CPU time of the parametric and the dichotomic method improves. Since all of the methods solve the small instances very fast this improvement in the CPU time is not immediately apparent for the small examples, but by increasing the size of instances the improvement of using the change of variables method becomes more obvious.

Name	$ \mathcal{Y}_{ex} $			CPU time			
	Total avg	Part1 avg	Part2 avg	Pa	Di	Change of variables	
						Pa	Di
U-N01	62.60	18.77	44.97	0.13	0.49	0.13	0.44
U-N02	93.27	30.70	63.87	0.19	0.88	0.18	0.74
U-N03	125.60	38.83	88.33	0.27	1.45	0.25	1.14
U-N04	142.57	45.77	98.47	0.46	2.20	0.46	1.80
U-N05	211.00	66.07	147.57	0.80	3.84	0.72	3.18
U-N06	278.67	87.70	194.30	1.17	5.97	1.12	5.04
U-N07	219.93	71.30	150.83	1.18	4.96	1.04	4.13
U-N08	342.60	108.00	238.47	2.08	9.50	1.75	7.39
U-N09	465.53	147.03	325.73	3.12	15.61	2.54	11.59
U-N10	311.10	208.61	209.63	2.22	9.37	1.92	7.39
U-N11	489.47	161.70	334.40	3.76	18.34	3.15	13.86
U-N12	653.13	205.03	459.50	6.10	29.86	4.57	22.11
U-F01	61.27	19.43	42.93	0.12	0.50	0.13	0.42
U-F02	90.03	30.23	61.00	0.18	0.87	0.18	0.71
U-F03	131.73	42.07	91.63	0.28	1.50	0.26	1.17
U-F04	130.90	40.57	92.03	0.41	1.97	0.42	1.59
U-F05	205.27	64.27	143.30	0.78	3.77	0.69	3.07
U-F06	274.73	85.97	192.33	1.18	6.01	1.11	4.82
U-F07	206.50	66.47	142.20	1.12	4.65	0.99	3.98
U-F08	325.43	100.67	228.30	2.02	8.97	1.68	6.93
U-F09	438.80	136.17	308.23	2.98	14.69	2.42	10.91
U-F10	276.63	185.41	187.17	2.04	7.84	1.72	6.43
U-F11	448.13	145.83	308.23	3.64	16.66	2.94	12.40
U-F12	627.83	217.03	420.90	5.95	27.55	4.47	20.75

Table 7: Results for undirected NETGEN instances.

Name	$ \mathcal{Y}_{ex} $			CPU time			
	Total avg	Part1 avg	Part2 avg	Pa	Di	Change of variables	
						Pa	Di
U-G01	44.83	7.40	38.50	0.09	0.36	0.09	0.30
U-G02	109.77	16.10	94.97	0.38	1.69	0.34	1.30
U-G03	176.60	26.57	151.57	0.99	4.29	0.80	3.18
U-G04	254.57	38.87	218.17	1.92	8.40	1.46	5.89
U-G05	180.57	26.13	156.17	0.99	4.33	0.79	3.23
U-G06	184.40	27.47	158.57	1.01	4.45	0.81	3.33
U-G07	147.43	23.83	129.90	0.96	3.66	0.76	2.81
U-G08	168.90	27.93	143.93	0.99	4.12	0.77	3.10
U-G09	251.63	36.87	217.33	1.91	8.37	1.45	5.86
U-G10	252.60	169.34	217.47	1.91	8.38	1.41	5.87
U-G11	201.17	35.60	174.83	1.85	6.77	1.37	4.91
U-G12	234.07	35.80	202.27	1.88	7.81	1.40	5.54

Table 8: Results for undirected grid instances.

Name	\mathcal{Y}_{ex}			CPU time			
	Total avg	Part1 avg	Part2 avg	Pa	Di	Change of variables	
						Pa	Di
U-L010	310.30	75.00	236.33	1.53	7.90	0.98	4.89
U-L02	484.23	110.67	374.63	3.13	17.44	2.30	11.45
U-L03	595.60	89.90	506.87	5.10	30.52	3.55	17.98
U-L04	712.37	170.23	543.37	7.35	38.27	4.79	21.64
U-L05	1081.50	253.23	829.60	15.09	78.19	9.95	49.94
U-L06	1476.20	361.70	1116.33	24.08	136.07	17.88	89.79
U-L07	1100.80	256.93	845.27	17.95	84.35	11.07	51.94
U-L08	1720.10	410.40	1311.93	35.25	183.83	24.92	116.47
U-L09	2331.90	566.80	1768.13	62.44	323.95	43.30	211.96
U-L10	1525.10	344.73	1181.80	31.34	150.08	20.55	95.63
U-L11	2376.47	562.37	1816.63	68.81	341.17	46.44	220.99
U-L12	3206.57	768.70	2442.30	120.33	603.85	67.06	373.57

Table 9: Results for undirected high density instances.

5 Conclusion

In this paper, we model the bi-objective multi-commodity minimum cost flow problem as a linear program. We solve the problem with two methods, the bi-objective parametric simplex method and the dichotomic method. We also implement a change of variables methods for solving the bi-objective undirected two-commodity minimum cost flow problem. In future research we will address combining research on single-objective multi-commodity flow problems, with algorithms for bi-objective single-commodity problems to make progress on the problem at hand. Furthermore, we will extend this research for problems with nonlinear objective functions.

References

- Ahuja, R.K., T.L. Magnanti, and J.B. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Englewood Cliffs: Prentice Hall.
- Cohon, J.L. 1978. *Multiobjective Programming and Planning*. Academic Press, New York.
- COIN-OR. 2012. Computational Infrastructure for Operations Research Home Page. <http://www.COIN-OR.org/>.
- Ehrgott, M. 2005. *Multicriteria Optimization*. Berlin/Heidelberg: Springer-Verlag.
- Hamacher, H.W., C.R. Pedersen, and S. Ruzika. 2007. “Multiple objective minimum cost flow problems: A review.” *European Journal of Operational Research* 176 (3): 1404–1422.
- Isermann, H. 1974. “Proper Efficiency and the Linear Vector Maximum Problem.” *Operations Research* 22 (1): 189–191.
- Klingman, D., A. Napier, and J. Stutz. 1974. “NETGEN: A Program for Generating Large Scale Capacitated Assignment, Transportation, and Minimum Cost Flow Network Problems.” *Management Science* 20 (5): 814–821.
- Raith, A., and M. Ehrgott. 2009. “A two-phase algorithm for the biobjective integer minimum cost flow problem.” *Computers & Operations Research* 36 (6): 1945–1954.
- Sedeño-Noda, A., C. González-Martín, and S. Alonso-Rodríguez. 2005. “The biobjective undirected two-commodity minimum cost flow problem.” *European Journal of Operational Research* 164 (1): 89–103.

Modelling Offshore Outsourcing of Software Testing Services: A Telecom New Zealand Case Study

Parvathy Muraleedharan¹ and Arun A. Elias²

¹MBA Student,

Victoria Business School,

Victoria University of Wellington, New Zealand

Phone: +64-278589958

Email: mparvathy@gmail.com

²Senior Lecturer

School of Management

Victoria University of Wellington, New Zealand

Phone: +64-4-4635736

Email: arun.elias@vuw.ac.nz

Abstract

Advancement in information technology and globalisation has witnessed a significant growth in offshore outsourcing of software services to countries that offer these services at a cheaper rate. Telecom New Zealand became a part of this growth with increasing number of outsourced projects and investments in outsourcing. But currently, this organisation is facing some challenges and risks, inhibiting its offshore outsourcing growth. This article analyses the risks involved in offshore outsourcing of testing services for organisations like Telecom New Zealand. Based on a set of ten semi structured interviews with client and service providers, the problem situation is structured systemically and a systems model developed to understand the various behaviours and trends that seemed to have an effect on IT offshore outsourcing at Telecom New Zealand. Based on the findings from this research, the study proposes two strategic interventions that identify opportunities to improve the communication practices between client and service provider and to increase service provider engagement. These interventions are expected to have a positive impact on the overall system and are expected to improve the relationship between the organisations.

Key words: Offshore Outsourcing, Systems Thinking and Modelling, Risk Management

1. Introduction

With the growth in globalisation and advancement in information and communication technologies, offshore outsourcing of IT work gained pace to seamlessly distribute work to different parts of the world with a primary intention to generate overall value (Lacity & Rottman, 2008). This value creation was expected to be realised by lowering costs, improving quality, increasing production, reducing delivery time and through risk sharing amongst client and vendor (Lacity & Rottman, 2008). The concept that any work that can be digitised could be globally sourced largely influenced the sourcing decision of IT offshore outsourcing (ITO) managers (Friedman, 2005). Based on the extent of offshore outsourcing carried out across organisations worldwide, researchers identifies six factors or enablers that influence growth of offshore outsourcing as rising wage differentials, growth of offshore labour pool, business friendly climate, globalisation in trade and services, software commoditization and drop in telecom costs (Carmel & Tija, 2005).

Yet, in this journey, organisations face significant challenges or risks that seem to inhibit their offshore outsourcing growth. Previous literature gives adequate insight into such risks faced by organisations and discusses several reasons as to why cost savings and quality of deliverables were hard to achieve. Some studies identify that such risks were distributed among clients as well as service providers at various stages of offshore outsourcing (Aron et al., 2005). While there seems to be ample information available describing client side risks, relatively fewer studies exposes service provider risks and a much lesser data is available to address both client and service provider risk perspectives together. Also, not many studies explain how these risks can be managed to build a value based client-service provider partnership by developing a holistic approach to improve relationship between client and service providers engaged in IT offshore outsourcing.

Telecom New Zealand is an organisation who has been engaged in IT offshore outsourcing work for nearly a decade. Although Telecom has been involved in outsourcing of their testing projects, very few cost reduction initiatives and operational improvements were realised in their journey. Access to skilled IT professionals and high end technology from service provider companies added to the attractiveness of the already perceived cost and quality benefits of offshore outsourcing venture. However, they found that even with a rising number of testing projects and proportional investment figures, service provider engagements in these projects were declining. Telecom is seeking to improve their relationship with service providers in anticipation of generating operational improvements and significant cost reductions.

Scholarly attention to offshore outsourcing practices from New Zealand context and testing as an offshore candidate has been relatively low compared to information available for other countries (Elias & Mathew, 2011). So, there is a perceived gap in literature that provided an opportunity to identify offshore outsourcing practices relevant to a New Zealand based organisations. This research will focus to build a 'holistic approach' specifically to improve 'relationship' between client and service providers engaged in offshore outsourcing of IT software 'testing'. Based on a systemic analysis of preliminary data and detailed information collected from interviews, a systems model is developed in this paper to propose strategies to improve the relationship between Telecom New Zealand and their service providers involved in offshore outsourcing of IT software testing services.

Particularly, this research seeks to understand the main risk factors faced by Telecom in their offshore outsourcing initiatives of IT software testing services and also

tries to identify the main risk factors faced by the main service providers of Telecom in their offshore outsourcing initiatives of IT software testing services.

2. Review of Literature

IT outsourcing is defined in literature as *significant contribution by external vendors towards physical and/or human resources associated with entire or specific components of IT infrastructure in the user organisation* (Loh & Venkatraman, 1992), whereas IT offshore outsourcing is defined as *specific type of outsourcing where firm contracts for services with external vendors located in a different country* (Poston et al., 2010). In that sense, IT Offshore Outsourcing is an offspring of outsourcing, which has its history rooted to early 1960s (Weber, 2004).

The interest towards IT outsourcing intensified after 1989 when Eastman Kodak signed a 500 million contract with IBM to build and operate a data centre for Eastman Kodak (Loh & Venkatraman, 1992). ‘Kodak-effect’ marked a critical event in the history of IT outsourcing and provided visibility into administrative practices required for such successful IS management (Lacity & Hirschheim, 1993). Thereafter, more companies began to join the “outsourcing bandwagon” to lower costs, increase service levels and attain flexible IT management (Lacity & Hirschheim, 1995).

While early forms of outsourcing were focussed only to a small part of the budget, over the last two decades, outsourcing evolved to span across multiple systems and represent a large part of the revenue for the company. Previous studies provide a strong research base for IT sourcing decisions (Aron *et al.*, 2005), describe prescriptive frameworks based on technological, economic and business factors (Ranganathan & Balaji, 2007; Lacity & Wilcocks, 1998), provide guidelines to make best use of outsourcing market (Lacity *et al.*, 1995,1998) and also discuss emerging market trends such as selective outsourcing, flexible contracts, strategic alliances etc. (Lacity & Wilcocks, 1998).

One of the dominant perspectives discussed in most of the outsourcing literature is about Transaction Cost Theory (TCT) coined by Williamson (1975). TCT describes that firm’s economising nature on transaction cost and researchers have used this framework as a foundation to address the cost efficiency attributes of contractual agreements. Lacity and Wilcocks (1998) used an interpretive lens to critically develop upon TCT and developed a heuristics for actual cost savings versus expected cost savings constructs relevant to outsourcing relationships. Carmel and Tija (2005) based their construct on wage differentials and captured the notion of transaction cost in ITO with respect to Total cost of Offshoring and Total cost of engagement, which they believed is useful to determine Total Savings of Offshore Strategy (TSoS).

Our review also found that researchers used models to understand various aspects of outsourcing. For example, Ranganathan and Balaji (2007) identified a set of critical capabilities and formulated a framework consisting of four key categories that accommodates 10 critical capabilities. The four key categories are (i) Systemic thinking on offshore sourcing (2) Global IS vendor management (3) Global IS Resource Management (4) IS Change Management

The goal of the framework is to provide evaluation criteria to IS managers so that they can gauge the level of offshore outsourcing capabilities and improve firm’s performance both at macro and micro levels. However, this framework lays more emphasis on client perspective than vendor perspective and doesn’t entirely provide an opportunity to build a value based partnership for both the parties.

An offshore stage model developed by Carmel and Tija (2005) takes a companywide view to assess the relative degree of maturity and sophistication for an organisation undertaking IT offshoring initiatives. The model identifies four main stages in offshore development that are characterised by strategic imperatives and organisation's internal dynamics: (1) Offshore Bystander – where organisations watch other companies and study various aspects of offshoring; (2) Experimental Stage – where organisations start to explore their offshore development; (3) Cost Strategy Stage – where there is an increased cost savings from offshoring and organisations invest more in offshoring; (4) Leveraging Offshore – organisations focus on value addition and continuous improvement from offshoring.

It provides a good starting point to determine the firm's position in the offshore market. The progressive nature of the model by itself acts as a motivational tool to drive performance. The model is framed using a linear approach, whereas in reality, such a sequential approach may be difficult to achieve.

In addition, Simchi-Levi *et al.* (2007), developed a 'Make versus Buy' model, categorising the fundamental reason for outsourcing into the following: (i) dependency on knowledge (ii) dependency on capacity (iii) modular product and (iv) integral products. The framework helps to assess independent tasks, components and sub systems separately, i.e. they were able to split the offshore decisions to smaller chunks and address them separately. The application of the model to IT offshoring was very limited compared to other models related to ITO sourcing decisions (Aron *et al.*, 2005).

We also found a Spiderweb chart or a Radar chart that can serve the purpose to be a focal point for an organisation to conduct an internal assessment of its offshore readiness and to conduct a risk assessment with stakeholders (Carmel & Tija, 2005). Spiderweb charts are simple, intuitive and easy to understand and reveal the strengths and weaknesses of an organisation at a glance. However, the choice of variables or performance indicators and dimensions are largely subjective. The chart does not represent quantitative data hence ratings are relative and factors can vary with perception of people.

The promising side of IT offshore outsourcing is however crippled by a set of delimiters, generally described as risks in IT offshore outsourcing that account for the slow pace of the offshoring growth. More than 18 conceptual papers have been published about ITO risks elements (Lacity *et al.*, 2008) detailing the various risk elements based on client perspectives (Aron *et al.*, 2005; Aundhe & Mathew, 2009). Academic research on IT offshoring risks from a service provider perspective is very limited and only few researchers have identified risk elements from vendor perspective (Aundhe & Mathew, 2009). Much of their research was focussed on application development and US-India country context.

Significant information about offshore outsourcing models and practices for various countries can be obtained from literature. However, compared to this, only limited information was available for a New Zealand client perspective. Therefore, this research aims to address that gap by presenting a case study that encapsulates the offshore outsourcing relationship between NZ client and Indian vendors engaged in software testing projects.

3. Methodology

This research uses a systems thinking approach (Senge, 1990) to study the relationship between Telecom and their service providers. This approach provided a good

foundation to address the complexities, interdependencies and risks involved in offshore outsourcing for client and service provider organisations (Maani & Cavana, 2007).

The methodological framework used in this research is described in two phases: Problem Structuring and Causal Loop Modelling. In Phase 1, the problem situation was structured methodically and a 'behaviour over time graph' was prepared to capture the problem situation. The main factors identified from phase 1 were inputs to phase 2, where a casual loop model was developed.

For the purpose of data collection, 10 semi structured interviews were conducted with 5 representatives from Telecom and 5 representatives from service provider organisations for an average duration of 30 minutes. This comprised of face to face and telephonic interviews. A set of questionnaire was prepared along with some open ended questions to enquire about various elements relevant to this research such as motives for offshore outsourcing testing activities, comparison of offshore outsourcing and outsourcing, benefits and risks involved in offshore outsourcing and finally improvements and suggestion to improve relationship between client and service provider organisations. Interviews were recorded using a voice recorder and later transcribed for the purpose of this report. The organisations selected for the interview were based on: (1) organisation's involvement in IT software testing ; (2) offshore outsourcing practices and organisation's involvement in outsourcing and offshore outsourcing of testing projects with various organisations around the world. The interview participants were selected from both client and service provider organisations based on their experience working in Telecom's IT outsourced and offshore outsourced testing projects.

3.1 Phase 1: Problem Structuring using a Behaviour Over Time Graph

In phase 1, the problem situation was defined systematically into collection of preliminary data, preparation of behaviour over time graphs and collection of detailed interview data. Preliminary data was collected from senior managers and key decision makers involved in Telecom testing services. Information on number of outsourced testing projects, Telecom's investments in outsourcing and offshore outsourcing initiatives and number of successful projects engaged by the vendor were collected.

Later, the preliminary data collected was used to prepare 'Behaviour Over Time' (BOT) graph (Refer Figure 1). This provided a good starting point to address the objective reality and to capture the main problem and influencing factors that were affecting the behaviour of outsourcing and offshore outsourcing work at Telecom. BOT graph captures a behavioural trend over a period of time and identifies the directions, variations and trends in the variable of interest such as growth, decline, oscillations, or a combination. In BOT graph, performance indicator is plotted along vertical axis and time in months/years is plotted along horizontal axis. They are only drawn to a rough sense and is not plotted against numerical values offering the flexibility to represent several variable to one graph (Maani & Cavana, 2007).

It was observed from the BOT graph that number of outsourced projects at Telecom was increasing along with their investments for outsourcing/offshore outsourcing testing activities. However, counter-intuitively, the engagement rate with service providers were declining as the number of successful projects handles by the service providers were decreasing. This is represented by the declining behaviour in the BOT graph.

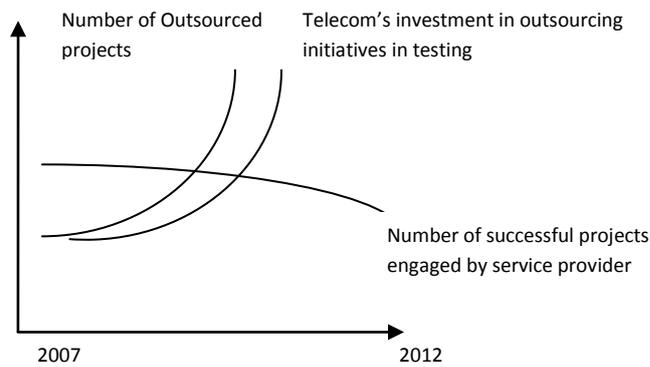


Figure 1 : BOT graph

3.2 Phase 2: Causal Loop Model

Based on the problem structure and variables identified from interview data, a causal loop model was developed to understand the factors that influence the offshore outsourcing growth at Telecom (Refer Figure 2). A causal loop diagram explains how “structure determines behaviour” and helps us to identify behavioural patterns that help us to address the overall structural problem (Maani & Cavana, 2007).

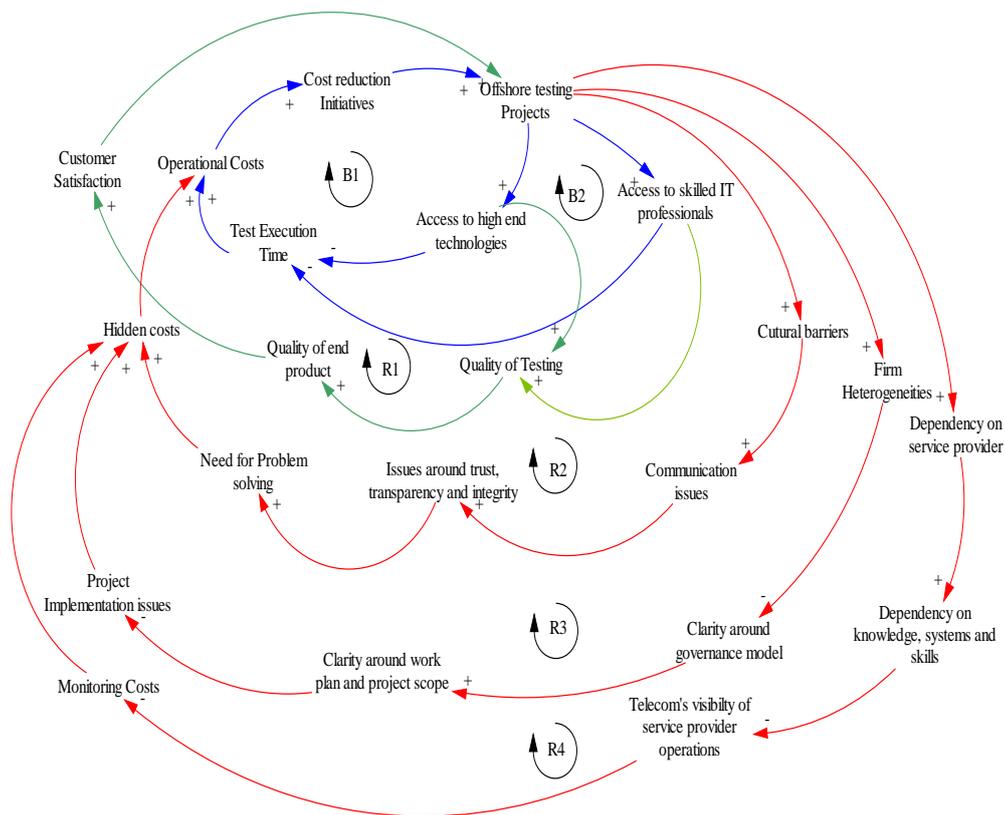


Figure 2 : Causal Loop Model

Telecom has witnessed that with a growing number of testing projects, investments in testing had started to increase proportionally. Operational costs involved in testing activities were beginning to rise branding testing as an expensive affair to the organisation. Various cost reduction initiatives and ‘cost cutting measures’ started to gain pace. Cost benefits of offshore outsourcing drew the attention of IS decision makers. Telecom also anticipated the benefits of gaining access to high end access to technology and access to highly skilled IT professionals from Indian companies to do

their testing work and expected these factors to reduce the current test execution time and the overall operational costs (Loops B1 and B2). There are an 'odd' number of negative polarities in Offshore Attractiveness Loops 1 and 2 making it negative feedback or a Balancing loops.

Based on systems literature (Maani & Cavana, 2007), it can be inferred that, when operational cost reduces, the 'goal seeking nature' compensates by increasing the initiatives to offshore outsource more projects as an 'ideal cost cutting measure' for Telecom.

While cost reduction might have been one motive for offshore outsourcing, access to skilled labour pool and high end technologies were also estimated to increase the quality of testing which and quality of end product. Rise in quality will increase the overall customer satisfaction, once again making offshore outsourcing a preferable option for cost reduction (Loop R1 - 'Quality Benefits Loop'). All the positive polarity links indicate R1 is positive feedback loop or a Reinforcing Loop. Now, if were to consider only client side expectations with no external impacts, then offshore outsourcing would have continued in this direction to achieve the goal of cost reduction. But in reality, the analysis revealed that emerging service provider behaviour patterns have influenced the underlying dynamics of offshore outsourcing in the long run. This resulted in three main feedbacks loops which are discussed in the next section.

3.2.1 Discussion of feedback loops

Based on the data collected from interviews, three feedback loops appear to have emerged after a period of time and they are as follows:

Feedback Loop 1: Cultural and Communication Barrier loop (Loop R2 in the diagram)

With an offset of Offshore Outsourcing of testing projects, the client and service provider companies witnessed an increase in cultural barriers between their employees. When cultural barriers increased, communication issues increased leading to lack of integrity, trust and transparency. Issues around bounded rationality and incomplete information from service providers increased the need for problem solving for Telecom. Telecom had difficulties to estimate these additional expenses that were hard to audit. This eventually led to rising hidden costs and operational costs, fading the overall cost benefit of Offshore outsourcing in the long run. This is a Reinforcing Loop as the overall loop consists of positive polarities only (Loop R2), but the resultant behaviour of this loop is for the number of successful projects engaged by the service provider to decline.

Feedback Loop 2: Firm Heterogeneities loop (Loop R3 in the diagram)

The client and service provider organisations follow different organisational culture, processes and routines and there was a gap prevailing in the organisations to understand each other's management structure and work patterns. Telecom representatives felt that their organisational structure and processes were not well understood by service providers and that these service providers often implemented processes that didn't work well at Telecom. Such failed attempts from the service provider started to increase operational costs. Gradually, Telecom began to lose confidence in service providers and the current contractual relationship began to deteriorate.

Service provider representatives' concerns around firm heterogeneities were that Telecom's organisational restructure and changing policies had introduced additional cost. In that sense, both the organisations witnessed that when firm heterogeneities

increase, there is an increased need for clarity around governance model. When roles and responsibilities were not clearly stated, distribution of work between the organisations became difficult. Telecom representatives identified that there were no clear resourcing strategies and claimed that service providers often provided staff who were less competent to work with Telecom's systems and software. Service providers, on the other hand, claimed that they often received very little information and a clear problem statement from Telecom management. Such factors reduced the clarity around work plan and project scope. Therefore, when clarity around governance model decreased, clarity around work plan and project scope also decreased.

When clarity of work plan reduced, service providers were sceptical about contract extension. They made fewer investments towards training and up skilling of their resources. With this the project implementation issues had started to increase and Telecom had to make additional investments towards training service provider representatives. This led to an increased transaction costs and resource contention that eventually increased the hidden cost of offshore testing.

There are even number of negative elements in the loop R3 which means this is a reinforcing loop. Once again the hidden costs and operational costs are increasing here, therefore the effect of this loop reduces the attractiveness of offshore outsourcing.

Feedback Loop 3: Dependency on Service Provider Loop

Offshore outsourcing of testing reaped quick wins to reduce operational cost but over time Telecom began to rely more on service providers for solutions and their resources. While the focus shifted to final output or deliverables from service providers, Telecom started to lose visibility around service providers' efforts. There were no common knowledge repositories and there were no up skilling for Telecom's staff. Therefore, when dependency on service providers increased, dependency on their skilled resources, systems and tools also increased. But visibility of service provider operation decreased during this period. When Telecom could not see the level of effort required and monitor the tasks executed by service provider, reviewing progress and daily work became difficult. As a result, monitoring cost and hidden cost increased.

This is a reinforcing loop. The effect of this loop contributes to the increasing trend of Telecom's investment in outsourcing initiatives of testing projects.

4. Conclusions

In this research, we explored few offshore outsourcing practices from earlier literature, gathered and analysed data collected from interviews and developed a systems model to explain the problem situation using simple cause-effect relationship (Maani & Cavana, 2007). This model has made an attempt to explain key strategies behind Telecom's decision for offshore outsourcing testing activities and feedback loops reveal that some emergent behaviour have affected the dynamics of offshore outsourcing growth. These feedback loops explain that short term fixes (e.g. various cost reduction initiatives) may not always be a remedy to address the primary symptom of the problem (reduce operational cost) in the long run. To induce long term changes in the behaviour, long term structural changes need to be devised (Senge, 1990). Therefore, a few improvements and suggestions in the form of two strategic interventions are proposed here to induce long term structural changes to change the behaviour of the system.

The first intervention develops strategies around relationship governance model and draws a set of communication practices required to distribute and manage work between

client and service providers. Primarily, this strategy focuses on two key aspects: Developing a resource strategy and developing a training plan. A sound resource strategy helps to identify, acquire and develop a team with a good resource mix from client and service provider side locations. Service providers could also engage Telecom managers in resource selection to ensure people with the right skills are chosen. Similarly, a training plan could be developed at the planning phase of the project to ensure informational flows are bidirectional and these knowledge and training materials are captured into central repositories. These artefacts are likely to increase clarity of work making it is easier for distributed work management.

The second strategic intervention discusses improvement measures to reduce hidden costs and increase vendor performance, with the intention to identify mechanisms to effectively increase service provider engagement.

At a micro level, Telecom could undertake an internal assessment to determine its internal capabilities, processes and systems. An assessment using Spiderweb chart/Radar chart can be a good starting point (Carmel & Tija, 2005) to evaluate the offshore readiness of the organisation against factors such as: maturity of project management, existing capacity and knowledge, complexity of project, IT infrastructure dependencies etc. This intervention also suggests an early engagement of service providers in decision making process will provide them sufficient clarity towards upfront work plan and this information could be utilised to prepare statement of work and resourcing strategy for Telecom. At a macro level, Telecom could consider setting up a Vendor Management office with an in-house team to review and manage vendor progress. Telecom could include additional provisions in the service level agreements to accommodate elements such as: management of IP risks, quality checks and measures for monitoring performance, clauses for distributed work management, provisions for contract flexibility to accommodate additional efforts, provisions to measure continuous improvements, etc.

Vendor audits are likely to open opportunities for international bidding where Telecom can project their expectations to multitude of service providers worldwide attracting competitive offers.

The first intervention will address the negative reinforcing loop, R2 where now, fewer communication issues will reduce issues around trust and transparency and therefore leading to lowers hidden costs and operational costs. The second intervention will influence negative reinforcing loops, R3 and R4, where minimising firm heterogeneities and dependencies on service provider will eventually reduce hidden costs and operational costs. To summarise, these strategic interventions can introduce long term changes to the overall offshore outsourcing structure and can influence the behaviour of the system to improve the relationship between client and vendors and once again increase the attractiveness of offshore outsourcing.

In this research, an effort has been made to construct a causal loop model to evaluate IT offshore outsourcing practices at Telecom New Zealand. The feedback loops explained in the systems model helps to capture the risk factors involved in offshore outsourcing and the strategic interventions proposed in this paper helps to mitigate these risks, thereby fostering a value based partnership between client and vendor.

5. References

- Aron, R., E. Clemons and S. Reddi. 2005. "Just right outsourcing: understanding and managing risk." *Journal of Management Information Systems* 22(2): 37-55.
- Aundhe, M.D. and S. Mathew. 2009. Risks in offshore outsourcing: a service provider perspective. *European Management Journal* 27: 418-428.
- Carmel, E. and P. Tija. 2005. *Offshoring Information Technology*. Cambridge: Cambridge University Press.
- Elias, A.A. and S.K. Mathew. 2011. Offshore IT outsourcing between India and New Zealand: a systemic analysis. Proceedings of the 25th Annual Australian and New Zealand Academy of Management (ANZAM) Conference, 7-9 December 2011, Wellington, ISBN: 978-1-877040-87-0, pp. 1-18.
- Friedman, T.L. 2005. *The World Is Flat: A Brief History of the Twenty-first Century*. New York: Farrar, Straus and Giroux.
- Lacity, M. and R. Hirschheim. 1995. *The Information Systems Outsourcing Bandwagon*. Chichester: Wiley.
- Lacity, M.P. and J.W. Rottman. 2008. *Offshore Outsourcing of IT Work: Client and Vendor Perspectives*. New York: Palgrave Macmillan.
- Lacity, M.C. and L.P. Willcocks. 1998. "An empirical investigation of information technology sourcing practices: lessons from experience." *MIS Quarterly* 22(3): 363-408.
- Lee, J. and Y. Kim. 1999. "Effect of partnership quality in IS outsourcing success: conceptual framework and empirical investigation." *Journal of Management Information Systems* 15(4): 29-61.
- Loh, L. and N. Venkatraman. 1992. "Diffusion of information technology outsourcing: influence sources and Kodak effect." *Information Systems Research* 3(4): 334-338.
- Maani, K.E. and R.Y. Cavana. 2007. *Systems Thinking, System Dynamics: Managing Change and Complexity*, 2nd edition. Auckland: Pearson Education New Zealand.
- Poston, R.S., J.C. Simon and R. Jain. 2010. "Client communication practices in managing relationships with offshore vendors of software testing projects." *Communications of the Association for Information Systems* 27(9): 129-148.
- Ranganthan, C. and S. Balaji. 2007. "Critical capabilities for offshore outsourcing of information systems." *MIS Quarterly Executive* 6(3): 147-164.
- Senge, P. 1990. *The Fifth Discipline: The Art and Science of the Learning Organization*. New York: Currency Doubleday.
- Simchi-Levi, D., P. Kaminsky and E. Simchi-Levi. 2007. *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies* 3rd edition, McGraw-Hill: New York.
- Weber, R. 2004. "Editor's Comments: Some implications of the Year-2000 era, dot-com era, and offshoring for information systems pedagogy." *MIS Quarterly*, 28(2): pp. iii-xi.
- Williamson, O. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications. A Study in the Economics of Internal Organization*. New York: Free Press.

A Qualitative System Dynamics Analysis of Airline Safety in New Zealand

Ong Su-Wuen, Robert Y. Cavana, Mondher Sahli
Victoria Business School, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand
su-wuen.ong@corrections.govt.nz

Abstract

This paper presents a preliminary analysis of airline safety in New Zealand using the system thinking tools of qualitative system dynamics.

Evidence is presented of the dynamic nature of airline safety. Based on the works of Rose and Rhoades et al, we theorise that different commercial pressures lead airlines to different safety responses. We start by adapting relevant aspects of Cooke's coal mine safety model, Salge and Milling's airline commercial model and Moizer's generic occupational safety model.

Using Phillips and Talley, we include aircraft characteristics, flight crew, weather conditions and airport conditions as important factors. Finally we include payrates and other commercial factors, based on Rhoades and Waguespack, and Wilson.

Combining all these elements yields our causal loop diagram. A number of the important feedback loops will be discussed. Finally some concluding comments will be provided about the value of qualitative system dynamics in theory building for airline safety.

Key words: Systems Thinking, Qualitative System Dynamics, Airline Safety, Causal Loop Diagram.

1 The New Zealand commercial aviation safety scene

As at 17 November 2012, the New Zealand safety regulator, the Civil Aviation Authority (CAA) has 1982 fixed wing and 792 helicopters on its registry (CAA, 2012).

There were 180 organisations licensed to carry fee-paying passengers or freight, called Part 119 operators. There were also 30 "adventure aviation operators" and 102 agricultural aircraft operators.

Many of the 180 "Air Operators" probably do not perform many passenger carrying flights. For the calendar year 2004, four operators only flew one such flight. At the other end of the scale, three operators each flew over 50,000 passenger flights each.

The CAA keeps counts of two measures that are directly safety-related. The first is a count of accidents. These are aircraft-related occurrences where the aircraft was damaged, gone missing or humans were seriously injured. The second is a count of 'incidents'. These are occurrences where safety was, or could have been, affected.

The reporting of both accidents and incidents is mandatory in New Zealand. The definition of accidents is very clear, so there is little leeway for not reporting an accident. However, the definition of an incident is relatively loose and open to interpretation. It is possible that the database does not capture the vast majority of

incidents. Fig. 1 shows the number of commercial flight accidents between 1995 to 2011 (Large aeroplanes greater than 13,608 kg; medium aeroplanes between 5,670 and 13,608 kg; small aeroplanes less than 5,670 kg).

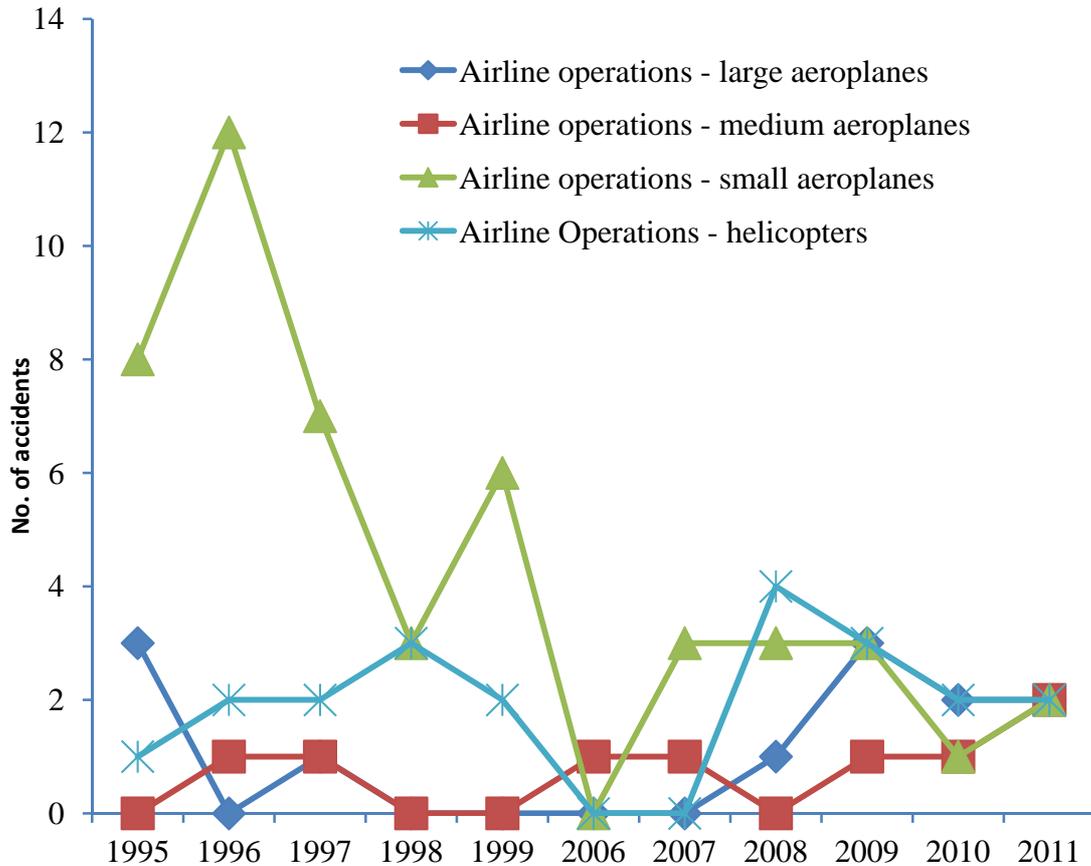


Figure 1. Number of all commercial flight accidents, from CAA website (17 Nov 2012)

2 Developing a model of airline safety

At an organisational level, safety is a dynamic issue with a tension between profitability and safety. This tension is nicely illustrated by Reason (1997) who points out that his safety model is a dynamic system – time being as the line within the graph (see Fig. 2).

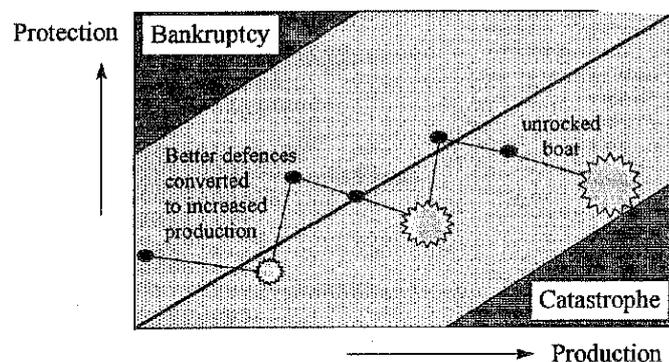


Figure 2. How a ‘catastrophes’ occur in companies, from Reason (1997, Fig 1.3, p5)

Cooke (2003) published a system dynamics coal mine safety model which was based on Sterman’s (2000) inventory control and order fulfilment archetype. Cooke (2003) had 4 distinct sub-systems –Human Resources, Production, Mine Capacity, and Safety (see Fig. 3). We used Cooke’s (2003) model as our primary source when building our model. The other main models we used as a source of ideas and inspiration were by Salge and Milling (2004) and Moizer (1999).

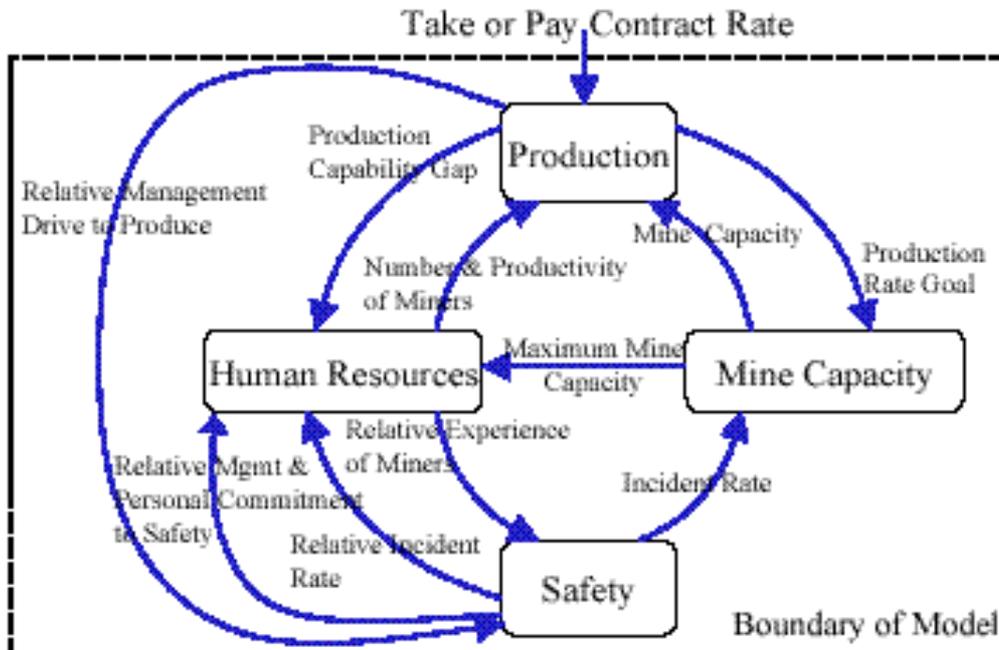


Figure 3. Subsystems of the Westray Mine Model, from Cooke (2003, Fig. 1, p144)

We combined the Production and Mine Capacity sub-systems of Cooke’s model into a ‘Business Operations’ sub-system. This would be the part of the model that simulates the commercial operations side of an airline or air operator. The modified subsystem view of the model is shown in Fig. 4.

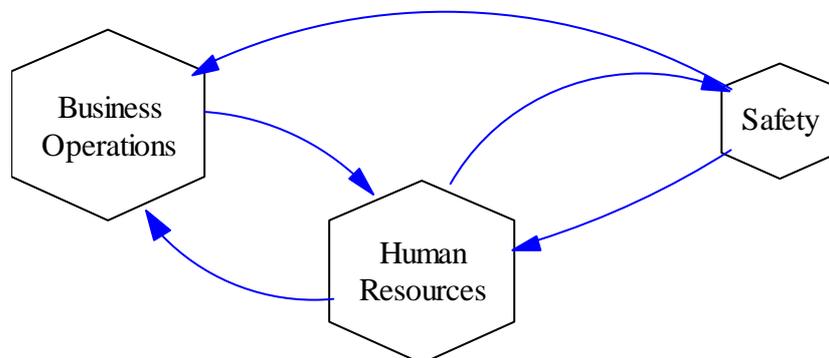


Figure 4. Overview diagram for an Airline Safety model

In the rest of this paper, we develop the causal loop diagram (CLD) model for airline safety, using the system dynamics methods outlined, for example, in Forrester (1961), Richardson & Pugh (1981), Coyle (1996), Sterman (2000) or Maani & Cavana (2007).

3 Causal connections in the Business Operations subsystem

We began constructing the causal loop diagram by starting at the Business Operations subsystem. The causal loop diagram of Salge and Milling's (2004) model of airline business, in Fig. 5, was a good starting point in building our CLD.

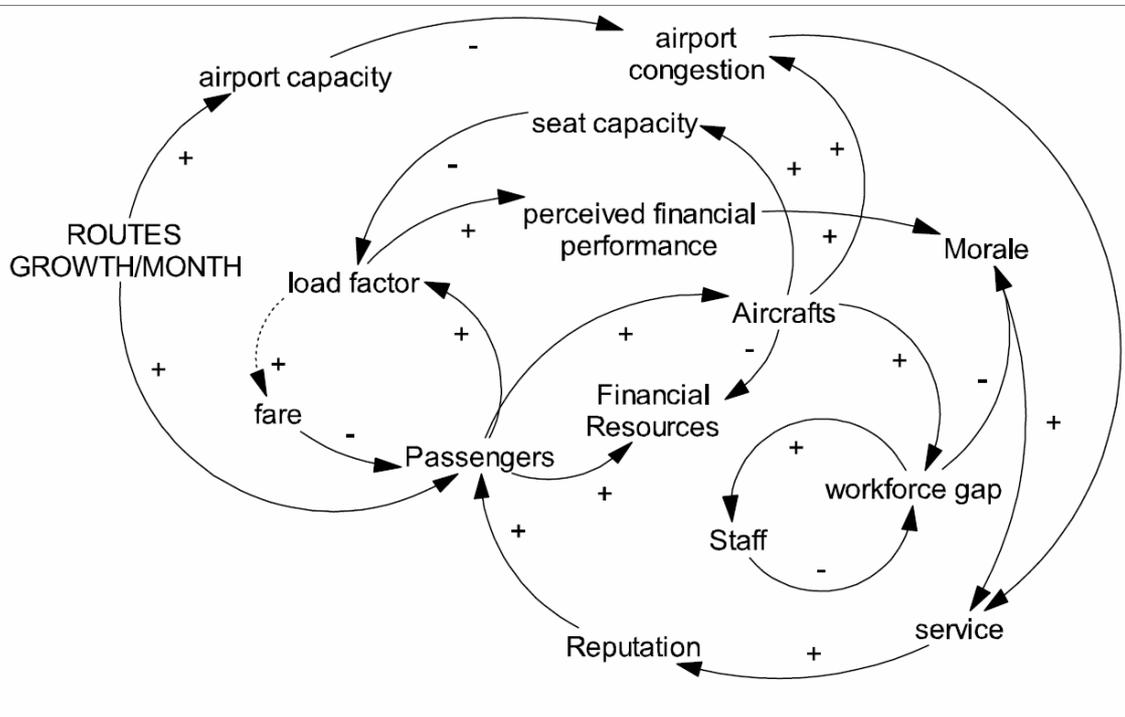


Figure 5. Salge and Milling's (2004, Fig 3, p7) airline operations causal loop diagram

Starting with 'aircraft', Salge and Milling (2004) have 'aircraft' affecting 'financial resources'. We use that but introduce 'maintenance expenditure' to highlight how aircraft affect an airline's cash balance.

Similarly, Salge and Milling (2004) have 'passengers' affecting 'financial resources'. Again we introduce an intermediate variable of 'revenue' as a clarification.

Fig. 6 shows these connections, which form the basis of the business operations subsystem.

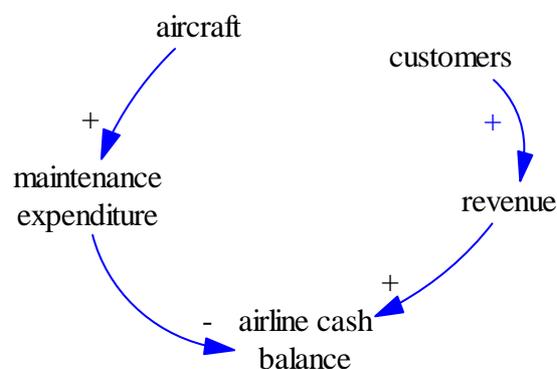


Figure 6. Initial causal connections in Business Operations subsystem

4 Causal connections in the Safety subsystem

We began by adapting Moizer's (1999) model into an airline safety model (see Fig. 7). Moizer (1999) makes a connection between 'accidents' and 'costs'. Direct cost of a crash is mostly covered by insurance, according to Reason (1997) and Rose (1992). So we introduced 'insurance' as an intermediate step.

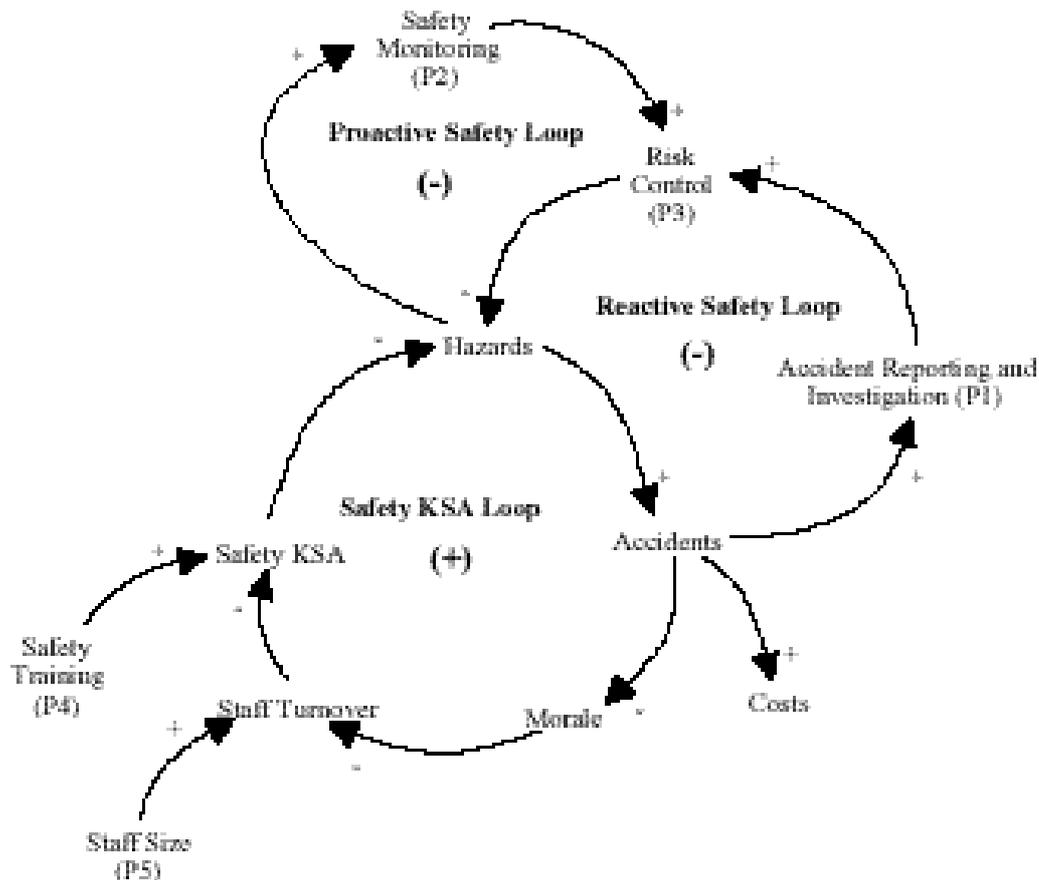


Figure 7. Safety causal loop diagram from Moizer (1999, Fig 5.3, p128)

Accidents could alter client demand for that airline. Salge and Milling's (2004) model shows that 'service' affects 'reputation' which then affects 'passengers'. For a safety model, accidents affect 'reputation' which then affects 'customers'.

What affects 'accidents'? Rhoades et al (2005) conjectured that fleet mix, fleet age, aircraft utilization and maintenance training could have an effect on safety. We used the terms "aircraft suitability" and "crew ability" to cover these factors. Phillips and Talley (1992) also have aircraft and crew characteristics and introduce weather and airport conditions.

None of the cited articles specifically mention the role of the safety regulator. However it is obvious that a greater number of accidents would lead to more oversight activity by the regulator. We thus incorporate a positive connection between accidents and oversight by the regulator. Fig. 8 shows the connections discussed in the safety subsystem.

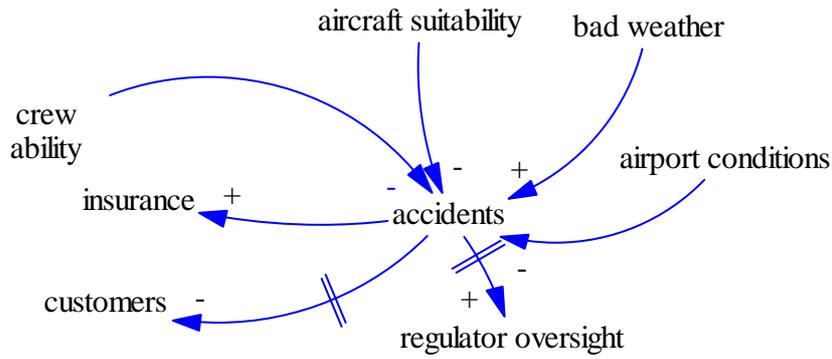


Figure 8. Initial causal connections in the Safety subsystem

5 Causal connections in the Human Resources subsystem

Crew characteristics (flight crew and maintenance) of relevance to safety are the experience and the training of the personnel. Training can be provided by the company itself. Experience can only be ‘procured’ by employing experienced pilots, or retaining pilots long enough to gain sufficient experience. In both cases, payrates are a major factor in the retention and recruitment of experienced staff (Cavana, et al., 2007).

Wilson (1997) wrote about how the US regulator was concerned about Valujet’s payrates. Rhoades and Waguespack (2000) go further and directly associate the lower pay of regional carriers compared to major airlines as a reason for the worse safety record of regional carriers. This is shown in Fig. 9.

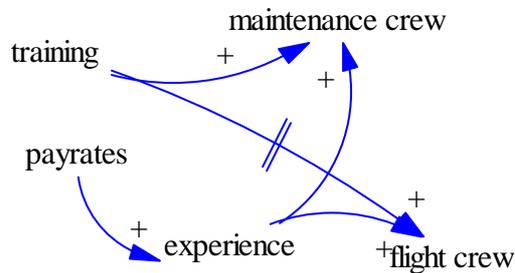


Figure 9. Initial connections in the Human Resources subsystem

6 Completing the airline safety causal loop diagram

The three subsystems are connected via the common variables, as shown in Fig. 10.

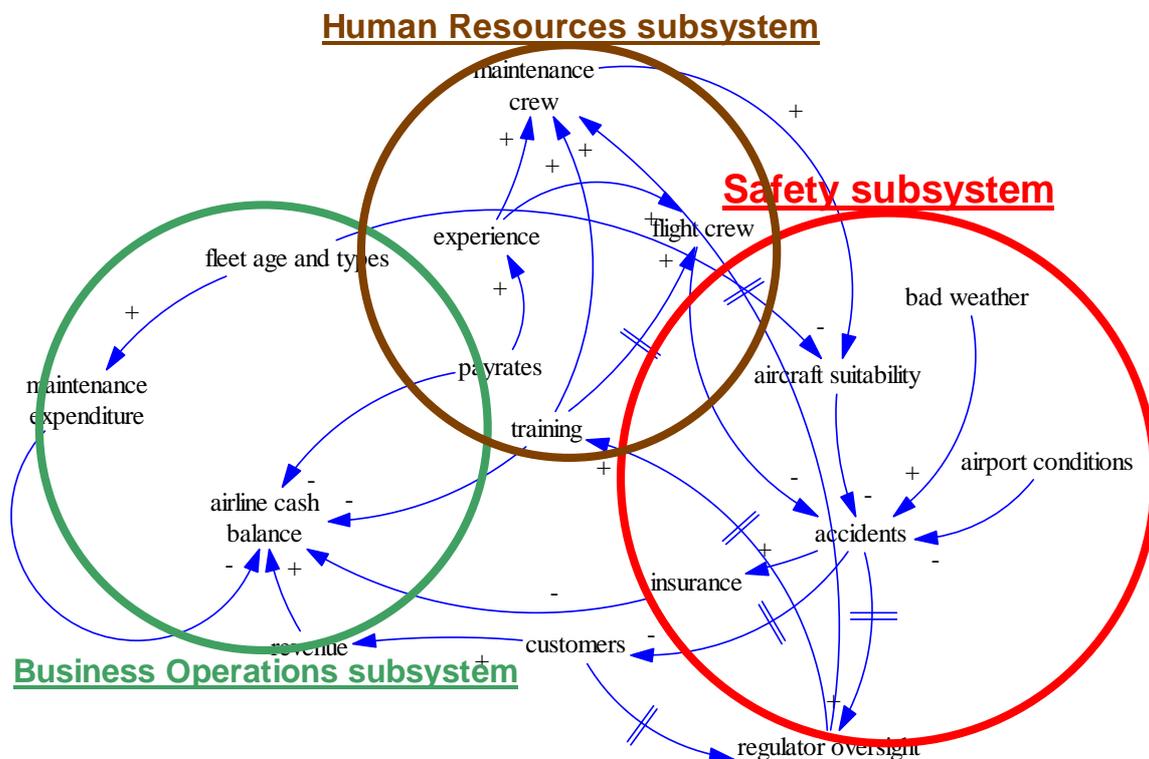


Figure 10. Connecting the 3 initial subsystems

Other connections are now made:

Insurance, payrates and training are connected to the airline cash balance variable. The age of the airline fleet and the number of different aircraft types would have an effect on the cost of maintenance – see Rhoades et al (2005) and Easdown and Wilms (2002).

The Regulator Oversight must extend to the functions that the CAA has authority over. This would include the maintenance performed on aircraft and the provision of training.

Maintenance – as carried out by the maintenance crew – would affect the condition of the aircraft for flying. This, in turn, would affect how and if accidents occurred.

The commercial aspect of the model is incomplete. There must be a feedback loop between the ‘airline cash balance’ and the various expenditure items – maintenance, training, payrates. The capital costs of aircraft procurement or leasing must also be added.

The CAA gets the bulk of its income from levies of airline customers (CAA, 2012). This is reflected by a connection between ‘customers’ and ‘safety regulator’.

Figure 11 shows the final CLD which contains 40 loops. Three such loops are highlighted. The one coloured green shows a balancing loop (B1) which involves ‘accidents’ and ‘regulator oversight’ (ie maintenance quality loop). The one coloured red identifies a reinforcing loop (R1) that involves the same 2 variables plus ‘training’ and ‘payrates’ (ie the training cost implications loop). A second balancing loop (B2)

operates to reduce accidents after a delay by additional training for airline staff (staff training loop).

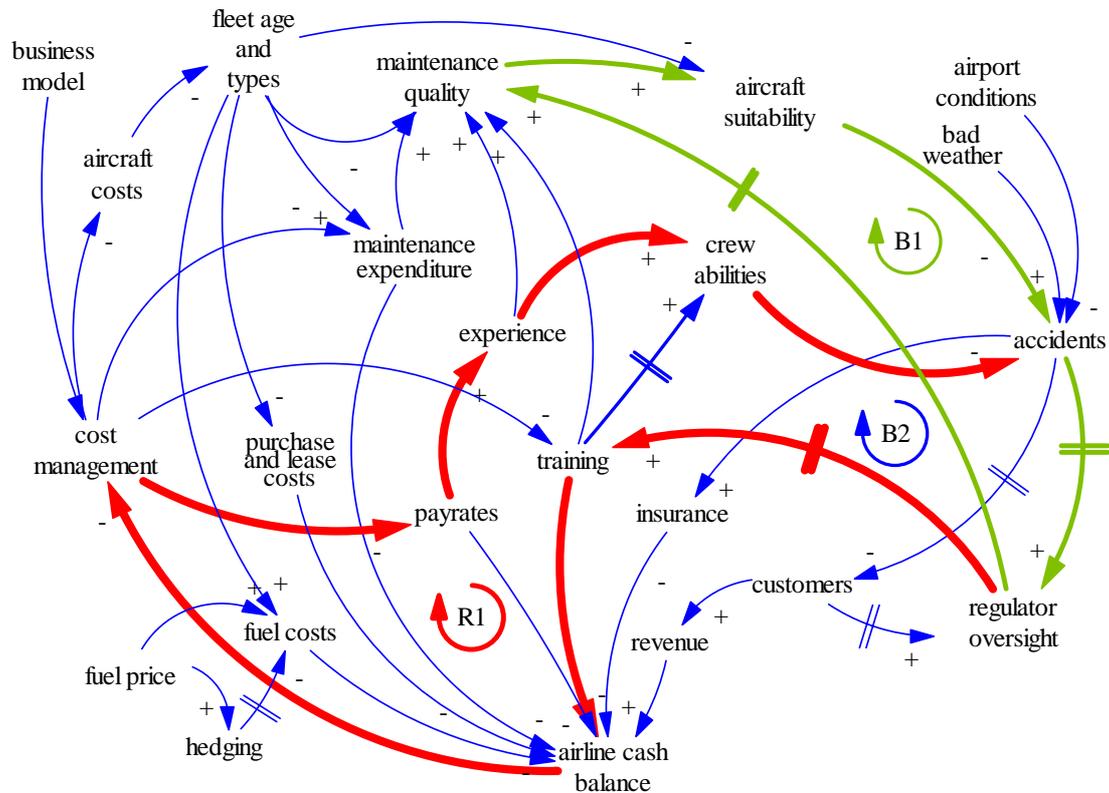


Figure 11. Airline safety CLD with feedback loops highlighted

In the balancing loop (B1), an increase in accidents leads, after a delay, to an increase in regulator oversight activity. This subsequently results in an increase in maintenance quality and aircraft suitability, which thereby decreasing the future accident rate.

The reinforcing loop (R1) is more elaborate. As in loop B1 above, an increase in accidents leads to an increase in regulator oversight activity. This also leads to an increase in training, which decreases airline cash balances, thus putting additional strain on the airline's cost management. Often the response to this situation is a decrease in payrates, leading to decreases in the experience of retained and recruited staff, further decreasing crew ability and subsequently leading to an increase in accidents.

Although another balancing loop (B2) evident in this diagram does result in a reduction in accidents due to the additional training airline staff receive, there is evidence of a 'Fix that Fails' systems archetype (Senge, 1990) operating here. This balancing loop would be offset by the adverse effects of reinforcing loop R1, unless other measures are put in place to prevent these adverse unintended consequences occurring.

7 Concluding Comments

Due to the complexity of the airline safety causal loop diagram outlined in this paper, it is extremely difficult to analyse the expected dynamic behaviour of the main variables

over time. Hence the 'obvious' conclusion is to go the next step to build a system dynamics simulation model. However, we think it would be difficult to get all the data for the 'full' CLD as shown. Nevertheless an initial model for a New Zealand airline could be developed based on available data, as a step towards a more comprehensive simulation model.

Finally, we consider that the causal loop diagram developed here, including human resources, business operations and safety subsystems does go some way towards developing a theory of airline safety using system dynamics methods as outlined in the special issue of *Systems Research & Behavioral Science* edited by Lane & Schwaninger (2008).

8 References

- CAA. 2012. Civil Aviation Authority of New Zealand, Wellington, New Zealand. http://www.caa.govt.nz/safety_info/safety_reports.htm
- Cavana, R.Y., D. Boyd, and R. Taylor. 2007. "A systems thinking study of retention and recruitment issues for the New Zealand Army electronic technician trade group." *Systems Research & Behavioral Science* **24**: 201-216.
- Cooke, D. L. 2003. "A system dynamics analysis of the Westray mine disaster" *System Dynamics Review* **19**:139–166.
- Cooke, D. L. 2004. *The Dynamics and Control of Operational Risk*, unpublished PhD thesis, Haskayne School Of Business Calgary, Alberta, Canada.
- Coyle RG. 1996. *System Dynamics Modelling: A Practical Approach*, Chapman & Hall: London.
- Easdown, G and P. Wilms 2002. *Ansett. The Collapse*, Thomas C. Lothian Pty Ltd., Melbourne, Australia.
- Forrester JW. 1961. *Industrial Dynamics*, MIT Press, Cambridge, MA (now available from Pegasus Communications: Waltham, MA).
- Lane, D.C and M. Schwaninger 2008. "Theory building with system dynamics: topic and research contributions." *Systems Research & Behavioral Science* **25**:439-445.
- Liehr, M; A. Großler; M. Klein and P. Milling 2001. "Cycles in the sky: understanding and managing business cycles in the airline market" *System Dynamics Review* **17**:311–332.
- Maani, K.E., R. Y. Cavana 2007. *Systems Thinking, System Dynamics: Managing Change and Complexity* 2nd ed., Pearson Education NZ Limited, Auckland, New Zealand.
- Moizer, J. D. 1999. *System Dynamics Modelling of Occupational Safety. A Case Study Approach*, unpublished PhD. Thesis, University of Stirling, UK.
- Phillips, R. A., W. K. Talley 1992. "Airline safety investments and operating conditions: determinants of aircraft damage severity." *Southern Economic Journal* **59**:157-164.
- Reason, J. 1997. *Managing the Risks of Organisational Accidents*, Ashgate Publishing Limited, Aldershot, England, UK.
- Rhoades, D. L., B. Waguespack Jr. 2000. "Judging a book by it's cover: the relationship between service and safety quality in US national and regional airlines." *Journal of Air Transport Management* **6**:87 –94.
- Rhoades, D. L., R. Reynolds, B. Waguespack, Jr. and M. Williams 2005. "The effect of line maintenance activity on airline safety quality." *Journal Of Air Transportation* **10**:59 – 71.

- Richardson, G.P., A.L. Pugh III. 1981. *Introduction to System Dynamics Modelling with DYNAMO*, Productivity Press, Mass.
- Rose, N. I. 1992. "Fear of flying? Economic analyses of airline safety." *Journal of Economic Perspectives* **6**:75-94.
- Salge, M. and P. M. Milling 2004. "The pace or the path? Resource accumulation strategies in the US airline industry", *2004 International Conference of the System Dynamics Society Proceedings*, Keble College, Oxford, England.
- Senge P. 1990. *The Fifth Discipline: The Art and Practice of the Learning Organization*, Doubleday/Currency, New York.
- Sterman, J. D. 2000. *Business Dynamics. Systems Thinking and Modeling for a Complex World*, McGraw-Hill Higher Education, Boston, USA.
- Wilson, M 1997. "Safety concerns of startup airlines" *Journal of Air Transportation World Wide* **2**:38-45.

Solving Bi-objective Traffic Assignment Based on Time Surplus Maximisation

O. Perederieieva
Department of Engineering Science
The University of Auckland
New Zealand
o.perederieieva@auckland.ac.nz

Abstract

Traffic congestion is an issue in most cities worldwide. One way to model and analyse the effect of congestion on route choice behaviour is traffic assignment (TA). Conventional TA models are based on the assumption that all drivers minimise their travel time or generalised cost, which usually represents a linear combination of time and monetary cost. This approach is not general. It allows to find only a subset of all equilibrium solutions.

We propose to use a conceptually different approach inspired by the multi-objective definition of optimality – the bi-objective user equilibrium. It considers two objectives separately and allows multiple solutions. In order to model user preferences, we apply the time-surplus maximisation model (TSMaXBUE) which can identify one of the solutions.

Time surplus is defined as the maximum time a user is willing to spend minus the actual time spent. The maximum time a user is willing to spend is modelled as an indifference curve - a non-linear function that depends on the path toll.

The TSMaXBUE model allows to obtain various traffic patterns by changing the indifference curves. We observe that this framework is general enough to cover any situation with a flow dependent and a flow independent component of path cost.

Key words: Traffic assignment, bi-objective user equilibrium, time surplus.

1 Introduction

Due to the fast development of cities and road networks the role of transportation planning grows every day since it provides tools for developing effective transportation spending and policies. As presented in (Ortúzar and Willumsen 2001), transportation planning allows to: analyse the current usage of a road network; predict the impact of potential projects and policies; control traffic (level of congestion, emission, toll revenue etc). In order to achieve these goals, a model that can realistically describe travel decisions made by drivers is required. However, in order to create such a model one should know how many people are travelling, from where to where they are travelling, which travel mode they choose (car, bus, bicycle etc) and which

routes they prefer. Since it is very difficult to predict how a particular individual will travel, the usual approach to tackle this difficulty is to make some assumptions on how people usually choose routes and to find a flow pattern satisfying these assumptions. The most well-known such assumptions are the ones following Wardrop's first principle that is also called *user equilibrium condition* (Wardrop 1952): *The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.*

This principle models the behaviour of travellers by assuming that all drivers are selfish and they tend to choose the fastest routes going from their origin to their destination. As a result an *equilibrium state* is achieved, when no one has an incentive to switch to another route. The mathematical model based on Wardrop's first principle is called traffic assignment (TA) problem. Given that a transportation network is defined as a directed graph $G(N, A)$ where N is a set of nodes and A is a set of links, TA can be stated as explained below (Florian and Hearn 1995).

Indices

a = link: $A = \{1, 2, \dots, |A|\}$;

p = origin-destination pair: $Z = \{1, 2, \dots, |Z|\}$ is a set of origin-destination (O-D) pairs;

k = path: $K_p = \{1, 2, \dots, |K_p|\}$ is a set of possible paths between O-D pair p , $K = \cup_{p \in Z} K_p$ is the set of all paths.

Parameters

δ_a^k equals one if link a belongs to path k , and zero otherwise;

D_p = demand of O-D pair p (*demand* represents how many vehicles are travelling from an origin to a destination);

$C_k(\mathbf{F})$ = is a function that describes travel time on path k , it depends on path flows $\mathbf{F} = (F_1, F_2, \dots, F_{|K|})$, i.e. the number of vehicles per time unit on each path;

$U_p = \min_{k \in K_p} C_k(\mathbf{F})$ is the minimum travel time for O-D pair p ;

Decision variables

F_k = flow on path k , $F_k \in \mathbf{F}$.

Model TA

$$F_k (C_k(\mathbf{F}) - U_p) = 0, \quad \forall k \in K_p, \forall p \in Z, \quad (1)$$

$$C_k(\mathbf{F}) - U_p \geq 0, \quad \forall k \in K_p, \forall p \in Z, \quad (2)$$

$$\sum_{k \in K_p} F_k - D_p = 0, \quad \forall p \in Z, \quad (3)$$

$$F_k \geq 0, \quad \forall k \in K_p, \forall p \in Z, \quad (4)$$

$$U_p \geq 0, \quad \forall p \in Z. \quad (5)$$

Explanation The objective is to find a flow pattern \mathbf{F}^* such that conditions (1)-(5) are satisfied, i.e. it is an *equilibrium problem*. Equations (1) and (2) state that if the flow F_k^* is positive (i.e. path k is used), then the travel time on this path $C_k(\mathbf{F}^*)$ is equal to minimum travel time U_p^* , and if the travel time $C_k(\mathbf{F}^*)$ is greater than U_p^* , then the flow F_k^* must be equal zero. Equation (3) is simply a conservation of flow constraint, and equations (4) and (5) are non-negativity constraints.

Function $C_k(\mathbf{F})$ might represent a measure different from travel time (for instance, linear combination of travel time and cost). As a result $C_k(\mathbf{F})$ is often called *generalised cost function* or simply *cost function* or *cost*. In the following all these terms are used interchangeably. The way the function $C_k(\mathbf{F})$ is defined decides how TA can be alternatively formulated and how difficult it is to solve.

Let $\mathbf{f} = (f_1, \dots, f_{|A|})$ denote a vector of link flows. They are related to path flows by the expression $f_a = \sum_{p \in Z} \sum_{k \in K_p} \delta_a^k F_k$. In order to ensure the existence of a solution of TA all path cost functions $C_k(\mathbf{F})$ must be *positive and continuous*, and to ensure the uniqueness of the solution of TA in terms of link flows \mathbf{f} all path cost functions $C_k(\mathbf{F})$ must be *strictly monotone* (Florian and Hearn 1995). In the following it is assumed that these requirements are satisfied.

2 Conventional Traffic Assignment Problem

The conventional model of the traffic assignment problem is based on two important assumptions that allow to formulate and solve it as a mathematical program. These assumptions are stated as follows:

1. *Additivity* of path cost functions: travel time of each path is the sum of travel times of links belonging to this path, i.e. $C_k(\mathbf{F}) = \sum_{a \in A} \delta_a^k \cdot c_a(\mathbf{f})$, where $c_a(\mathbf{f})$ is a function representing time needed to go through link a ;
2. *Separability* of link cost functions: travel time of each link depends only on flow on this link, i.e. $c_a(\mathbf{f}) = c_a(f_a)$.

If these assumptions are satisfied, solving the following optimisation problem results in the link flows satisfying the user equilibrium condition (Sheffi 1985):

$$\begin{aligned}
 & \min \sum_{a \in A} \int_0^{f_a} c_a(x) dx \\
 & \sum_{k \in K_p} F_k = D_p, \quad \forall p \in Z, \\
 & F_k \geq 0, \quad \forall k \in K_p, \forall p \in Z, \\
 & f_a = \sum_{p \in Z} \sum_{k \in K_p} \delta_a^k F_k, \quad \forall a \in A.
 \end{aligned} \tag{6}$$

As can be seen, model (6) is based only on travel time. Therefore, it is assumed that all drivers make their travel decisions regarding travel time only. However, this is not true in general. According to empirical studies other important factors are travel time reliability and monetary cost (Wang, Ehrgott, and Chen 2012). This fact motivates researchers to introduce models based on multiple criteria rather than on a single one. The next section describes how the conventional TA can be extended to a multi-objective framework.

3 Bi-objective Traffic Assignment

As discussed in (Raith 2009), in the literature on the TA problem where two or more objectives are explicitly distinguished the majority of the models form a weighted

sum of the objectives, or introduce several user classes with different weighting factors in each class, or assume a distribution of weighting factors. These ideas seem to be very natural since when two objective functions are combined into a single one via a linear relation, model (6) still results in the user equilibrium condition and, hence, can be used to solve TA. Moreover, by changing weights in the linear combination of objectives and solving the TA problem again, different flow patterns can be obtained. This strategy even follows the main idea of multi-objective optimisation where instead of one optimal solution a set of *efficient* solutions is generated (Ehrgott 2005).

However, it is well-known from multi-objective optimisation that in general linear combination of objectives allows to find only a certain subset of the set of efficient solutions (Ehrgott 2005). They are called *supported efficient solutions*, all other solutions are called *non-supported*. As presented in (Raith 2009), the same conclusion applies in case of TA, i.e. by combining objectives into a single one via a linear relation only supported efficient solutions can be found. Moreover, the TA model is an equilibrium problem, and not an optimisation one. Therefore, it seems more appropriate to redefine the user equilibrium condition instead of trying to integrate multiple objectives into a mathematical program (6).

The *bi-objective user equilibrium (BUE)* condition can be stated as follows: *under bi-objective user equilibrium condition, traffic arranges itself in such a way that no individual trip maker can improve either of their objectives or both of them without worsening the other component by unilaterally switching routes* (Wang, Raith, and Ehrgott 2010). As can be seen, the BUE condition is a direct extension of Wardrop's first principle.

Let us first introduce the definition of efficiency. Let $C_k^{(1)}(\mathbf{F})$ and $C_k^{(2)}(\mathbf{F})$ denote the objectives that each driver takes into account during travel planning. Then a cost vector $\mathbf{C}_k(\mathbf{F}) = \left(C_k^{(1)}(\mathbf{F}), C_k^{(2)}(\mathbf{F}) \right)$ is associated with each path $k \in K_p, \forall p \in Z$. Path k is called *efficient* if for a given flow solution \mathbf{F} and a given O-D pair p , there is no path whose cost vector dominates the cost vector of path k . Here vector \mathbf{x} *dominates* vector \mathbf{y} if $x_i \leq y_i, \forall i = 1, \dots, |\mathbf{x}|$ and at least one inequality is strict (Ehrgott 2005), this is denoted by $\mathbf{x} \leq \mathbf{y}$. An example of efficient paths is presented in Figure 1: O-D pair p is under consideration, the flow \mathbf{F} is assumed to be given and fixed, circles and squares denote cost vectors of all available paths for O-D pair p , circles correspond to the efficient paths and squares correspond to the dominated ones. Therefore, a feasible vector \mathbf{F}^* satisfies the BUE condition if for each O-D pair p the following statement holds (Raith 2009):

$$\mathbf{C}_k(\mathbf{F}^*) \leq \mathbf{C}_{k'}(\mathbf{F}^*) \Rightarrow F_{k'}^* = 0, \quad \forall k \in K_p, \forall k' \in K_p, k' \neq k, \quad (7)$$

i.e. flow is positive only on efficient paths, dominated paths must carry zero flow. There might also exist *equivalent* paths, i.e. paths with identical cost vectors. Therefore, under the BUE condition any path with positive flow must be efficient or equivalent to another efficient path. It means that all drivers are assumed to travel on efficient paths instead of fastest ones as is the case in the conventional TA.

As usual in the multi-objective philosophy, the BUE condition allows multiple solutions, potentially infinitely many of them (Raith 2009). In practice, however, we are interested in finding only one solution. Moreover, it is desired to find a solution that describes an actual flow pattern occurring in reality. In order to achieve this goal we propose to apply the concept of time surplus maximisation bi-objective user

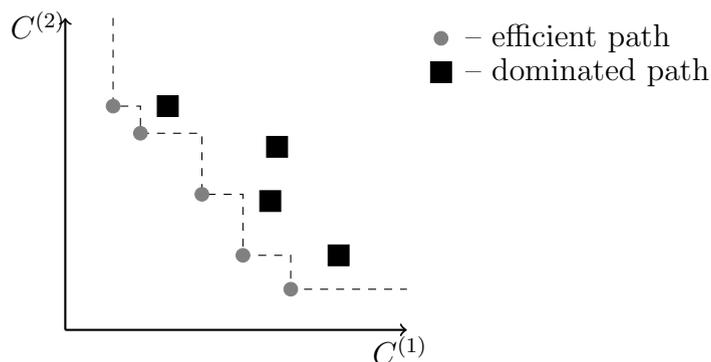


Figure 1: An example of efficient paths

equilibrium (TSMaXBUE) that is presented in the next section.

4 Time Surplus Maximisation Model

In TA each driver minimises path cost $C_k(\mathbf{F})$. In the conventional TA this cost is simply the travel time. In (Wang and Ehrgott 2011) the authors introduced the TSMaXBUE model which defines path cost $C_k(\mathbf{F})$ in a way that takes into account two criteria: travel time $t_k(\mathbf{f})$ and toll T_k . In the following it is assumed that only travel time depends on flow, and travel toll is independent of flow.

Each driver *maximises time surplus* which is defined as the maximum time a user is willing to spend minus the actual time $t_k(\mathbf{f})$ spent. The maximum time a user is willing to spend is modelled as an *indifference curve* $G_k(T_k)$ – a non-linear function that depends on the path toll. Therefore, each driver makes a decision based on travel time and toll which are combined into one criterion via an indifference curve. The most important observation here is that the indifference curve is a non-linear function. This allows to generate not only supported efficient flow patterns, but also non-supported ones (Raith 2009).

Under the assumption that all drivers have the same indifference curve which still might depend on the O-D pair, we obtain TSMaXBUE with one user class. When all drivers are divided into classes with different indifference curves, we obtain TSMaXBUE with multiple user classes. In the following, we consider the model with a single user class only.

Assumptions of the model can be summarised as follows: 1) All link cost functions are separable: $c_a(\mathbf{f}) = c_a(f_a)$; 2) Travel time of path k is additive: $t_k(\mathbf{f}) = \sum_{a \in A} c_a(f_a) \cdot \delta_a^k$; 3) Toll of path k does not depend on flow and is additive: $T_k = \sum_{a \in A} \tau_a \cdot \delta_a^k$ where τ_a is a fixed toll of link a ; 4) Indifference curves $G_p(T_k)$ are non-negative continuous and strictly decreasing functions of path toll T_k . Thus, in the TSMaXBUE model the time surplus of path k is defined as follows:

$$C_k(\mathbf{f}) = G_p(T_k) - t_k(\mathbf{f}), \quad \forall k \in K_p, \forall p \in Z. \quad (8)$$

For convenience, we transform the TSMaXBUE model into minimisation problem. This can be easily done by multiplying by -1 the generalised cost (8): $C'_k(\mathbf{f}) = -C_k(\mathbf{f}) = t_k(\mathbf{f}) - G_p(T_k)$. Since $G_p(T_k)$ is strictly decreasing, the function $\overline{G}_p(T_k) = -G_p(T_k)$ is strictly increasing. We also have to ensure that $\overline{G}_p(T_k)$ is non-negative. This can be achieved by adding to it a big enough constant, for example: $\overline{G}_p(T_k) =$

$G_p(0) - G_p(T_k)$. Thus, the path cost function of the TSMaXBUE model has the form:

$$C_k(\mathbf{f}) = t_k(\mathbf{f}) + \overline{G}_p(T_k) = t_k(\mathbf{f}) + G_p(0) - G_p(T_k), \quad (9)$$

If measure (9) is used to find an equilibrium solution and the assumptions presented earlier in this section are satisfied, then there is an equivalent mathematical program (MP) that allows to find this user equilibrium solution (Larsson et al. 2004):

$$\min \sum_{a \in A} \int_0^{f_a} c_a(x) dx + \sum_{p \in Z} \sum_{k \in K_p} F_k \cdot \overline{G}_p(T_k) \quad (10)$$

$$\sum_{k \in K_p} F_k = D_p, \quad \forall p \in Z, \quad (11)$$

$$F_k \geq 0, \quad \forall k \in K_p, \forall p \in Z. \quad (12)$$

After analysing this MP we can conclude that the algorithms used for conventional traffic assignment can be applied to solve it provided that a *non-additive shortest path algorithm* is available and *path flows* are used as explicit variables. Therefore, only *path-based approaches* (when the solution is represented by path flows) can be applied. This is due to the fact that the cost of each path can be evaluated only on the path level.

As mentioned earlier, a non-additive shortest path algorithm is required. Due to the assumptions made on functions $G_p(T_k)$ and additivity of both travel time and toll, the corresponding shortest path in terms of measure (9) is one of the efficient solutions of the bi-objective shortest path problem with objectives: $z_1 = \min_{k \in K_p} \sum_{a \in A} \delta_a^k \cdot c_a$ and $z_2 = \min_{k \in K_p} \sum_{a \in A} \delta_a^k \cdot \tau_a$ (Tsaggouris and Zaroliagis 2004). This problem is known to be NP-complete, e.g. (Ehrgott 2005). In order to solve it we implement the bi-objective label setting algorithm as presented in (Ehrgott 2005).

In order to solve the TSMaXBUE model we propose to apply the following path-based algorithms: path equilibration (PE) (Florian and Hearn 1995), projected gradient (PG) (Florian, Constantin, and Florian 2009) and gradient projection (GP) (Chen and Jayakrishnan 1998). Previously these methods were applied to solve the conventional traffic assignment problem with travel time as the objective and were reported to have a very promising performance. This motivates us to test the performance of adopting these methods to solve the more complicated TSMaXBUE model.

5 Numerical Study

All algorithms were implemented in the C++ programming language and compiled using g++ 4.6.3 (Ubuntu/Linaro 4.6.3-1ubuntu5). All runs were performed under the following environment: Ubuntu Release 12.04 64-bit, Kernel Linux 3.2.0-24-generic; Intel Core i5-2500 CPU, 4 Core, 3.30GHz; 7.7 GB RAM. As a convergence criterion the following measure was used: $\max_{p \in Z} \{ \max_{k \in K_p^+} C_k(\mathbf{f}) - \min_{k \in K_p^+} C_k(\mathbf{f}) \}$.

5.1 Small Example

In order to validate the TSMaXBUE model, we first perform several tests on a small instance from (Raith 2009). This instance has grid structure (see Figure 2a),

25 nodes, 81 links and one O-D pair, all links are untolled except for the central ones highlighted in Figure 2a. Each tolled link has the same toll of \$5 (tolled links are faster than untolled ones). Other characteristics of this instance can be found in (Raith 2009).

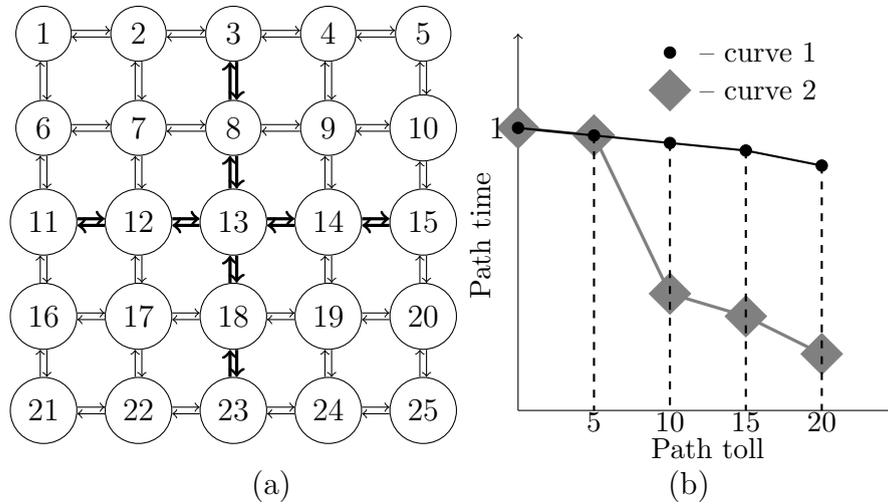


Figure 2: (a) – Tested instance; (b) – Indifference curves

The conventional user equilibrium solution (when tolls are not considered) is presented in Figure 3. Near each link there are two numbers: the number in brackets is link flow and the second number is the travel time of the corresponding link. The amount of flow is also visualised by the thickness of the links. The links that carry zero flow are not presented in the figure. Since link tolls are not taken into account the entire demand is mainly distributed on two fastest paths and these paths have the biggest toll of \$20. Because of this, the links on the fastest paths are highly congested whereas the remaining links do not carry flow at all or carry only a small amount of it.

In order to demonstrate how the solution depends on an indifference curve we created two curves presented in Figure 2b. Both of them are piecewise linear functions. The solution presented in Figure 4 corresponds to curve 1 that describes the situation when drivers are not concerned about tolls (the values of travel time the drivers are willing to spend are almost the same for all possible values of toll). As a result, they choose to travel on the fastest paths resulting in a flow pattern similar to the conventional one. The solution presented in Figure 5 corresponds to curve 2 that describes the situation when the drivers are willing to spend much less time compared to toll-free paths, if they have to pay more than 5\$. As a result, in this case all drivers choose the paths with toll of 5\$. This flow pattern is very different from the one corresponding to curve 1. Therefore, we can conclude that by changing the indifference curve it is possible to generate different flow patterns.

5.2 Bigger Instances

In order to compare performance of path-based algorithms we perform tests on the following instances: Sioux-Falls, Anaheim and Barcelona that are available at the web-site: <http://www.bgu.ac.il/~bargera/tntp/>. Link tolls were generated using the marginal cost approach, see (Ortúzar and Willumsen 2001), and then scaled (letters after each instance name correspond to different scaling strategies).

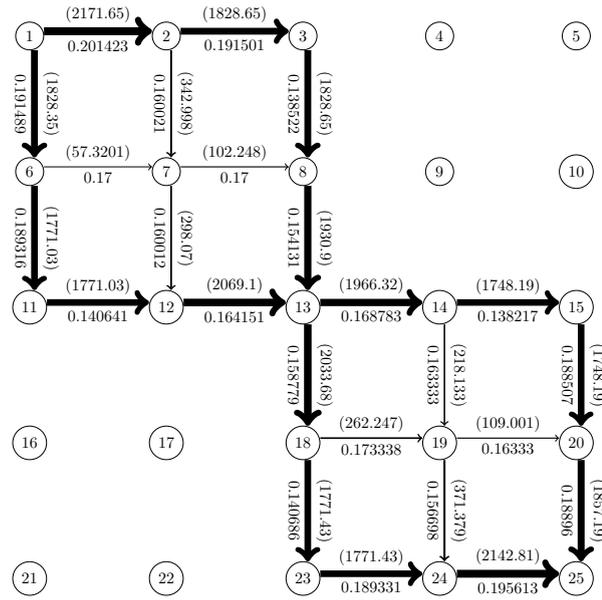


Figure 3: Conventional user equilibrium solution

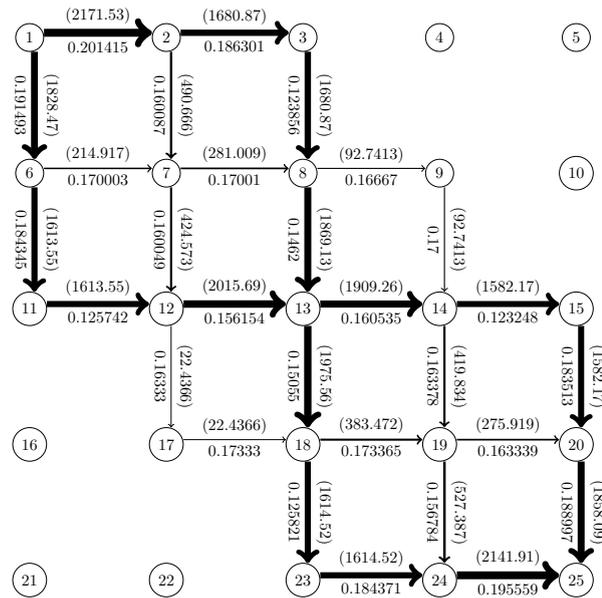


Figure 4: Solution corresponding to Curve 1

Indifference curves $G_p(m_k)$ were generated by assigning zero flow to all links and finding efficient paths, cost vectors of which were used as breakpoints of piecewise linear functions. The comparison of tested algorithms is presented in Figure 6, according to which the algorithms show similar performance, except GP with value of its specific parameter equal to 1.

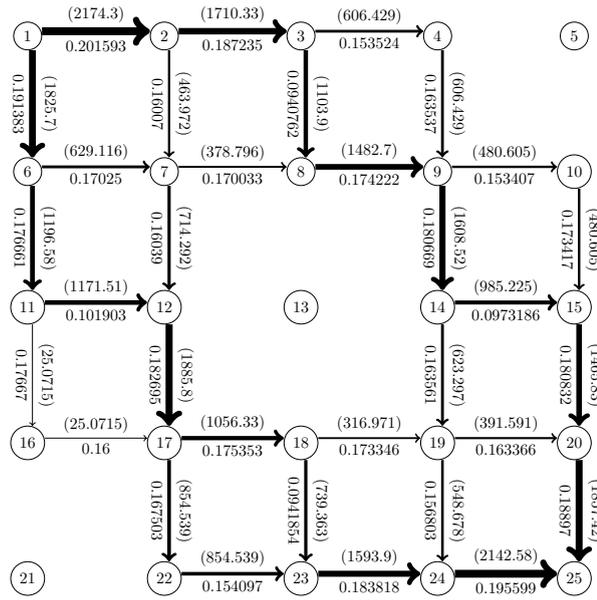


Figure 5: Solution corresponding to Curve 2

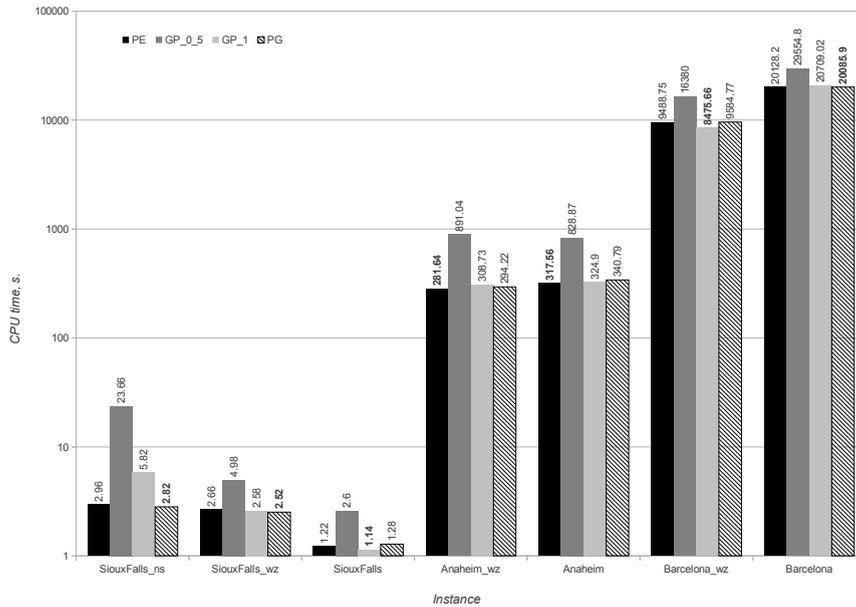


Figure 6: Performance of algorithms

6 Conclusion

The BUE condition is conceptually different from the approaches based on aggregation of objectives. It is inspired by the multi-objective definition of optimality and allows multiple solutions (potentially infinitely many of them). We apply the TSMaXBUE model as a possible way to represent route choice behaviour in tolled road networks. In case of one user class it can be solved by optimisation-based algorithms. To solve the TSMaXBUE model we adopt some path-based algorithms used for conventional traffic assignment, compare their performance and study how the solution space depends on the chosen indifference curve. The TSMaXBUE model

allows to obtain various traffic patterns by changing the indifference curve that represents different preferences among user groups. We observe that this framework is general enough to cover any situation with a flow dependent and a flow independent component of path cost.

References

- Chen, A., and R. Jayakrishnan. 1998. "A path-based gradient projection algorithm: effects of equilibration with a restricted path set under two flow update policies." *Transportation Research Board Annual Meeting*.
- Ehrgott, M. 2005. *Multicriteria Optimization*. Berlin: Springer.
- Florian, M., I. Constantin, and D. Florian. 2009. "A New Look at Projected Gradient Method for Equilibrium Assignment." *Transportation Research Record: Journal of the Transportation Research Board* 2090:10–16.
- Florian, M., and D. Hearn. 1995. "Network equilibrium models and algorithms." Chapter 6 of *Network Routing*, edited by C.L. Monma M.O. Ball, T.L. Magnanti and G.L. Nemhauser, Volume 8 of *Handbooks in Operations Research and Management Science*, 485 – 550. Elsevier.
- Larsson, T., P. Lindberg, M. Patriksson, and C. Rydergren. 2004. "On Traffic Equilibrium Models with a Nonlinear Time/Money Relation." In *Transportation Planning*, edited by M. Patriksson and M. Labbé, Volume 64 of *Applied Optimization*, 19–31. Springer US.
- Ortúzar, J. D., and L.G. Willumsen. 2001. *Modelling Transport*. Chichester New York: J. Wiley.
- Raith, A. 2009. "Multiobjective Routing and Transportation Problems." Ph.D. diss., The University of Auckland.
- Sheffi, Y. 1985. *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Englewood Cliffs, N.J.: Prentice-Hall.
- Tsaggouris, G., and C. D. Zaroliagis. 2004. "Non-additive Shortest Paths." *12th Annual European Symposium, Bergen, Norway, September 14-17, 2004. Proceedings*. 822–834.
- Wang, J. Y. T., and M. Ehrgott. 2011. "Modelling stochastic route choice with bi-objective traffic assignment." *International Choice Modelling Conference 2011, Leeds, U.K., 4-6 July 2011*.
- Wang, J. Y. T., A. Raith, and M. Ehrgott. 2010. "Tolling Analysis with Bi-objective Traffic Assignment." In *Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems*, edited by M. Ehrgott, B. Naujoks, T. J. Stewart, and J. Wallenius, Volume 634 of *Lecture Notes in Economics and Mathematical Systems*, 117–129. Springer Berlin.
- Wang, J.Y.T., M. Ehrgott, and A. Chen. 2012. "A Bi-objective User Equilibrium Model of Travel Time Reliability in a Road Network." *Paper to be presented at The 5th International Symposium on Transportation Network Reliability to be held in Hong Kong, 18-19 December 2012*.
- Wardrop, J.G. 1952. *Some theoretical aspects of road traffic research*. Institution of Civil Engineers.

Efficient Timetable Modifications in University Course Timetabling

Antony Phillips, Matthias Ehrgott, David Ryan
Department of Engineering Science
University of Auckland
New Zealand
antony.phillips@auckland.ac.nz

Abstract

University course timetabling is a large resource allocation problem, in which both times and rooms are determined for each class meeting. Due to the difficulty of the problem, it is often handled in two stages: timetable generation followed by room allocation. Many universities will internally generate a timetable by hand, to suit their (often) complex and institution-specific needs. However they may use a software package to allocate rooms, for which efficient exact methods exist. It is clear that the feasibility (and quality) of the room allocation problem is highly dependent on the structure of the timetable. In most cases the original timetable will not allow for a feasible room allocation to exist, and will require modification.

This work focusses on using optimisation methods to automate the timetable modification process, while minimising the disruption to the original timetable. Using structured information gained from attempting the (infeasible) room allocation problem, the key congested areas and features of the timetable can be identified. This information is used to keep the model size small, by means of an optimistic methodology which attempts to resolve the infeasibility in as small a local neighbourhood as possible. Computational results on full size datasets are presented and practical issues are discussed.

Key words: University Timetabling, Integer Programming, Decision Support Systems.

1 Introduction

The University Course Timetabling Problem (UCTP) requires finding a time and a room for every class meeting (or *contact*) out of limited university resources. The problem has always drawn significant interest from the academic community, and increasingly the private sector. Despite continued advancement in computational capabilities, the large size of modern universities remains highly limiting on the use of automated methods.

The majority of algorithms proposed for the UCTP in the literature are meta-heuristics which build a full timetable from scratch, dealing with the times and rooms for contacts in the same algorithm. An alternative approach to the problem

is to decompose it into a timetabling component (contacts-to-times) and a room allocation component (contacts-to-rooms). Algorithms for dealing with the UCTP are less well studied, despite this decomposition being common in practice.

Many universities choose to internally generate a timetable using existing knowledge of their particular requirements, and then use an automated method to allocate rooms. However, because the timetable generation does not take the available room resources into consideration, this timetable may not allow for a feasible room allocation to exist. Usually this is due to many contacts assigned in “peak” timeslots in the centre of the day, but it can also occur when rooms with specialised equipment are requested. Traditionally, the timetable is then manually modified until a feasible room allocation can be found.

This paper introduces an efficient automated algorithm to modify the timetable as little as possible while achieving feasibility in the room allocation. It is clear that the difficulty of solving the Timetable Modification will depend on the structure of the timetable, and the resulting degree of infeasibility in the Room Allocation. This paper assumes that the timetable has been generated to meet potentially complex unknown quality metrics and so it is undesirable to modify the timetable any more than necessary. We test our algorithm on data from 2010 at the University of Auckland, where this was the method for generating a timetable.

2 The Room Allocation Algorithm

The Timetable Modification problem arises following an attempted solve of an infeasible Room Allocation Problem. Although the room allocation is not the central focus of this paper, the algorithm outlined here is designed to reveal key information about the nature of the room infeasibility, which will be used to modify the timetable.

Our room allocation algorithm takes a timetable as input and solves a sequence of binary integer programs (BIP), each modelling feasible contact-to-room assignments with different objective functions. The first problem, “Contact Hours”, simply attempts to find a suitable room (in size and attributes) for as many contacts as possible. If a room can be found for all contacts, then this room allocation is considered feasible in a practical context. However, if the BIP terminates with any contacts unallocated, timetable modification is necessarily required. This certainty of (in)feasibility is an advantage of using an optimisation algorithm.

At this stage, although it is known whether or not the timetable will require modification, we continue to focus on the room allocation. Several further BIPs are solved to improve the quality of the partial room assignment, by deciding specifically which contacts to leave unallocated and which room to assign to each contact. The “Seated Student Hours”, “Seat Utilisation” and “Building Preference” expressions are each used in turn as objective functions to the room allocation. In each case, previous objectives are maintained at their optimal values by means of constraints in the current model. These objectives help to give precedence to large contacts and provide good fits of contact-to-room in terms of size and building.

This process is shown in Figure 1 whereby each box represents solving an optimisation model with previous objectives fixed. After timetable modification when a feasible room allocation is found, a similar sequential objective fixing is performed to decide the final room allocation.

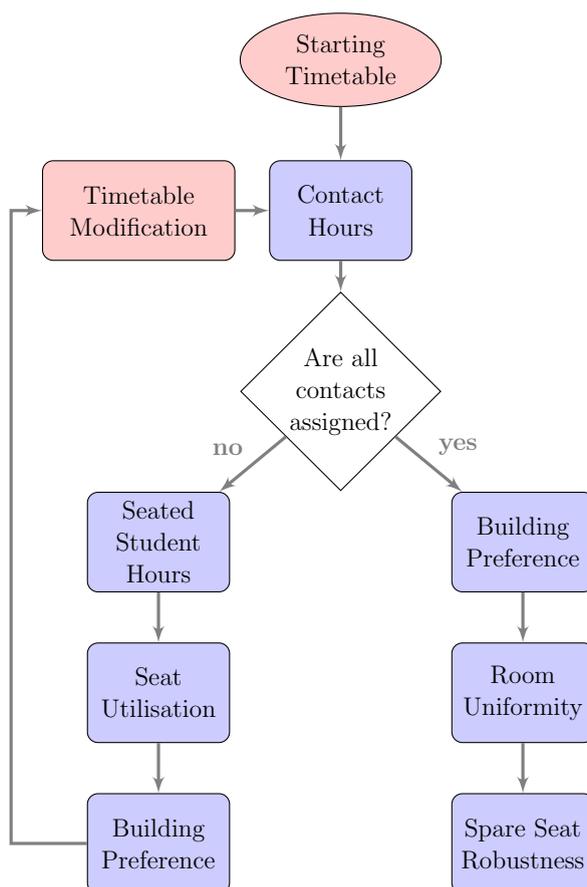


Figure 1: UoA Timetabling Process

3 The Timetable Modification Algorithm

3.1 Expanding Neighbourhood Algorithm

Timetable Modification requires a consideration of both the timeslot and room for each contact, rather than handling them separately. However, building a model with variables indexed over contacts, timeslots and rooms could easily result in millions of binary variables. Therefore it is important to limit the number of variables as much as possible.

Based on room utilisation studies (Beyrouthy et al. 2007), it is reasonable to assume for practical problems that there does exist a timetable with a feasible room allocation. In fact, it is likely that the timetable will not need to be significantly changed, particularly if the timetable is based on one which was implemented in a previous year.

The fundamental idea of our approach is to consider each infeasible timeslot separately, and to try to resolve the infeasibility in as small a local *neighbourhood* as possible. Because we know most contacts will stay in their existing time and room, the *neighbourhood* refers to all contacts which we are allowing to move. We will initially generate a very small neighbourhood, such as one comprising of all large contacts in the timeslots one hour before and after the infeasible timeslot. Then an optimisation problem is solved inside this neighbourhood, to attempt to find a feasible timetable and room allocation while minimising disruption (as detailed in Section 3.2). Due to the use of an exact algorithm, we will know with certainty whether a feasible solution exists in this neighbourhood or not. If this model is

found to be infeasible, we will progressively expand the neighbourhood to other contacts in nearby timeslots until a feasible solution can be found. This algorithm is presented as Algorithm 1.

By using an optimistic approach i.e. starting with as small a neighbourhood as possible, the optimisation models are also kept small. Although several optimisation problems may need to be solved until a feasible solution is found, these will necessarily be small and relatively fast. The idea of an expanding neighbourhood has also been demonstrated as an efficient way of overcoming combinatorial dimensionality issues in other applications (Rezanova and Ryan 2010).

Algorithm 1 Expanding Neighbourhood Algorithm

```

for all its in Timetable.getInfeasibleTimeslots() do
  N ← generateInitialNeighbourhood(its, Timetable.contactInformation)
  searching ← True
  while searching do
    P ← generateNeighbourhoodBIP(N)
    P.solve()
    if P.isFeasible() then
      Timetable.update(P.solution)
      searching ← False
    else
      N.expand()
    end if
  end while
end for

```

3.2 Neighbourhood BIP Model

Within a given neighbourhood, a Binary Integer Program (BIP) is used to attempt to find a feasible allocation of contacts to timeslots and rooms, using the formulation given below. As mentioned in Section 3.1, significant attention is paid to the scope of the neighbourhood, as represented in the “Sets” over which variables are generated.

Although the timetabling and room allocation problems are both represented in this formulation, the structure is such that many of the complex constraints can be handled implicitly through the variable generation. The definitions of C_t and C_r (or equivalently T_c and R_c) control the scope of when and where contacts can be reallocated which offers significant modelling power. In particular, for many streams it is required that a maximum of one contact is held per day. If a *fixed* contact (or a ‘non-neighbourhood’ contact) from a stream is held on day d , then no timeslots on this day will be in T_c , which the decision variables are indexed over. Similarly, a *curriculum* is a group of streams which cannot be held in the same timeslot (as they are taken by a common group of students). If a contact from curriculum cu is fixed in a timeslot within the neighbourhood, then no contacts from streams in this curriculum can be moved to this timeslot. Finally, timetable elements which span multiple consecutive timeslots (e.g. two-hour lectures) are generalised as sets of contacts known as *multis*. If a multi lies partially in the neighbourhood, the

contacts inside the neighbourhood are confined to their timeslots to ensure the time continuity.

Sets

C = contacts in neighbourhood N ;	S = streams
C_t = contacts feasible for timeslot t ;	T = timeslots in neighbourhood N ;
C_r = contacts feasible for room r ;	D = days in neighbourhood N ;
$C_{tr} = C_t \cap C_r$	T_d = timeslots on day d ;
C_s = contacts part of stream s ;	T_c = timeslots feasible for contact c ;
C_{mf} = contacts in multis fully in N ;	R_c = rooms feasible for contact c ;
C_{cu} = contacts in curriculum cu	R_t = rooms available in timeslot t ;
CU = curricula	$R_{ct} = R_c \cap R_t$

Decision Variables

$$x_{ctr} = \begin{cases} 1 & \text{if contact } c \text{ is held at time } t \text{ in room } r \\ 0 & \text{otherwise} \end{cases} \quad \forall c \in C, t \in T_c, r \in R_{ct}$$

Parameters

dp = penalty for moving a contact by one day (default: 3);
 pp = penalty for moving a contact by one period (default: 1);
 t_init_c = timeslot for contact c in initial timetable;
 r_init_c = room for contact c in preliminary room allocation;

$$w_{ctr} = \begin{cases} dp * |\text{dayDiff}(t, t_init_c)| + \\ pp * |\text{periodDiff}(t, t_init_c)| & \text{if } t \neq t_init_c \\ 0.01 & \text{if } r \neq r_init_c \text{ \& } t = t_init_c \\ 0 & \text{if } r = r_init_c \text{ \& } t = t_init_c \end{cases} \quad \forall c \in C, t \in T_c, r \in R_{ct}$$

Model

$$\text{Minimise } \sum_{c \in C} \sum_{t \in T_c} \sum_{r \in R_{ct}} w_{ctr} * x_{ctr}$$

$$\sum_{c \in C_{tr}} x_{ctr} \leq 1 \quad \forall t \in T, \forall r \in R_t \quad (1)$$

$$\sum_{t \in T_c} \sum_{r \in R_{ct}} x_{ctr} = 1 \quad \forall c \in C \quad (2)$$

$$\sum_{\substack{c \in C_s, \\ c \notin C_{mf}[1 \rightarrow]}} \sum_{t \in T_c \cap T_d} \sum_{r \in R_{ct}} x_{ctr} \leq 1 \quad \forall s \in S, \forall d \in D \quad (3)$$

$$\sum_{c \in C_{cu}} \sum_{r \in R_{ct}} x_{ctr} \leq 1 \quad \forall cu \in CU, \forall t \in T \quad (4)$$

$$\sum_{r \in R_{ct}} [x_{ctr} - x_{(c-1)(t-1)r}] = 0 \quad \forall c \in C_{mf}[1 \rightarrow], \forall t \in T_c \quad (5)$$

Explanation

The objective function minimises the total timetable “distance” between the timetable solution, and the initial (infeasible) timetable. The relative sizes of the penalty parameters may depend on individual preferences, and can even be customised for every potential contact-time-room variable. A small disincentive in the objective is made for contacts which remain in their own timeslot to change to a different room. This ensures that the only room swaps made by contacts within a timeslot are necessary for a better solution. For example, they can be made to “free-up” a room in a particular timeslot for a contact from another timeslot to use. The disincentive helps to maintain the structure and quality of the preliminary room allocation, as originally obtained by the iterations of the Room Allocation algorithm (Figure 1).

Constraint 1 ensures that each room in each timeslot is used by a maximum of one contact, and conversely Constraint 2 ensures that all contacts are assigned to one room in one timeslot. Constraint 3 ensures that two contacts from the same stream cannot move to timeslots on the same day. The exception to this rule is where the contacts are part of a multi, so only the first contact of any multi is included in the constraint. Constraint 4 ensures that two contacts from the same curriculum cannot move to the same timeslot. Lastly, Constraint 5 enforces the time-continuity on the contacts of a multi. In some formulations, room continuity is enforced as a hard constraint on the contacts of a multi. This is done with an alternative version of Constraint 5, where the room summation becomes one constraint per room. However in this case we treat room-continuity as a quality measure and so it is handled in the final room allocation.

4 Defining the neighbourhood

As mentioned in Section 3.1, it is desirable to keep the neighbourhood as small as possible. This necessitates a careful definition of the neighbourhood at each stage of expansion to include the most promising variables (i.e. contact-time-room allocations) in the model first.

If we consider the properties of the *unallocated contacts* (contacts which were not assigned rooms in the room allocation), we not only have knowledge of which timeslots are infeasible, but we also can deduce the reason for the infeasibility.

For example, if several large contacts are unallocated, we could assume that this timeslot has too many large contacts scheduled for the available rooms. However if we only run the “Contact Hours” solve during the room allocation, then a shortage of two medium-size rooms could result in the unallocated contacts arbitrarily being either medium or large-sized. As a result, we may be searching for large rooms in the neighbouring timeslots, when only a medium-sized room is required. By maximising the “Seated Student Hours”, this ensures that larger contacts have preference to be allocated, and so the contacts left unallocated will be the smallest possible.

The “Contact Hours” and “Seated Student Hours” solves choose which contacts get be allocated a room (and which remain unallocated), in the infeasible timeslots. However, they have no influence on the configuration of contacts to rooms out of those contacts which are allocated. Solving “Seat Utilisation” chooses this configuration such that any unused rooms are the large possible, and that contacts get

the best possible fit into the rooms. For the unused rooms, if a timeslot has many medium-sized contacts occupying the largest rooms but some medium rooms free, the timeslot would not be such an obvious candidate for the neighbourhood if we were searching for large rooms. Furthermore, in order to facilitate the large rooms being potentially “freed-up” in the BIP, we would need to generate many extra variables. Achieving a good utilisation fit between contacts and rooms can also help reduce the neighbourhood size, since contacts which fit well into their rooms are unlikely to need to change either timeslot or room.

Finally, room characteristics should be considered when formulating the neighbourhood. In some timeslots, contacts can remain unallocated due to rooms with particular characteristics being in shortage. In this case, it may be a good idea to allow any contacts occupying a room of this nature into the neighbourhood, regardless of the utilisation.

5 Computational Results

The Timetable Modification Algorithm was tested on data from the second semester of 2010 at the University of Auckland City Campus. The starting timetable contains 2234 contacts across 45 weekly timeslots, with 72 available rooms.

To test the concepts explained in Section 3 and 4, two neighbourhood definitions are used. In the former (denoted ‘N1’), all contacts in each neighbourhood timeslot are permitted to move to any other timeslot or room. In the latter (denoted ‘N2’), the neighbourhood is more tightly controlled, firstly being restricted to contacts which occupy a room of approximately equal size to that of the unallocated contacts. Furthermore, any contact with a utilisation of 80% or more in its current room will not enter the neighbourhood. However, an exception is made for contacts occupying a room which possesses any special characteristic requests of the unallocated rooms. All rooms that were free are also automatically in the neighbourhood.

In both cases the expansion of the neighbourhood progressively allows more timeslots to be included. This is done by a simple ordering of which timeslots to consider next, based approximately on the objective distance of moving from the centre timeslot. An expansion typically involves adding two or four new timeslots located symmetrically around the central infeasibility, although frequently less timeslots can be added when this expansion would extend outside the timetable.

Timeslot	Mon 10am	Tue 10am	Tue 2pm	Wed 12pm	Thu 9am	Fri 1pm
Shortage	large rooms	large rooms	tut. rooms	all rooms	large rooms	large rooms
Objective	5	14	141	9	8	2
Expansions	1	3	2	0	1	0
Timeslots	5	11	7	3	4	3
Contacts	230	532	366	167	191	146
Variables (k)	27.5	127.0	47.9	11.8	18.3	10.3
Constraints (k)	0.6	2.3	1.2	0.4	0.5	0.3
Solve Time (s)	0.6	4.4	3.4	0.6	0.7	0.3

Table 1: Large Neighbourhood (N1) Method Results

In this problem although there were 13 timeslots (out of 45) which were infeasible, only 6 timeslots were needed to remove the infeasibility. This is because as the neighbourhood expands, it expectedly crosses into other infeasible timeslots.

Timeslot	Mon 10am	Tue 10am	Tue 2pm	Wed 12pm	Thu 9am	Fri 1pm
Shortage	large rooms	large rooms	tut. rooms	all rooms	large rooms	large rooms
Objective	5	16	19	9	13	2
Expansions	1	3	2	0	3	0
Timeslots	5	11	7	3	10	3
Contacts	65	181	178	75	105	18
Variables (k)	4.1	20.9	13.8	2.7	9.6	0.4
Constraints (k)	0.3	1.1	0.6	0.2	0.8	0.1
Solve Time (s)	0.0	0.5	0.4	0.1	0.3	0.0

Table 2: Restricted Neighbourhood (N2) Method Results

For each infeasible timeslot Tables 1 and 2 first state whichever room size or attribute is in shortage, and thus responsible for the infeasibility. Next is the objective value of the solution found after the listed number of expansions (beyond the starting neighbourhood), then timeslots and contacts required in the neighbourhood to reach feasibility. The number of variables and constraints for the largest neighbourhood subproblem solved are also listed. Despite the large size of several of these problems, all N1 and N2 problems were solved in under 5 seconds and 1 second respectively. This is because the structure of the matrix for these problems is naturally integer.

Comparing the methods in terms of the objective function reveals similar values, except for the Tuesday 2pm timeslot where the N1 method has an extremely bad objective. An investigation showed that this is related to its solve in the earlier Tuesday 10am timeslot, whereby a good solution was obtained by adding further congestion to the Tuesday midday timeslots. This meant that solving on Tuesday afternoon in the local 7 timeslots still yielded a feasible solution, but only through significant re-shuffling of contacts of all sizes. The N2 method does not (and cannot) face this situation, as it is only able to move the contacts of relevant size, and will expand the neighbourhood if no feasible solution can be found through moving these at this level of expansion. Although the N2 method resulted in a slightly poorer objective at Tuesday 10am, the solution at Tuesday 2pm is much better.

It would be unfair to state that the extra variables in N1 are a disadvantage, since if the algorithm permitted further expansion of the neighbourhood (i.e. beyond the first feasible solution), the N1 theoretically will find the best possible solution. However, in a practical context, this would create an even larger BIP subproblem. Therefore, given computational limitations, it is more useful to model with only the most relevant variables.

The overall quality of the final timetables, in terms of contact shifts from the original is given in Table 3. Although the data is skewed by one timeslot, it is fair to say that the N2 method of limiting the variables in the neighbourhood is at least non-detrimental to the objective function. At the same time, it is clearly shown to significantly reduce the problem size.

Shift	N1 Method	N2 Method
0 days, 1 period	30	23
0 days, 2 periods	3	3
1 day, 0 periods	8	2
1 day, 1 period	1	1
1 day, 2 periods	23	5
Total Disruption Penalty	179	64

Table 3: Practical Solution Quality

6 Conclusion & Ongoing Research Directions

This paper has outlined an efficient optimisation-based algorithm for minimally modifying a timetable with no feasible room allocation. Although the algorithm has undergone preliminary testing on the University of Auckland data, it would also be worthwhile experimenting with some more heavily infeasible instances. The algorithm may be useful as part of a full automated timetabling solution, because it would allow the timetable generator to create a timetable which is not strictly room-feasible.

The most important result is that the choice of neighbourhood definition (and choice of expansion) can play significant roles in controlling the size of the problem and quality of solution. These have not been fully explored, and are the subject of ongoing research.

In particular, the direction of the neighbourhood expansion appears to be very influential. The neighbourhood could be discouraged from expanding in the direction of other infeasible timeslots, as this causes congestion in the feasible timeslots which makes the later solves more difficult. There are also related issues with the neighbourhood ‘accumulating’ infeasible timeslots as it expands. When this happens it usually necessitates an even greater degree of expansion, and the new infeasibility is at the edge (rather than centre) of the neighbourhood space.

Broader definitions for the neighbourhood and expansion algorithm also need to be considered. An opposite but equally valid approach could be to have a large number of timeslots in the original neighbourhood, but with a greatly reduced number of contacts. The number of free contacts could then be expanded sequentially. Even more ambitious neighbourhood structures could include the movement of entire streams, which is an important detail not captured in the present formulation. Typically it would be more desirable to move all contacts of a stream by a certain amount, rather than just one. Although this will introduce many more timeslots into each neighbourhood, stream considerations are an important next-step in development of the algorithm.

Acknowledgments

Primarily I would like to thank my PhD supervisors and co-contributors Professor David Ryan and Professor Matthias Ehrgott. I would also like to thank Dr Hamish Waterer, Dr Andrew Mason and Dr Michael O’Sullivan for their general assistance.

References

- Beyrouthy, C., E. K. Burke, D. Landa-Silva, B. McCollum, P. McMullan, and A. J. Parkes. 2007. “Towards improving the utilization of university teaching space.” *Journal of the Operational Research Society* 60 (1): 130–143.
- Rezanova, N. J., and D. M. Ryan. 2010. “The train driver recovery problem - A set partitioning based model and solution method.” *Computers & Operations Research* 37 (5): 845–856.

Notes on a UK Water Market Demonstration

John F. Raffensperger
JFR Decision Research, Ltd.
Christchurch, New Zealand
john.raffensperger@canterbury.ac.nz

Darren Lumbroso
HR Wallingford, Ltd.
Howbery Park, Wallingford,
Oxfordshire, United Kingdom

Abstract

In this presentation, we will summarise the development and key lessons of a water market demonstration in the Upper Ouse and Bedford Ouse Catchment in East Anglia, in the UK. This project was done from May to November 2012 (and is being completed as of this writing). We will first describe the development of the hydrological optimization from the source data, including an overview of the key scripts and various complications. Of special interest was the aim to drive the market clearing as directly as possible from source databases, in particular the GIS database of the river catchment. We will describe the linear program used to clear the market, with its strengths and weaknesses.

We will demonstrate the user interface, and we will discuss users' reactions to it, along with the surprising price-finding process that actually occurred. Finally, we will discuss the changes in users' mindset that will likely be required to implement such a market.

Key words: Smart markets, water allocation, linear programming.

Exploring Bi-Objective Column Generation

A. Raith, S. Moradi, M. Ehrgott and M. Stiglmayr
Department of Engineering Science
University of Auckland, New Zealand
and Bergische Universität Wuppertal, Germany
a.raith@auckland.ac.nz

Abstract

The bi-objective simplex method is a well-known technique to solve bi-objective linear programmes (BLP). It is based on the simplex method in iteratively moving between feasible bases. In every iteration of the simplex method a variable with negative reduced cost is selected to enter the basis, and another one leaves the basis, which corresponds to a step towards the optimal solution of a standard linear programme. Column generation is a technique to solve large-scale linear programmes. It has the benefit that not all variables are in the problem formulation initially but instead they are generated by solving a column generation subproblem that finds a negative reduced cost variable to enter the basis. A bi-objective simplex method on the other hand needs to find (all) efficient solutions of the (BLP). This is done by initially finding an optimal solution with respect to the first objective function. The method then iteratively selects a variable to enter the basis with maximum ratio of improvement of the second objective and deterioration of the first one yielding another efficient solution. This continues until all efficient solutions are obtained. We explore the applicability of a column generation approach to the bi-objective simplex method.

Key words: Bi-objective linear programming, column generation, simplex algorithm.

1 Background

Throughout this paper we will consider the following linear programmes. Firstly, we consider a single-objective linear programme (LP)

$$\begin{aligned} \min \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{LP}$$

with variables $x \in \mathbb{R}^n$, and problem parameters $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$. A bi-objective linear programme (BLP) has the form

$$\begin{aligned} \min \quad & Cx = \begin{pmatrix} (c^1)^\top x \\ (c^2)^\top x \end{pmatrix} \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0, \end{aligned} \tag{BLP}$$

where we now have two objective components given by objective matrix $C \in \mathbb{R}^{2 \times n}$. We also write the objective function as two separate components $y^1 = (c^1)^\top x$ and $y^2 = (c^2)^\top x$ with $c^1, c^2 \in \mathbb{R}^n$. The feasible set of both linear programmes is $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b \text{ and } x \geq 0\}$ and the feasible set of (BLP) in objective space is $\mathcal{Y} = \{y \in \mathbb{R}^2 : y = C^\top x \text{ and } x \in \mathcal{X}\}$. In a bi-objective optimisation problem we seek the set of efficient solutions which will be defined in the following.

A point $y \in \mathbb{R}^2$ dominates $\tilde{y} \in \mathbb{R}^2$ if $y \leq \tilde{y}$ which means that

$$y_1 \leq \tilde{y}_1, y_2 \leq \tilde{y}_2 \text{ and } y \neq \tilde{y}.$$

A feasible solution $x^* \in \mathcal{X}$ is efficient if its image Cx^* is not dominated by the image $C\tilde{x}$ of another feasible solution $\tilde{x} \in \mathcal{X}$. The set of efficient solutions is $\mathcal{X}_E \subset \mathcal{X}$ and the corresponding set of nondominated objective vectors is $\mathcal{Y}_N = \{y \in \mathbb{R}^2 : y = Cx, x \in \mathcal{X}_E\}$. When solving (BLP) we want to obtain a complete set of efficient solutions $\bar{\mathcal{X}}_E \subset \mathcal{X}_E$ which contains at least one efficient solution per nondominated objective function vector.

1.1 Bi-objective simplex

The well-known simplex method (Dantzig and Thapa 1997, for example) for Linear Programmes works by iteratively moving from one basic feasible solution to the next. In every iteration of the simplex method a variable with negative reduced cost \bar{c} is selected to enter the basis, and another one leaves the basis, which corresponds to a step towards the optimal solution of an LP.

The bi-objective simplex method initially obtains a solution which is minimal with respect to the first objective component. If there is more than one such solution, the one with minimal second component among them is obtained (which is also known as a lexicographically optimal solution, or lexmin solution in short). To solve (BLP) the entering variable in the simplex is selected to ensure the algorithm moves from one efficient solution to the next. This is illustrated in Figure 1a that shows \mathcal{Y} , the feasible region of (BLP), in objective space and the image of the initial solution.

Then, in each iteration, an entering variable is selected which results in a maximum ratio of improvement of the second objective function over deterioration of the first. Nonbasic variable x_i is chosen to enter the basis if both (1) and (2) hold:

$$\bar{c}_i^2 < 0 \text{ and } \bar{c}_i^1 \geq 0 : x_i \text{ in the basis improves } c^2 \text{ and worsens } c^1, \tag{1}$$

$$i \in \operatorname{argmax} \left\{ \frac{-\bar{c}_i^2}{\bar{c}_i^1 - \bar{c}_i^2} \right\}, \tag{2}$$

where reduced costs have a component \bar{c}^1, \bar{c}^2 for each objective vector and \bar{c}_i^1, \bar{c}_i^2 refers to the reduced cost of x_i . The variable leaving the basis is selected as in the standard simplex method. By moving from one efficient solution to the next a complete set of

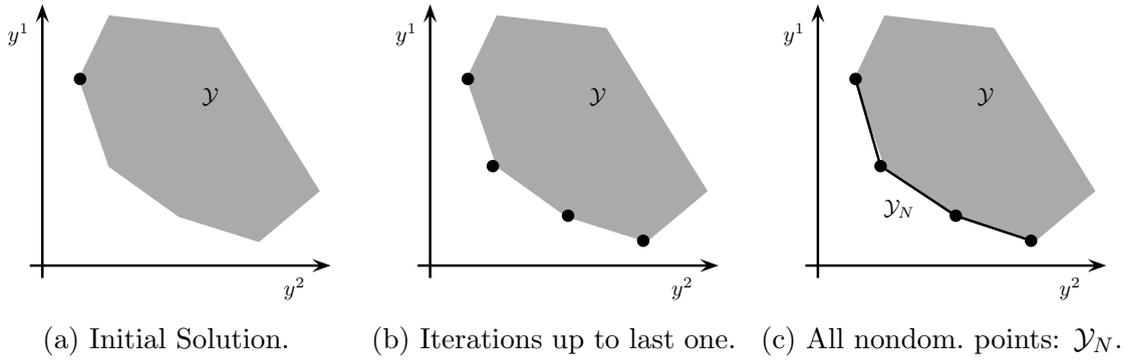


Figure 1: Bi-objective simplex illustration in objective space.

efficient solutions is ultimately obtained when the algorithm arrives at the solution that minimises the second objective function. At this point there are no more variables with negative reduced cost in the second component, see Figure 1b. The obtained complete set of efficient solutions (and nondominated points) consists of the convex combinations of all obtained efficient solutions with neighbouring images in objective space, see Figure 1c. Details on this parametric simplex method for (BLP) can be found in Ehrgott (2005) and on the more general multi-objective version in Evans and Steuer (1973).

It should be noted that no speed-up techniques such as partial pricing can be used in the simplex method for (BLP) as it must be guaranteed that ratio (2) is maximised for an entering variable. Hence, all nonbasic variables must be considered in each simplex iteration. It would be beneficial to improve the efficiency of this step in the bi-objective simplex algorithm.

1.2 Column generation

Lübbecke (2010) provides an excellent introduction to column generation, and we follow the notation in this paper. Again, (LP) is considered (now shown in slightly altered format), which is also referred to as master problem (MP):

$$\begin{aligned} \min \quad & \sum_{j \in \mathcal{J}} c_j x_j \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}} a_j x_j = b \\ & x \geq 0, \end{aligned} \tag{MP}$$

where \mathcal{J} is the set of variables with $|\mathcal{J}| = n$ and a_j are the columns of constraint matrix A in (LP). Instead of (MP) a restricted master problem (RMP) is considered with only a subset $\mathcal{J}' \subseteq \mathcal{J}$ of variables. In the simplex method a nonbasic variable with negative reduced cost enters the basis in each iteration. In column generation a new column (or variable) with negative reduced cost is identified to enter the basis after (RMP) is solved to optimality. If such a column is identified it is included in (RMP) which is re-solved. Otherwise, the optimal solution of (RMP) is also optimal for (MP).

Next we explain how a new column is generated. Once (RMP) is solved to optimality giving solution x^* with dual π^* such a non-basic variable with negative reduced cost \bar{c}_j is found by solving a subproblem (SUB):

$$\begin{aligned} \min \quad & c_j - \pi^* a_j \\ \text{s.t.} \quad & j \in \mathcal{J} \end{aligned} \tag{SUB}$$

If the optimal objective function value of (SUB) is negative for the optimal solution x_{j^*} , this variable is added to (RMP) and it enters the basis of the current solution of (RMP). This process continues iteratively until no more entering variables are identified in (SUB). If the set \mathcal{J} can be described as feasible set $\mathcal{X}_{\mathcal{J}}$ of an optimisation problem instead, (SUB) becomes an optimisation problem:

$$\begin{aligned} \min \quad & c(\lambda) - \pi^* a(\lambda) \\ \text{s.t.} \quad & \lambda \in \mathcal{X}_{\mathcal{J}}, \end{aligned} \tag{SUBopt}$$

where $c_j = c(\lambda_j)$ and $a_j = a(\lambda_j)$ and variables are $\lambda_j \in \mathcal{X}_{\mathcal{J}}$. Hence, the variable to enter the basis is identified by solving (SUBopt) in every simplex iteration for (RMP), and subsequently updating (RMP).

Column generation is especially beneficial if (MP) has many variables (columns) and (SUB) is relatively easy to solve. (SUB) is often a shortest path problem, for example when solving multi-commodity network flow problems (Ahuja, Magnanti, and Orlin 1993), a resource constrained shortest path problems, for example in airline scheduling problems (Klabjan 2005), or a knapsack problem (Glimore and Gomory 1961).

2 Bi-objective column generation

In order to incorporate column generation concepts from Section 1.2 into the bi-objective simplex from Section 1.1 the main concern is to ensure that the entering variable that is selected in every iteration satisfies equations (1) and (2). This question will be addressed in Section 2.1. We present a bi-objective version of the simplex algorithm with column generation in Section 2.2.

2.1 Bi-objective column generation subproblem

We start by defining the bi-objective restricted master problem (BRMP)

$$\begin{aligned} \min \quad & \begin{pmatrix} \sum_{j \in \mathcal{J}'} c_j^1 x_j \\ \sum_{j \in \mathcal{J}'} c_j^2 x_j \end{pmatrix} \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}'} a_j x_j = b \\ & x \geq 0, \end{aligned} \tag{BRMP}$$

where \mathcal{J}' is again a subset of the full set of variables \mathcal{J} of (BLP). We assume that the current basic feasible solution of (BRMP) is efficient. This means we do not (yet) consider the case of finding the first efficient solution, which is a lexmin solution as outlined in Section 1.1 here. When applying the concept of column generation to the parametric simplex for (BLP) the most important aspect is that there must be a column generation subproblem that is capable of identifying an entering variable with maximum ratio (2). This will ensure that the algorithm moves from one efficient solution to the next one rather than to a dominated solution.

We assume that the current efficient (basic feasible) solution of (BRMP) is x^* with dual variables π^1, π^2 and reduced cost \bar{c}^1, \bar{c}^2 . For simplicity, we write the bi-objective equivalent (BSUB) of the column generation subproblem (SUB) in terms of the reduced costs \bar{c}^p instead of its equivalent $c^p - \pi^p a_j$ for both objective components

$p = 1, 2$. (BSUB) simply states conditions (1) and (2):

$$\begin{aligned} \max \quad & \frac{-\bar{c}_j^2}{\bar{c}_j^1 - \bar{c}_j^2} \\ \text{s.t.} \quad & \bar{c}_j^2 < 0 \text{ and } \bar{c}_j^1 \geq 0 \\ & j \in \mathcal{J}. \end{aligned} \tag{BSUB}$$

We can rewrite the fractional single-objective problem (BSUB) as a bi-objective problem (BSUB')

$$\begin{aligned} \min \quad & \begin{pmatrix} \bar{c}_j^2 \\ \bar{c}_j^1 - \bar{c}_j^2 \end{pmatrix} \\ \text{s.t.} \quad & j \in \mathcal{J}. \end{aligned} \tag{BSUB'}$$

It should be noted that problems (BSUB) and (BSUB') would be rewritten as proper optimisation problems in a column generation context, similar to (SUBopt). We omit this here and in the following to simplify notation.

Firstly, we establish that the optimal solution of (BSUB) is among the efficient solutions of (BSUB').

Theorem 2.1 *The optimal solution of (BSUB) is among the efficient solutions of (BSUB') provided that x^* is an efficient solution of (BRMP) (and of (BLP)) and $\bar{c}^p, p = 1, 2$ are the corresponding reduced cost vectors.*

Proof We assume on the contrary that an optimal solution $j^* \in \mathcal{J}$ of (BSUB) exists which is not efficient for (BSUB'). This means for index j^* that there exists another index $i \in \mathcal{J}, i \neq j^*$ such that

$$\bar{c}_i^2 \leq \bar{c}_{j^*}^2 \quad \text{and} \quad \bar{c}_i^1 - \bar{c}_i^2 \leq \bar{c}_{j^*}^1 - \bar{c}_{j^*}^2, \tag{3}$$

with at least one strict inequality. Note that from $\bar{c}_{j^*}^2 < 0$ in (BSUB) it follows that $\bar{c}_i^2 \leq \bar{c}_{j^*}^2 < 0$. Also, as x^* is an efficient solution of (BLP), we know that the reduced cost cannot be negative in both components, i.e. we cannot have $\bar{c}_i^1 < 0$ and $\bar{c}_i^2 < 0$ as this would contradict the efficiency of x^* . Hence $\bar{c}_i^1 \geq 0$ as we know that $\bar{c}_i^2 < 0$. Consider

$$\frac{-\bar{c}_{j^*}^2}{\bar{c}_{j^*}^1 - \bar{c}_{j^*}^2} \stackrel{(3)}{\leq} \frac{-\bar{c}_i^2}{\bar{c}_{j^*}^1 - \bar{c}_{j^*}^2} \stackrel{(3)}{\leq} \frac{-\bar{c}_i^2}{\bar{c}_i^1 - \bar{c}_i^2}, \tag{4}$$

with at least one of these two inequalities being strict. This contradicts the optimality of j^* in (BSUB). \square

From Theorem 2.1 we conclude that, if the current solution x^* is efficient for (BRMP) and (BLP) it is sufficient to identify all efficient solutions of (BSUB') in order to find the maximum ratio in (BSUB). It is possible to further characterise which efficient solutions of (BSUB') need to be considered.

We observe that for all possible combinations of values of $\bar{c}^p, p = 1, 2$ the corresponding images of efficient solutions of (BSUB') are located in three quadrants of objective space, as highlighted in gray in Figure 2. We note that the solutions that are candidates to enter the basis are those with $\bar{c}^2 < 0, \bar{c}^1 \geq 0$ and hence $\bar{c}^2 < 0, \bar{c}^1 - \bar{c}^2 > 0$ in Figure 2. In the following we restrict our considerations to this area of objective space. The ratio in (BSUB) is represented by contour lines which are straight lines through the origin as can be seen in Figure 3, and this ratio increases with increasing slope of the contour line.

It follows that the maximum ratio (BSUB) corresponds to the contour line with maximum slope, hence a tangent to the feasible set in objective space, $\mathcal{Y}_{\text{BSUB}}$ (a

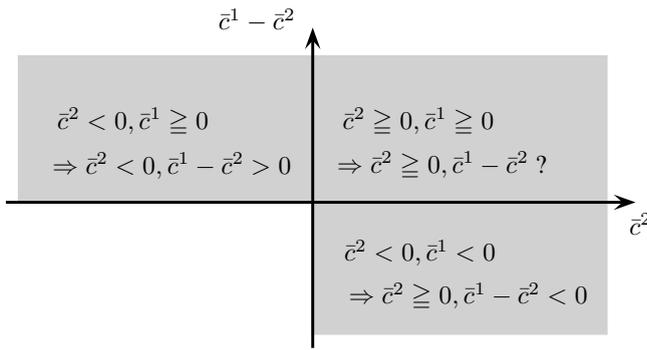


Figure 2: Location of images of efficient solutions of (BSUB') in objective space.

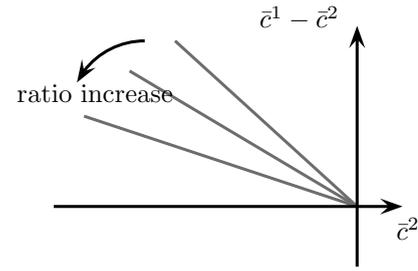


Figure 3: Contour lines of ratio in (BSUB) in objective space of (BSUB').

formal proof is omitted here). This is illustrated in Figure 4, where the corresponding efficient solution (as represented by its nondominated image) is unique. This nondominated point may not be unique if the tangent intersects with a facet of $\mathcal{Y}_{\text{BSUB}'}$. We note that in both cases an extreme point of $\mathcal{Y}_{\text{BSUB}'}$ is among the maximisers of (BSUB). From this, we can conclude that it is not required to find all efficient solutions of (BSUB'), instead we can start by minimising \bar{c}^2 and generating solutions with increasing value of \bar{c}^2 until a tangent of an extreme point of $\mathcal{Y}_{\text{BSUB}'}$ goes through the origin.

Theorem 2.2 *Let (BSUB') be a bi-objective linear programme and let x^* be an efficient solution of (BRMP) and of (BLP) and $\bar{c}^p, p = 1, 2$ the corresponding reduced cost vectors. Efficient solutions of (BSUB') can be calculated iteratively starting from the solution with minimum \bar{c}^2 component and then increasing this component. The maximum ratio of (BSUB) is obtained once a tangent of an extreme point of $\mathcal{Y}_{\text{BSUB}'}$ goes through the origin.*

Proof As (BSUB') is a bi-objective linear programme, $\mathcal{Y}_{\text{BSUB}'}$ is a convex set. Efficient solutions can be ordered such that their images in objective space $y_1, y_2, y_3 \dots$ have increasing first component y_i^1 and decreasing second component y_i^2 :

$$y_1^1 < y_2^1 < y_3^1 < \dots \quad \text{and} \quad y_1^2 > y_2^2 > y_3^2 > \dots$$

This can be seen in Figure 4. All of the above implies that the slopes $m_i = \frac{y_{i+1}^2 - y_i^2}{y_{i+1}^1 - y_i^1}$ of the facets between points y_i and y_{i+1} are increasing.

We consider all tangents to the set $\mathcal{Y}_{\text{BSUB}'}$ in an extreme point y_i , where tangents are supporting hyperplanes of the set. For each extreme point y_i define a cone \mathcal{C}_i obtained by considering straight lines from each extreme point along a possible tangent towards the direction where the tangents intersect the y^2 -axis (where the y^2 axis represents objective $\bar{c}^1 - \bar{c}^2$), as shown in Figure 5. We call the intersection of an extreme ray of \mathcal{C}_i and the y^2 -axis a y^2 -intercept. Due to the convexity of $\mathcal{Y}_{\text{BSUB}'}$ the y^2 -intercepts of the extreme rays of cones \mathcal{C}_i increase until the y^2 -intercepts ultimately become non-negative, as illustrated in Figure 5. The efficient solution corresponding to the extreme point y_i with cone \mathcal{C}_i that has one negative and one non-negative y^2 -intercept attains the maximum ratio (BSUB) as this cone represents point y_i with a tangent through the origin. If the y^2 -intercepts of the cone are

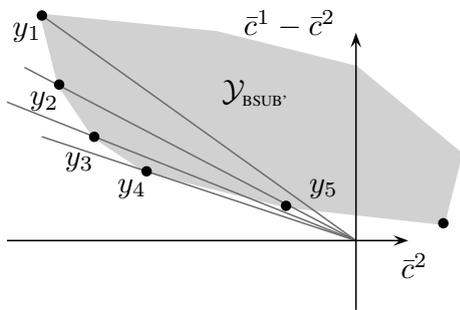


Figure 4: Maximum ratio among images of efficient solutions of (BSUB') in objective space.

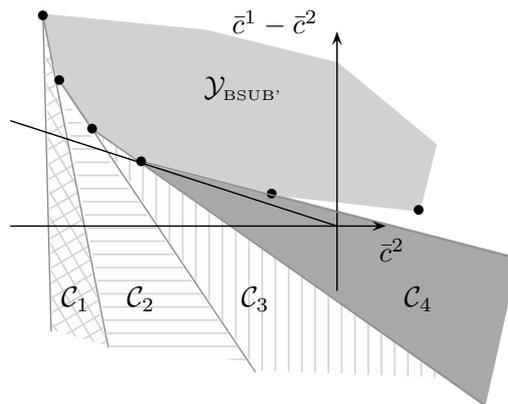


Figure 5: Cones \mathcal{C}_i for feasible set of (BSUB') in objective space.

negative and positive the maximum ratio is unique, whereas a negative and a zero y^2 -intercept indicate that the solution is non-unique. \square

Corollary 2.1 *The maximum ratio is obtained for an extreme efficient solution of (BSUB').*

Proof This follows along the lines of the proof of Theorem 2.2 as the tangent to the set $\mathcal{Y}_{\text{BSUB}'}$ intersects the set in either one extreme point of $\mathcal{Y}_{\text{BSUB}'}$ or in a whole facet, in which case the intersection contains two extreme points both of which maximise the ratio (BSUB). \square

2.2 Bi-objective column generation algorithm

Algorithm 1 outlines the bi-objective column generation principle. Step 5 requires solving (BSUB'), which may be done by finding all of its efficient solutions according to Theorem 2.1. Alternatively, one can also solve (BSUB') in a “left-to-right” manner until it is guaranteed that the maximum ratio is attained as in Theorem 2.2. It remains to show that Algorithm 1 does generate a complete set of efficient solutions of (BLP).

Theorem 2.3 *The efficient basic feasible solutions of (BRMP) obtained by Algorithm 1 are efficient basic feasible solutions of (BLP). The set \mathcal{E} together with the set of convex combinations of $x, \bar{x} \in \mathcal{E}$ where x, \bar{x} have neighbouring images (as in Figure 1c), is a complete set of efficient solutions.*

Proof We show this by induction. Base case where $i = 1$: The lexmin solution of (BLP) and (BRMP) is obtained using standard column generation and hence optimality (with respect to the problem with a single objective c^1) follows from standard column generation theory. (To reach this lexmin solution once an optimal solution with respect to c^1 is obtained, we follow up by some simplex iterations in direction of c^2 that do not worsen c^1 .) This lexmin solution $x^{(1)} = x_{\text{lex}}$ is efficient.

Now assume that $x^{(i)}$ is an efficient solution of (BRMP) and (BLP) and show that $x^{(i+1)}$ is also efficient, where $x^{(i)}$ denotes the i^{th} solution obtained by Algorithm 1. By contradiction we also assume that $x^{(i+1)}$, obtained by letting a variable with

Algorithm 1 Bi-objective simplex using column generation

- 1: **input:** A, b, C for (BLP).
 - 2: Obtain lexmin solution x_{lex} with standard column generation at which point we have (BRMP) with variables \mathcal{J}' .
 - 3: Calculate corresponding duals $\pi^p, p = 1, 2$; set $\mathcal{E} = \{x_{\text{lex}}\}$.
 - 4: **repeat**
 - 5: Solve (BSUB') for $\pi^p, p = 1, 2$.
 - 6: **if** New entering variable j^* is identified **then**
 - 7: Add x_{j^*} to (BRMP) and perform one simplex iteration obtaining x^* .
 - 8: Update duals $\pi^p, p = 1, 2$ based on x^* ; $\mathcal{E} = \mathcal{E} \cup \{x^*\}$.
 - 9: **else**
 - 10: Stop.
 - 11: **end if**
 - 12: **until** No entering variable is identified.
 - 13: **output:** Efficient solutions \mathcal{E} identified. Other efficient solutions are convex combinations of efficient solutions with neighbouring images in objective space.
-

index i^* that maximises (BSUB) enter the basis, is not efficient in (BLP). Note that by construction $x^{(i+1)}$ will be efficient in (BRMP).

There exists an efficient feasible solution $\tilde{x} \in \mathcal{X}$ of (BLP) whose image dominates that of $x^{(i+1)}$. There must be a basic feasible solution that is either equivalent to \tilde{x} (i.e. has the same image), or \tilde{x} is a convex combination of basic feasible solutions. We denote by \hat{x} this basic feasible solution in the first case, and the ‘‘left-most’’ (with smaller c^1 component) of the two solutions in the second case. It must be possible to reach \hat{x} through a sequence of simplex pivots. WLOG assume \hat{x} can be reached from $x^{(i)}$ in a single pivot (if not rename \hat{x} to be the first such basic feasible solution), and j^* is the variable entering the basis to get from $x^{(i)}$ to \hat{x} .

Note that efficient solutions of (BLP) are connected by sequences of simplex pivots, where each of the solutions in the sequence is efficient itself (Steuer 1985). With reduced costs $\bar{c}^p, p = 1, 2$ for solution $x^{(i)}$, we can distinguish three cases for the ratio in (BSUB), see also Figure 6a-6c:

1. $\frac{-\bar{c}_{j^*}^2}{\bar{c}_{j^*}^1 - \bar{c}_{j^*}^2} > \frac{-\bar{c}_{i^*}^2}{\bar{c}_{i^*}^1 - \bar{c}_{i^*}^2}$: This contradicts the optimal choice of entering variable with index i^* as maximiser of (BSUB).
2. $\frac{-\bar{c}_{j^*}^2}{\bar{c}_{j^*}^1 - \bar{c}_{j^*}^2} < \frac{-\bar{c}_{i^*}^2}{\bar{c}_{i^*}^1 - \bar{c}_{i^*}^2}$: Solutions $x^{(i)}$ and \tilde{x} must be connected by a sequence of basic feasible solutions that are efficient (Steuer 1985). In case 2 the solution obtained when j^* enters the basis of solution $x^{(i)}$ is dominated by a convex combination of $x^{(i)}$ and $x^{(i+1)}$, a contradiction.
3. $\frac{-\bar{c}_{j^*}^2}{\bar{c}_{j^*}^1 - \bar{c}_{j^*}^2} = \frac{-\bar{c}_{i^*}^2}{\bar{c}_{i^*}^1 - \bar{c}_{i^*}^2}$: The solution obtained when j^* enters the basis of solution $x^{(i)}$ is equivalent to a convex combination of $x^{(i)}$ and $x^{(i+1)}$. It follows that $C\hat{x}$ cannot be an efficient solution of (BLP) as it does not lie on the boundary of any \mathbb{R}_+ -convex set containing points $Cx^{(i)}$ and $C\tilde{x}$ (which dominates $Cx^{(i+1)}$) on its boundary, a contradiction.

We have now established that all extreme efficient solutions of BLP can be found using Algorithm 1. It follows that a complete set of efficient solutions is obtained

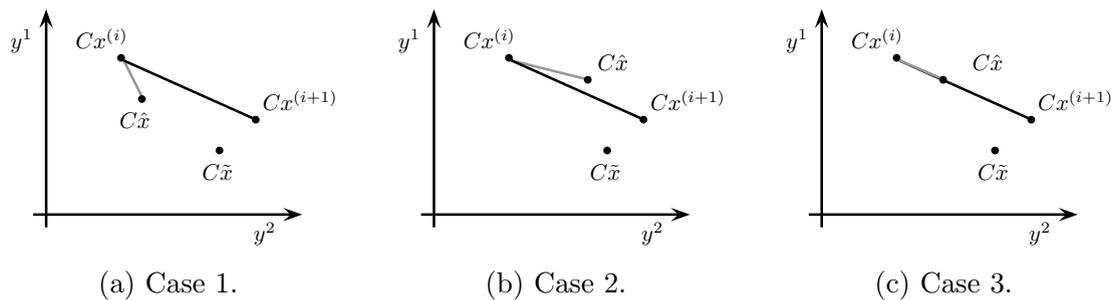


Figure 6: Cases in proof of Theorem 2.3

from the extreme solutions as set of convex combinations of all extreme efficient solutions with neighbouring images in objective space. \square

3 Conclusion and outlook

We present the first steps towards integrating column generation in the solution of bi-objective linear programmes. We show that it is possible to formulate a column generation subproblem, (BSUB) with fractional objective, that allows to obtain an entering variable. We show this subproblem can be replaced by finding all (or some) of the extreme efficient solutions of the bi-objective problem (BSUB'). We outline a column generation algorithm, Algorithm 1, for solving (BLP).

In the future we want to extend this work by testing the proposed approach on (BLP) instances. We will investigate the difficulty of the arising subproblems (BSUB'). Subproblems are often integer or combinatorial problems, and we will extend our approach those types of problems.

References

- Ahuja, R.K., T.L. Magnanti, and J.B. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall.
- Dantzig, G.B., and M.N. Thapa. 1997. *Linear Programming 2: Theory and Extensions*. Springer Series in Operations Research. Springer.
- Ehrgott, M. 2005. *Multicriteria Optimization*. Second Edition. Springer.
- Evans, J.P., and R. Steuer. 1973. "A revised simplex method for linear multiple objective programs." *Mathematical Programming* 5:54–72.
- Glimore, P.C., and R.E. Gomory. 1961. "A linear programming approach to the cutting-stock problem." *Operations Research* 9:849–859.
- Klabjan, D. 2005. Chapter Large-scale models in the airline industry of *Column Generation*, edited by G. Desaulniers, J. Desrosiers, and M.M. Solomon, 163–195. New York: Springer.
- Lübbecke, M. 2010. Chapter Column Generation of *Wiley Encyclopedia of Operations Research and Management Science*, edited by J.J. Cochran, 1–14. John Wiley & Sons, Inc.
- Steuer, R.E. 1985. *Multicriteria Optimization - Theory, Computation and Application*. Wiley.

Generalized CDDP for Reservoir Management

Rosemary A. Read, Shane Dye, E. Grant Read
Department of Management
University of Canterbury
New Zealand
rar78@uclive.ac.nz

Abstract

Constructive Dual Dynamic Programming (CDDP) has been successfully applied to reservoir management problems in hydro-dominated electricity sectors. In New Zealand, the SPECTRA and DUBLIN models use a two-reservoir CDDP as the basis for release decisions.

While the ‘curse of dimensionality’ ultimately applies, the computational efficiency of SPECTRA suggests that higher dimensional problems should be solvable. To date, however, such generalization has been inhibited by the two-dimensional representation of operational “guidelines” or marginal water surface contours used by SPECTRA. This two-dimensional representation is also quite specific to the reservoir configuration assumed in the New Zealand system. In this paper, we describe a generic approach to CDDP, which is more readily generalizable to alternative configurations and higher dimensions. We briefly describe its application to an upstream-downstream, two-reservoir configuration.

Key words: CDDP, Reservoir Management, Hydro-electric generation, SPECTRA

1 Introduction

In hydro-dominated electricity sectors reliability of supply and electricity pricing can rely heavily on effective release decisions based on solutions to the reservoir-management problem. A number of techniques have been applied to this problem, including Stochastic Dual Dynamic Programming (SDDP) and Constructive Dual Dynamic Programming (CDDP) (Periera & Pinto, 1991, Read & Hindsberger, 2010). SDDP represents future costs as piecewise linear functions, and uses these instead of enumerating the entire feasible region. CDDP rapidly produces useful release “guidelines” over the entire feasible region of reservoir level combinations. In principle the approach is very flexible, because any solution approach can be taken to the intra-period problem, provided it produces a monotone “demand curve (or surface) for release”. For example, Scott and Read (1996) applied a Cournot gaming model. And, in principle, the higher level inter-period problem can be solved very efficiently, since it only needs to produce primal “guideline” solutions corresponding to critical MWV levels, or relationships.

So far implementations have been limited to dealing with two-reservoir systems. Although the ‘curse of dimensionality’ must ultimately apply to any DP-based technique, the efficiency of these two reservoir implementations suggests that higher

dimensional problems should be solvable. The limitation, to date, has related more to the difficulties of conceptualising higher dimensional generalisations of the “guideline diagram” representation central to these models. The SPECTRA model, in particular, uses a representation which is quite specific to an assumed configuration of the New Zealand power system.

This first application of CDDP proved particularly useful while electricity generation decisions were centralized, and is still used by a number of major electricity companies in New Zealand. It allows for two reservoirs, each in a different island, with a limited HVDC link between the two islands. Losses on the link are accounted for in the CDDP implementation. SPECTRA also assumes alternative sources of electricity with constant marginal costs, in the North Island only, while South Island demand can either be met, or shortage will occur. In a preliminary benchmarking exercise against SDDP, SPECTRA was discovered to have very rapid computational times and created useful two-dimensional Marginal Water Value (MWV) surfaces closely approximating all possible reservoir storage level combinations (Miller, 2009).

The on-going usefulness of the SPECTRA model has been limited, though, by its lack of adaptability. One limitation has been its lack of intra-island transmission system modelling, but that limitation could be addressed by generalising the intra-period pre-computations, without fundamentally changing the way the CDDP implementation works. Higher reservoir numbers, and alternative reservoir configurations, cannot be accounted for by the current implementation. The RAGE/DUBLIN model takes a very different (gaming) approach to intra-period pre-computation, and accounts for risk aversion, but is still limited to dealing with the same reservoir configuration as SPECTRA (Read & Hindsberger, 2010).

These limitations were taken into consideration in the development of ECON BID for the Scandinavian industrial context. That model iteratively solves a number of two-reservoir stochastic problems, for every iteration, the release from a single reservoir alongside the sum of all other reservoir releases is considered (Read & Hindsberger, 2010). The iterative process, however, limits the accuracy of the MWV surfaces produced, since the optimal release decision for a given reservoir may be different for different distributions of the stored water in the rest of the system. Thus ECON BID, while a useful tool, does not fully exploit the potential of CDDP.

This paper describes a first step towards re-casting CDDP concepts in a way which is more readily generalizable, while limiting the associated increase in computational burden, as far as possible. The central concept is that release and storage surfaces can be defined in terms of a discrete set of critical marginal water values (MWVs) and MWV combinations, with associated storage and release values. Our goal has been to re-conceptualise the way SPECTRA deals with these surfaces, to allow generalisation to other system configurations, involving two or more reservoirs. SPECTRA is a stochastic model, and we have plans to improve both the accuracy and efficiency of SPECTRA’s treatment of uncertainty. This initial project deals only with the deterministic case. Models for basic single reservoir, dual reservoir and three reservoir cases have been implemented. Complex two-reservoir configurations including multiple efficiencies, losses, and region specific load requirements were also explored in the context of the project. However, here we focus mainly on the application of these concepts to a general upstream-downstream two reservoir case.

2 Conceptual Basis

The underlying problem for reservoir management in the electricity sector is that the demand for electricity must be met from a variety of generation sources, with varying costs. The total demand varies over time. For a given time period, there exists a load duration curve (LDC) which represents the fraction of time that the demand for electricity is expected to exceed any particular level. We assume that this demand is met by “filling the LDC” in “merit order”. That is, the load is met by the least expensive (lowest short run marginal cost) generation source first, as shown in the Figure 2 below. Determining the portion of the load to be met by hydro-electric generation is more difficult.

Water is a resource with no explicit cost so the cost associated with hydro-electric generation is the opportunity cost of being unable to use the same water for generation at a later point to replace more expensive thermal generation, or avoid shortage. Where there is no storage capacity, water cannot be used for later generation, hence there is no significant opportunity cost. In this circumstance hydro-generation is generally the least expensive energy source and so lowest in the merit order. Stations are scheduled from lowest to highest in the merit order. However, if a unit of water could be stored to reduce the cost of filling the LDC some later period by \$1000, then it should not be released now, unless it can be used to reduce the cost of filling the LDC in the current period by \$1000 or more. Thus the MWV is not constant. It depends very much on ever-changing circumstances, and must be determined by a long term optimisation model, which takes all future opportunities into account.

This valuation of water is represented through Marginal Water Value surfaces which we refer to as the Demand Surface for Release (DSR), and the Demand Surface for Storage (DSS). These surfaces correspond to multi-dimensional demand curves for release and storage. The DSR for a given time period represents the marginal value of using an additional unit of water for generation to displace a generation source with a higher marginal cost, or reduce shortage. The DSR could be stored as discrete release quantities, with their associated marginal values, as in the single reservoir model of Starkey et al. (2012). In higher dimensions, though, it is more efficient to store the DSR as a set of critical marginal water values, with their associated cumulative release quantities, as described in section 2.1 below. These cumulative DSR quantities represent the maximum quantity that it is worthwhile to release for each associated MWV.

The DSS similarly associates each additional unit of water with the marginal value obtainable by storing that water to displace generation in future time periods. In the deterministic case the value assigned to any increment of water in the DSS is determined by the value that increment will have in the DSR for some future period. The stochastic case the DSS value would differ in that it would be the expected value that increment may have in future period DSRs. So for deterministic inflows, we basically need to add up those future DSRs to form the DSS for any period. But we must also account for the fact that demand that will be met by future inflows need not be met by stored water. Also, reservoir limits imply that future demand cannot be met by water that cannot physically be stored. In practice, then, the DSS is formed as a sum of inflow adjusted DSRs for later time periods, truncated so as to lie within feasible limits. Bellman’s principle of optimality applies, if we assume that inflows, and benefits, are independent between periods. So the DSS for any period can be formed from the DSR for the current period and the DSS for the following period.

For the final time period, we assume a DSS representing the MWV of water stored beyond the planning horizon. CDDP is then solved using backwards induction. A general algorithm is stated on the LHS of the diagram below (Figure 1), and illustrated by the one dimensional diagrams on the RHS, where the “surfaces” (DSR and DSS) are one dimensional “curves”. Step 1 shows the interim DSS' for the beginning of the current period being formed by adding the intra-period DSR for that period, to the DSS for the end of the time period. That is, we add the recommended cumulative release and end-of-period storage quantities, at each MWV level, to obtain the recommended cumulative beginning-of-period storage quantity for that MWV. In steps 2 and 3, the DSS is adjusted for the expected inflow, and truncated to form the final DSS for the beginning of the period, and hence for the end of the previous period.

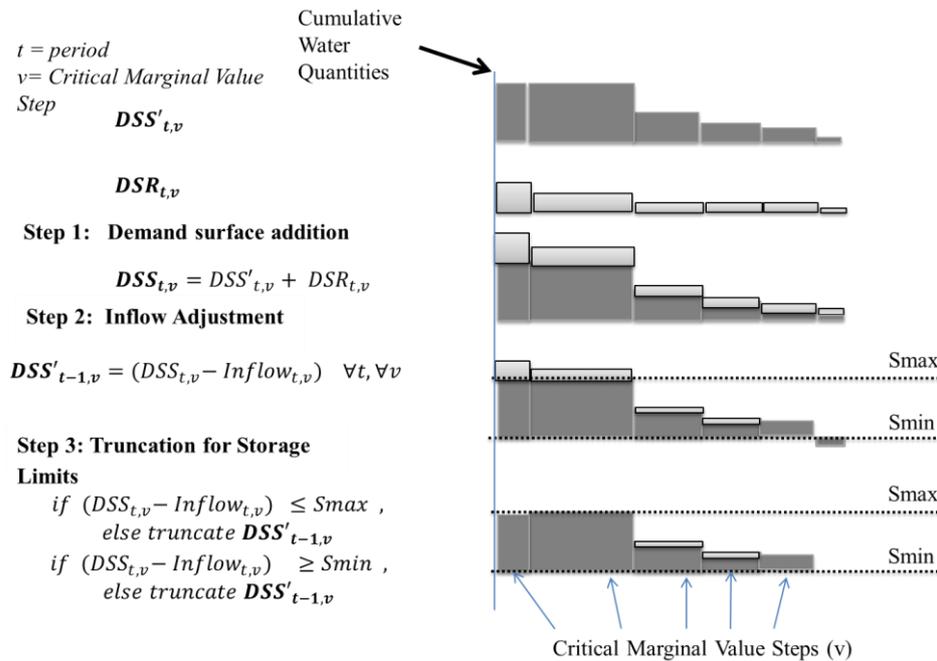


Figure 1: CDDP algorithm for forming Demand Surfaces for Storage.

Essentially the same process applies in higher dimensions, using n n -dimensional DSSs, representing the MWV of water stored in each reservoir, as a function of all reservoir levels. The computation required for each point is trivial, but the key source of computational burden is the number of points required to represent the surfaces. Rather than store, and update, n MWVs for each point in an n -dimensional grid, the algorithms described here only store and update the storage vectors corresponding to a limited set of critical MWV levels, and MWV ratios. SPECTRA, for example, formed North Island “storage guidelines”, along which the North Island MWV equalled critical North Island marginal cost levels, and South/North import/export transfer “guidelines” along which the South Island MWV equalled the North Island MWV +/- marginal HVDC losses. We have generalised this approach to deal with a range of two and three-dimensional cases incorporating upstream-downstream reservoirs, multiple efficiencies, losses, and region specific load requirements. Here we will first discuss a generic two-dimensional problem and then deal with the specific application to the upstream-downstream case.

2.1 Pre-computation for the Intra-period Problem

In order to allow the CDDP algorithm to be more readily generalizable to higher dimensionality, DSRs are pre-computed and representations are stored in a more compact form than in previous CDDP implementations. Examples of earlier

representations include SPECTRA; in which the “guidelines” for critical North Island marginal cost levels are still defined over an arbitrary grid of South Island storage levels; and that in Dye et al. (2012), where marginal water values are associated with a fine grid of regularly sized discrete water quantities. Here we only compute release and storage levels for a discrete set of critical MWVs, and MWV combinations.

For simplicity, we assume that reservoir storage is measured in energy terms. Ignoring Hydro 1, Figure 2 illustrates how release from Hydro 1 will decrease when the MWV of Hydro 1 rises above the marginal cost of a particular thermal generator. Clearly, each generator marginal cost is a critical MWV, at which the quantity of water released will change. At the critical MWV, any combination of the two sources which produces the same total generation has the same cost, and any intermediate release level can be optimal. Thus, increasing MWV from zero to infinity so that reservoir release starts at the bottom of the LDC and moves up to the top of the LDC, produces a monotone decreasing DSR, consisting of a series of steps, as in Figure 1 above.

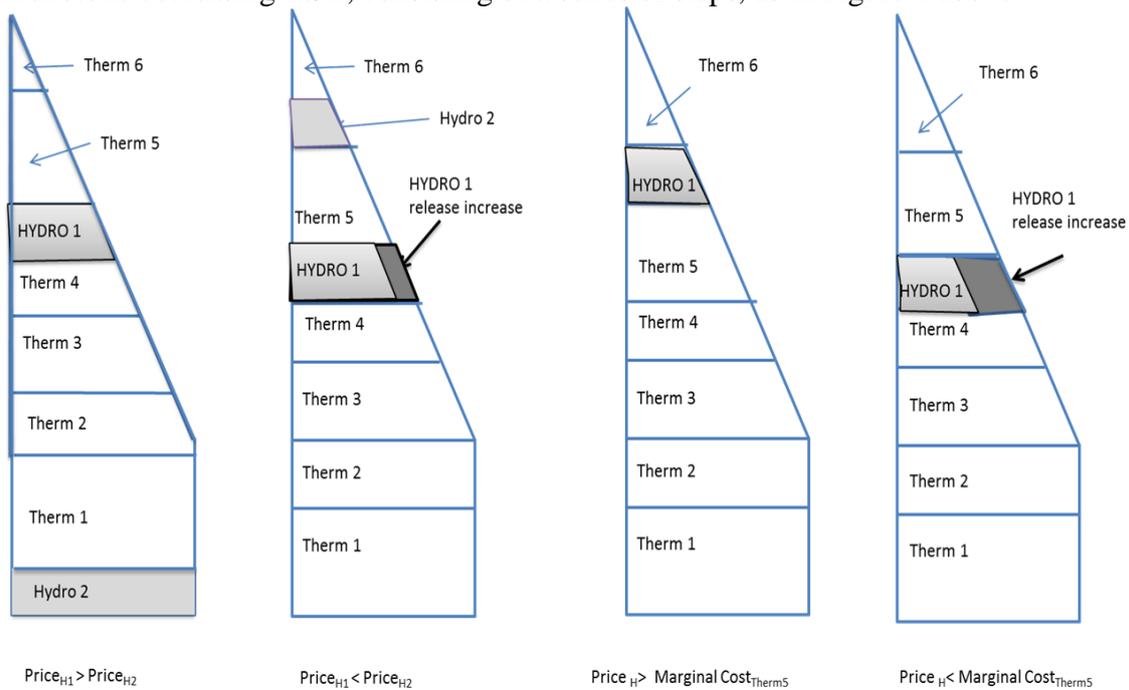


Figure 2: Single reservoir LDC Fill, and Double reservoir LDC Fill

Where there are two reservoirs, the release quantity for each reservoir will depend on whether its MWV is above or below that of the other reservoir, as in Figure 2 above. But it does not matter how much the MWVs differ by. So we only need to form two DSRs for each reservoir; one giving a set of “minimum” releases, assuming that the other reservoir is the bottom of the merit order; and the other giving a set of “maximum” releases, assuming that it is at the top. Ignoring transmission losses and assuming constant generation efficiency from each reservoir, storage and release can be measured in energy terms, and here we have assumed the critical MWV ratio is 1. For higher ratios the minimum DSR will apply for one reservoir and the maximum for the other. For lower ratios, the situation will be reversed. At the critical ratio, we are indifferent between release from the two reservoirs, and a range of intermediate solutions may be optimal, provided the same total energy is produced. This implies the compact representation of the DSR in terms of reservoir “guidelines” shown in Figure 3 below.

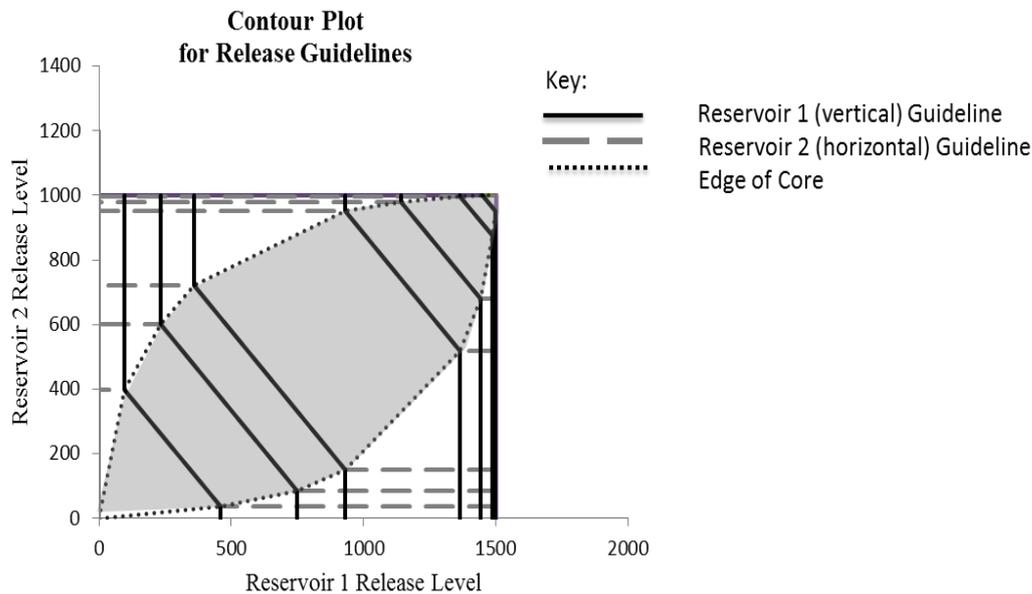


Figure 3: Surface and Contour Plots for Reservoir Release “guidelines”.

The contour plot in Figure 3 shows “release guidelines” corresponding to critical MWV levels on DSRs for two reservoirs, in decreasing order from the origin. The vertical “guidelines” represent contours of the DSR for reservoir 1, corresponding to release combinations where MWV1 is held constant at a critical level, while MWV2 varies. The horizontal “guidelines” represent contours of the DSR for reservoir 2, corresponding to storage combinations where MWV2 is held constant at a critical MWV, while MWV1 varies. The shaded area is the area in which the MWVs for both reservoirs are at the critical ratio, and an area of indifference occurs. The diagonal line segments crossing this “core” area are common to the “guidelines” for both. Outside that zone, all release “guidelines” are horizontal/vertical because the LDC area covered by release from each reservoir, does not change as the MWV of the other reservoir moves further away from the critical ratio.

Thus, this pair of DSRs can be entirely defined by the list of critical MWVs, and the set of points delineating each edge of the core, from which the entire grid of intersecting “guidelines” may be inferred. The upstream/downstream case is discussed further below, but the above discussion applies to any set of parallel reservoirs. But we may need several critical ratios, and associated “cores”, if we wish to model the impact of intervening transmission losses and/or the way in which generation efficiency falls with increasing release. The cores will be n -dimensional.

2.2 Recursive Solution of the Inter-period Problem

CDDP forms a DSS for each period, working backwards recursively from the end of the planning horizon. We first form a set of critical price vectors, from a master list of all critical MWV levels and ratios that may occur over the planning horizon. The detail of the MWV surfaces does not actually matter, between these critical levels/ratios because (by construction) it has no impact on the release. But, ignoring wastage, holding costs and discounting, no other MWV levels can actually occur, in this deterministic case, because each unit of water will eventually be used to satisfy demand at one of these

marginal cost levels. Rather than determine which level that will be, for some arbitrary set of storage points, the deterministic CDDP algorithm determines the boundaries of the set of storage pairs over which each possible MWV level applies. To do this, it only needs to determine the storage pair corresponding to each critical price pair, recursively, for each period. Figure 1 above, describes this process generically, in terms of adding “curves”. But our algorithm reduces this to a simple point-wise addition:

$$DSS_{t,d,v_1,v_2} = DSS'_{t,d,v_1,v_2} + DSR_{t,d,v_1,v_2}$$

Contour Plot for Reservoir Storage Guidelines

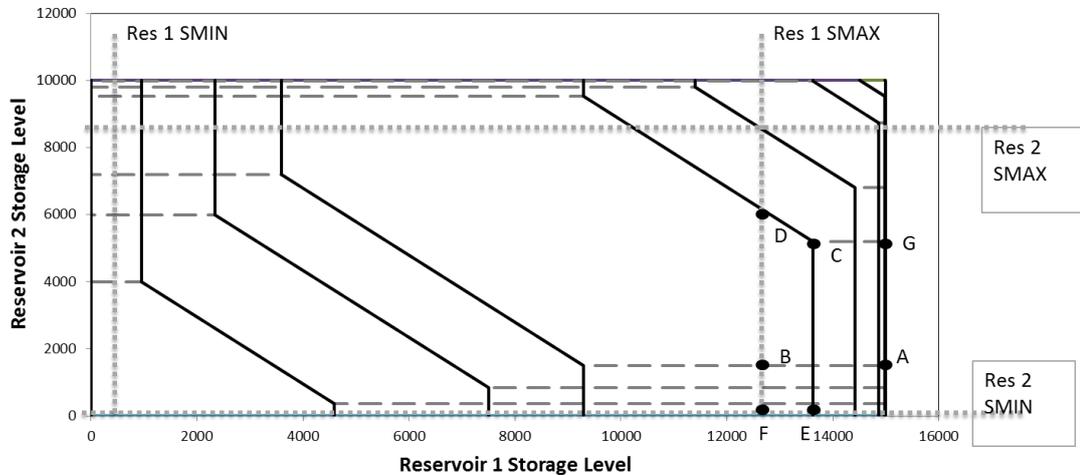


Figure 4: Truncation for Reservoir Storage “guidelines”

Applying this equation will produce a set of critical beginning-of-period storage pairs, defining an interim DSS over an expanded storage range, offset by the expected inflow pair, as shown in Figure 4 above. But points outside the feasible storage range represent solutions that, while desirable, are not possible. They imply storing water for future use that cannot be physically held, or drawing down water for current usage that is not yet there. Thus these interim “guidelines” must be truncated to respect storage limits, in a way that preserves both MWVs accurately, on the bounds.

Two cases are displayed in the diagram above. Point A simply moves to point B. This sets reservoir 1 storage for the end point of the corresponding horizontal (reservoir 1) “guidelines” to its maximum, and moves the remaining section of the corresponding vertical (reservoir 2) “guidelines” to the bound. But point C, and points such as G that are further along the “horizontal guideline”, move diagonally to point D, which now defines the point on the edge of the core, at which MWV2 and MWV1 both equal the critical MWV for this “guidelines”. The remainder of the corresponding vertical “guidelines” then lies along the bound, as above. Thus, for example, point E moves to point F.

This truncation alters the shape of the core, and creates multiple coincident points which may seem redundant, in this period. But the grid outside the core is still (trivially) formed by horizontal/vertical projection from these core edge points. And the algorithm treats them just like any other points, carrying them back into earlier periods where the combination of inflows and release opportunities will eventually move them back into the feasible storage area. This corresponds to a seasonal cycle in which generation options that may well be employed at some times of the year would not be contemplated at other times, even if the storage was already empty/full. For example, we might never use Otahuhu in Summer, even if storage was empty, but we would in Winter.

3 Application to the Upstream/Downstream Case

We have implemented CDDP using this compact representation and a generalised (two island) LDC-filling pre-computation, with “thermal” stations, or demand reduction, in both islands. The model has successfully been applied to a three reservoir case, and two-reservoir cases involving multiple efficiencies and inter-regional transmission losses. Here, though, we only discuss the upstream-downstream case, because it requires some alteration to the CDDP algorithm itself.

We ignore the possibility that upstream spill could by-pass downstream generation (as in the Tekapo/Pukaki system), and assume that all upstream release arrives in the downstream reservoir, in the same period. Generation efficiency is assumed to be constant, for each reservoir, but the same quantity of water produces a different quantity of energy at each reservoir. Thus upstream release quantities are scaled by an “efficiency ratio”, E , on receipt downstream. From a primal perspective, upstream release implies a “diagonal” shift on the storage diagram, reducing upstream storage, but increasing downstream storage. That same diagonal shift must be reflected in the way the DSR is interpreted in the CDDP algorithm.

If we interpret Figure 5 in terms of releases, it actually looks the same as Figure 3 above. In the electricity market, generation from these reservoir releases interacts in just the same way as any other generation, and the LDC filling pre-computation of Figure 2 proceeds identically. But, while release from the downstream reservoir is still driven by its MWV, release from the upstream reservoir is now driven by the *difference* between its MWV and that downstream. Thus the “guidelines” for the upstream reservoir now correspond to storage points at which the (efficiency adjusted) *incremental* MWV of upstream storage equals a critical level. If downstream MWVs are equal, that now means the incremental upstream value is zero, implying that we are indifferent as to whether upstream water spills, or not. The critical price relationship, at which we are indifferent between upstream and downstream generation is now when the downstream MWV equals the (efficiency adjusted) difference between upstream and downstream MWVs: At that storage balance point the MWV ratio is $(1+E)$, implying that energy stored in the upstream reservoir has just the same marginal value as energy stored in the downstream reservoir.

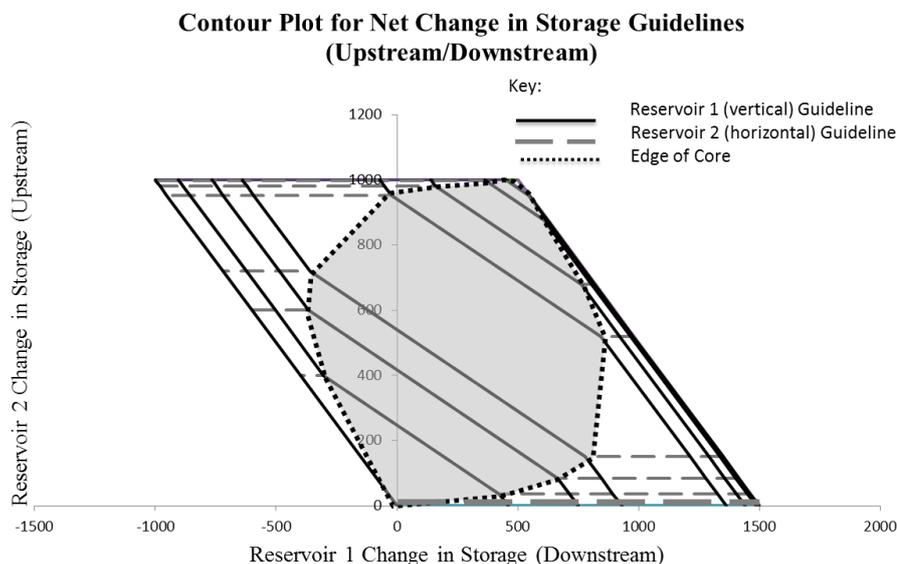


Figure 5: Net Change in Storage Contour Plot

But upstream releases also impact storage differently. So, in the change in storage diagram above, we get the sheared version of Figure 3 shown above, where reservoir 2 as the upstream reservoir and $E=1$. Note that the net downstream storage change can now be negative, if upstream release is greater than downstream release. The “horizontal guidelines” remain horizontal, but now correspond to a constant incremental MWV in the upstream reservoir. The “guidelines” corresponding to constant MWV in the downstream reservoir are no longer vertical, but have a slope reflecting the efficiency ratio between the two reservoirs. The “diagonal guideline” segments in the core remain, and still represent the locus of points at which we are indifferent between upstream and downstream release. However, they have a shallower slope as releasing more upstream water, not only raises downstream storage indirectly by backing off downstream release as before, but also now raises it directly. With that understanding, the general algorithm given above still applies, with two minor changes:

- First, the central recursive up-date equation can still be applied in a point-wise fashion, but Step 1 from Figure 1 for the downstream reservoir becomes:

$$\begin{aligned}
 & \text{Where } t = \text{period, } v_i = \text{Critical marginal water value res. } i, \\
 & i = 1, 2, E = \text{Efficiency ratio } u = \text{upstream res, } d = \text{downstream res,} \\
 & DSS_{t,d,v_1,v_2} = (MWVS_{t+1,d,v_1,v_2} - Inflow_{t,d} - DSR_{t,u,v_1,v_2} E) \quad \forall t, \forall v_1, \forall v_2
 \end{aligned}$$

- Second, the storage bounds to which “guidelines” must be truncated are not sheared. So “vertical guidelines” must now be truncated to the bounds on a angle, in a similar manner to “core guidelines”.

4 Conclusions

This report describes a re-conceptualisation of the way in which CDDP-based models, such as SPECTRA, deal with release and storage surfaces for hydro-electric generation. We focus on a representation in terms of critical MWVs to form a compact representation of surfaces that can be generalised to other system configurations, involving two or more reservoirs. A number of these system configuration models have been implemented, by modifying the intra-period pre-computations, and adding extra “cores” corresponding to extra critical MWV ratios. Of particular interest is the upstream-downstream model described above, which also required a slight modification to the CDDP algorithm, but added little to the computational burden.

We believe this simple, generalizable and compact form of CDDP will provide a functional and efficient basis for further research on the applicability of CDDP to stochastic reservoir release decisions for systems with higher modelled reservoir numbers. As the number of reservoirs increases the number of potentially critical ratios and hence cores, also increases, as does the number of “edge guidelines” for each core. But it should still be possible to model the number of significant long term storage reservoirs in New Zealand. Eventually, the curse of dimensionality will apply, though, and an exact representation will be unattainable. But it should be borne in mind that no other method attempts to provide an exact representation of the entire surface, and we plan to adopt a piece-wise linear approximation, as in SPECTRA, rather than the piece-wise constant approximation used here. This will also be necessary to deal with our other major challenge, which is improving on SPECTRA’s treatment of uncertainty.

Acknowledgments

The authors wish to thank the Electricity Authority for its support, and particularly Phil Bishop, Matthew Civil, John Culy, Roger Miller, and John Raffensperger.

5 References

- Dye, S., E.G. Read, R. A. Read, and S.R. Starkey. 2012. “Easy implementations of generalised stochastic CDDP models for market simulation studies”, *4th IEEE/Cigré International Workshop on Hydro Scheduling in Competitive Markets*, Bergen, Norway.
- Miller, R. 2009. “SDDP, SPECTRA and reality – a comparison of hydro-thermal generation system management”, *proceedings of EPOC*, New Zealand.
- Pereira, M.V.F. and L. Pinto. 1991. “Multi-stage stochastic optimization applied to energy planning.” *Mathematical Programming*, 52, 359-375.
- Read, E.G. and M. Hindsberger. 2010. “Constructive dual DP for reservoir optimization”, In *Handbook of Power Systems 1*. Edited by S. Rebennack, P.M. Pardalos, M.V.F. Pereira, and N.A. Iliadis, Vol. 1: Springer, 2010. 3-33.
- Scott, T.J. and E.G. Read. 1996. “Modelling hydro reservoir operation in a deregulated electricity market”, *International Transactions in Operational Research*, 3.3, 243-253.
- Starkey, S.R., Dye, S., E.G. Read, and R.A Read. 2012. “Stochastic vs deterministic water market design: some experimental results”, *4th IEEE/Cigré International Workshop on Hydro Scheduling in Competitive Markets*, Bergen, Norway.

Using Cognitive Mapping and Qualitative System Dynamics to Develop a Theory of Implementation in Primary Health Care

David Rees, Robert Y. Cavana & Jacqueline Cumming
Victoria Business School and Health Services Research Centre,
Victoria University of Wellington, New Zealand
david.rees@synergia.co.nz

Abstract

Over the last decade there has been a major emphasis on responding to the continuing rise in chronic conditions by delivering more health care in primary and community settings. To do so however, requires major changes in how care is managed and delivered, and research has managed to identify a number of factors that are key to bringing about these changes. Implementing the changes however, requires more than an understanding of key success factors. It also requires an understanding of how they interact over time in different contexts.

To develop this understanding, a number of health experts were interviewed, using cognitive mapping, to ascertain their thinking about the implementation challenge. The maps were then consolidated and developed further into a causal loop diagram, which describes a set of interlinked feedback loops representing the processes involved in implementing the required changes in primary care.

This causal loop diagram will then be developed into a quantitative system dynamics simulation model will be used to explore how the key factors, known to be important in implementing chronic care programmes in primary health care, interact and evolve over time.

Key words: cognitive mapping, qualitative system dynamics, causal loop diagrams, theory building, health.

1. Introduction

Much has been written about the process of model conceptualisation (for example see Cavana and Mares, 2004) but as acknowledged by Kopainsky and Luna-Reyes (2008) there is still a lot of room for research that illuminates the process of converting ‘real world’ information from the mental models of key informants into system dynamics simulation models. The prime purpose of this paper is to describe a process that was used to capture ‘expert theories’ of how to successfully implement programmes to improve care for people with long-term conditions and how that information contributed to the development of a qualitative system dynamics model. While there are a number of techniques available for model conceptualisation this research focuses on the use of cognitive mapping.

The idea of using cognitive mapping to help conceptualise system dynamics models was first articulated by Colin Eden in 1988 (Eden, 1988). Subsequent to that a number of papers have been written that describe the issues and themes involved (for example

see Andersen et al, 2007 and Eden et al, 2009). This paper builds upon that tradition and provides a specific example of how the idea can be applied in practice, using an explicit replicable process.

The importance of model conceptualisation was succinctly put by Eden (1994), in which he pointed out that defining the problem that the model is trying to solve is crucial and that effective model conceptualisation is, in the end, about, 'reducing the risk of finding the right solution to the wrong problem', (Eden, 1994, p 257). Taking this perspective, it was important in undertaking research about the implementation of health innovations to understand what the problem of implementation was about. This was done by focusing the initial research on developing an in-depth understanding of the views of seven people who are actively involved, at a senior level, in the design and implementation of initiatives to improve care for people with long-term conditions within New Zealand. The seven people interviewed were all involved at a national and regional level and four were also practicing clinicians, who combined their clinical practice with involvement in policy at both national and regional levels. The question that formed the basis of the interview was;

“What are the key issues that you consider to be important in the effective implementation of programmes for the care of people with long-term conditions?”

While there has been a lot of research that has identified a number of factors important to implementation (Carlfjord et al, 2010; Connolly et al, 2010) they generally have treated these factors as individual, independent items. So for example, clinical engagement is often mentioned as being important, but how it relates to and interacts with other important factors, such as information feedback, is less well understood. Lists of factors, while being useful to identify variables to consider, rarely provide any information about causality. This is, in part, because the world of implementation requires an understanding of the causal mechanisms at play and there is very little research that incorporates well specified causal mechanisms. Instead, while there is a plethora of research on the factors that are said to increase or decrease the successful implementation of an innovation, it is usually limited to qualitative or conceptual papers (Klein and Knight, 2005, Dewett, Whittier and Williams, 2007).

Such research also tends to ignore context. They assume that somehow there is a direct and isolatable causal link between, for example, 'clinical engagement' and more effective implementation of new health innovations. While it could be argued that clinical engagement is necessary, it is not sufficient and will only deliver more effective implementation of the new innovation if it is combined with other necessary conditions that enable the clinicians to become engaged and for their engagement to have impact.

To respond to this criticism of action lists this study develops a 'theory of implementation', that as well as describing key factors affecting the implementation of new innovative programmes for the care of people with long-term conditions also provides insight into the causal relationships between the factors and the contextual elements that affect those relationships. The initial themes were elicited in interviews with health experts using the cognitive mapping technique. These cognitive maps then provided information on the key variables that would be included in a causal loop diagram, which was used to describe the theory.

This paper describes the process used for eliciting the key issues involved in implementing programmes for improved care of people with long-term conditions from detailed interviews with health experts in New Zealand. The paper then describes how

those themes provided the basis for developing a theory of implementation, which is described using a causal loop diagram. Figure 1 shows the process used.

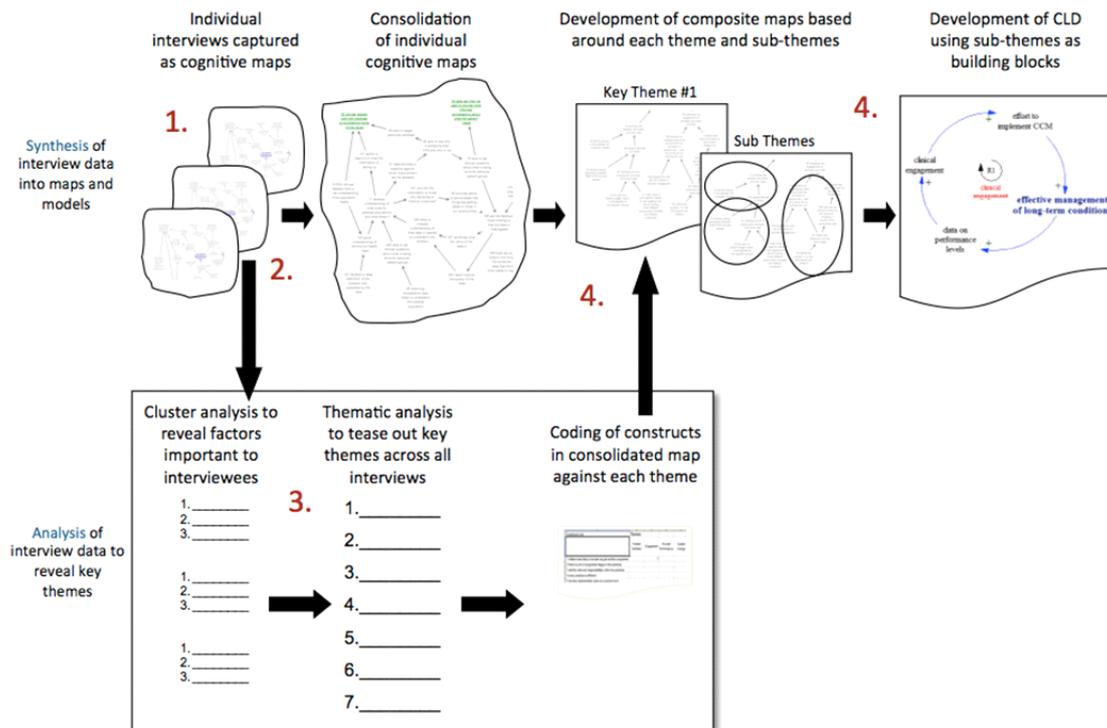


Figure 1. Using cognitive maps and causal loop diagrams to develop theories of implementation

2. Understanding mental models using cognitive mapping

The research began by interviewing seven experts within the New Zealand health sector. The aim was to obtain a good understanding of their mental models and do so in a way that allowed those models to be explored and tested. The interviews were structured using the cognitive mapping method, developed by Eden (1988). Cognitive mapping is a visual mapping technique used to elicit peoples’ description of a situation and/or issue; why it is the way they see it and why it is important to them. The interview process teases out the key ideas – termed constructs – related to the interview focus and through the use of unidirectional arrows depicts the line of argument. The arrows depict a line of influence. For example in Figure 2, construct 1 ‘develop a clear definition of the problem well supported by the data’ influences construct 5 ‘generates provider understanding of the gap between ‘what is’ and ‘what could be’’. Where there is a negative sign ‘-’ next to that arrow it indicates a negative influence. Thus meaning “...is not deduced from a semantic analysis but rather from the context of the construct – what it explains (consequences) and what explains it (causes)” (Eden, 1994, p 264). Cognitive maps also have an additional advantage in that by laying out the interviewees’ responses in the form of a visual map the interpretation of meaning is made explicit, able to be tested and therefore changed. To ensure that our interpretation of what was said in the initial interviews reflected what the interviewee was in fact trying to say, all people were interviewed twice. In the second interview the focus was on the cognitive map that was developed in the first interview, allowing it to be tested and refined. In all cases, the second interview led to further additions to the map,

elements they thought were not covered, or not covered in enough detail. It was rare to have any of the constructs in the first version deleted. In most cases the second interview provided the opportunity for a richer, more detailed discussion of key ideas. In all there were seven cognitive maps developed.

The cognitive maps were all inputted into 'Decision Explorer', a software tool developed by Colin Eden to display and analyse cognitive maps. Individual maps ranged in size from 25 to 53 constructs. Figure 2 provides an example of the cognitive maps developed in these initial interviews.

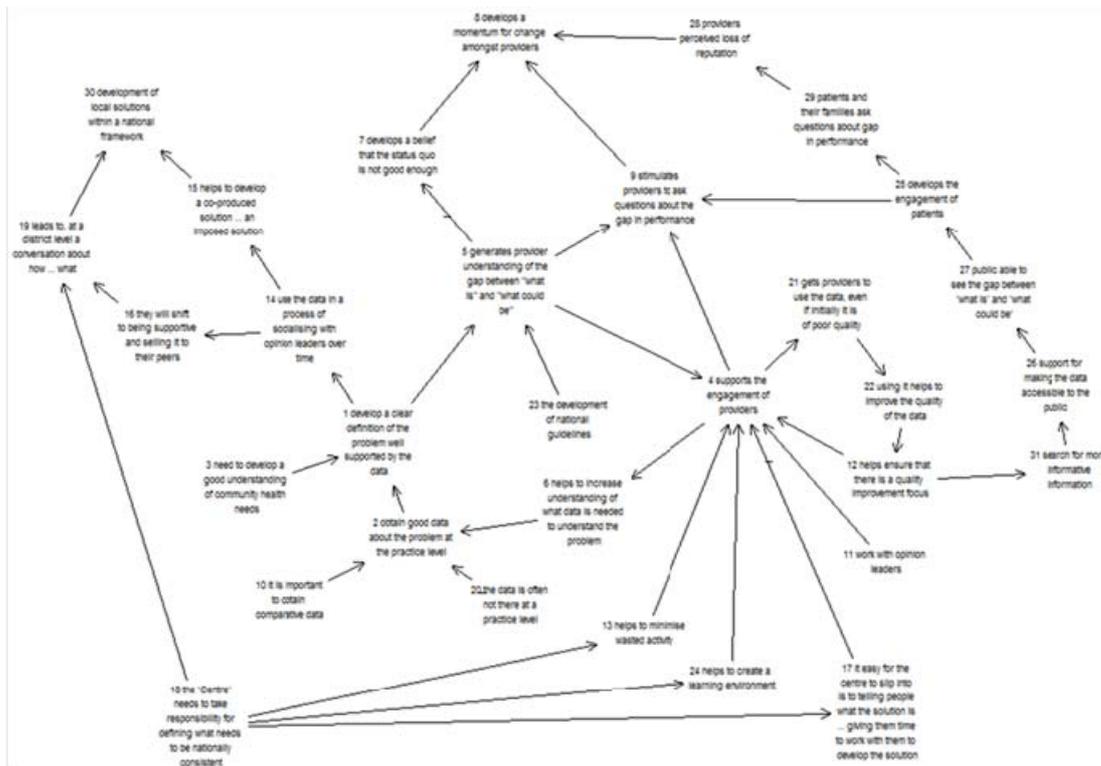


Figure 2. Individual cognitive map

2.1 Analysis of the Individual Cognitive Maps

The analysis of the maps was undertaken using a centrality analysis (Eden, 1994, p313). Centrality analysis highlights how central a construct is and, "...indicates the richness of meaning of each particular construct" (ibid, p 313), by calculating the number of in-arrows (causes) and out-arrows (consequences) from each construct. This is an important analysis as it pulls out, from the large number of connected constructs, those that are central to the ideas being presented by the interviewee. Using the software to do the analysis avoids preconceptions of the interviewer to determine what is, and is not important to the interviewee. What is important are those ideas that are densely connected, affecting and being affected by a large number of other ideas put forward during the interview process. Centrality analysis isolates core constructs and provides a method for developing a summary, or overview, of the total map that highlights the constructs having a significant importance to the interviewee. For example, in the cluster analysis conducted for the cognitive map shown in Figure 3, 'supports the engagement of providers' (construct 4) came through with a high score. Using the software to map other constructs directly linked to it revealed the following map:

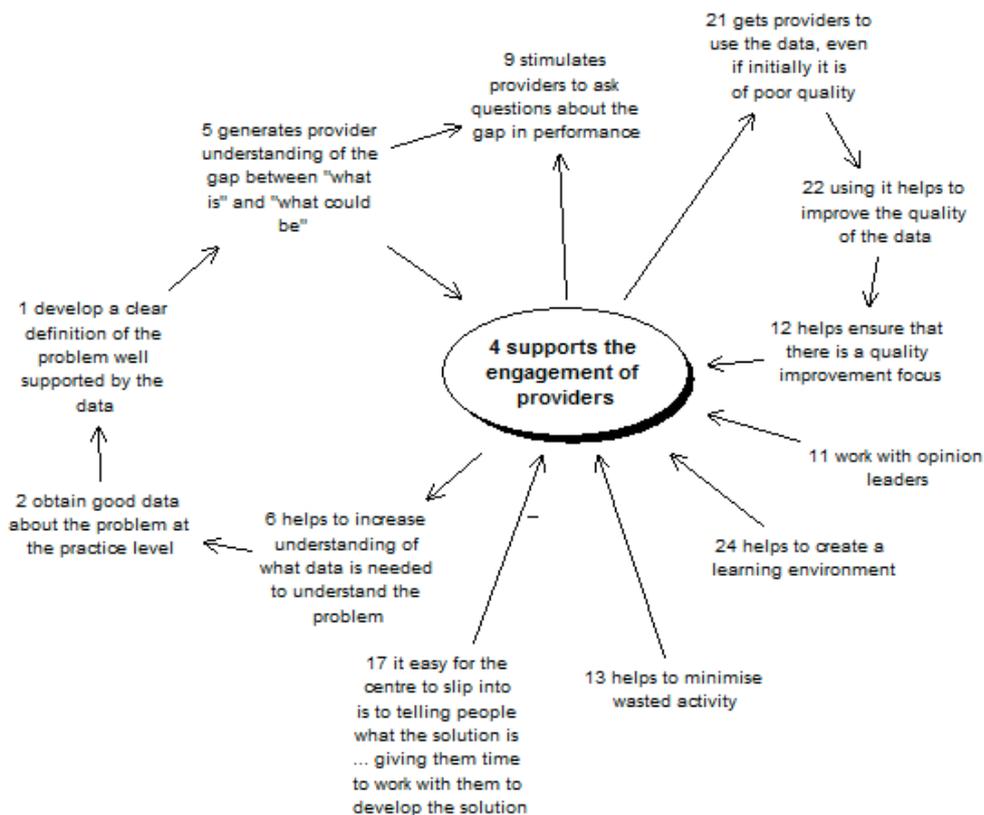


Figure 3. Cognitive map extract

Shown in the context of the map it becomes clear why this construct is considered important by the interviewee, and what is required if the meaning associated with it is to occur. As the map extract shows it is considered important by the person interviewed because it is a causal factor in increasing understanding of what data is needed to understand the problem (construct 6), supports the use of data, even when it is of poor quality (construct 21) and stimulates providers to question performance gaps (construct 9). To develop that engagement the interviewee considers it important to have a quality improvement focus (construct 12), minimise wasted activity (construct 13), develop a learning environment (construct 24), giving people time to work closely with you in developing the solution (construct 17), working with opinion leaders (construct 11) and developing provider understanding of what is and what could be (construct 5). In addition, there are also two important feedback loops at play. In the first, the engagement of providers promotes the use of data (construct 21), which enhances the quality of data available (construct 22), which in turn helps ensure a quality improvement focus (construct 12) that supports the further engagement of providers (construct 4). In the second, the engagement helps increase understanding of what data is required (construct 6), which supports the collection of good data (construct 2), which enables a clearer definition of the problem supported by good data (construct 1), which then develops greater understanding of the gap between 'what is' and 'what could be' (construct 5). That in turn reinforces the engagement of providers (construct 4).

Exploring a map in this way reveals what the interviewee considers important and what their line of argument is. It does provide a 'list' of key items but also uncovers the context within which they sit; how they link to other items and the meaning it has for

the interviewee. The use of cognitive maps begins to describe the causal theories of the interviewee, not just the factors considered important.

Each of the interviewees had a centrality analysis (Eden et al, 1992) conducted on their individual maps to ascertain those constructs that had a central position in their thinking. To ensure that the wider context of the construct was taken into account the centrality analysis was constructed to ensure that successive layers, or domains, were considered, that is, not just the constructs to which it is immediately linked, but also those that are further removed. Those that are further removed are given a diminishing weight i.e. those that are directly connected to the construct are given a weight of 1. Those that link into them, i.e. level two, are given a score of 0.5. Those that link into them, i.e. level three, are given a score of 0.25. The top 5 constructs for each person are shown in Table 1.

The centrality analysis enabled the authors to distil the key ideas from each of the seven interviewees. The 35 key constructs that emerged from this process were then coded, using the steps for conducting a content analysis outlined in Cavana et al, 2001, resulting in the emergence of seven key themes.

A check was done to see if any significant change in themes would occur if a greater number of constructs were included. To do this a further centrality analysis was done to include the top 7 constructs for each person, giving a total of 49 in all. When this analysis was done there were no new themes emerging. The only changes were slightly higher scores for each theme.

The themes and their scoring under the two options are shown in Table 1.

Table 1. Key themes arising out of the centrality analysis

Theme	Scoring of top 5	Scoring of top 7
Performance Feedback	6	8
Engagement	5	7
Provider Performance	5	7
System Change	5	6
Clinical Leadership	4	6
Organisational Leadership	4	6
Models of Care	3	6

Having now obtained the key themes from the initial interviews, the next step was to combine the data into an overall composite model that captured the constructs and their connections across all seven interviews.

2.2 Developing and analysing the composite cognitive map

A major benefit of utilising the Decision Explorer software is that it makes it possible to manage large amounts of qualitative data in a structured way. The next step was to combine all the individual maps into one overall composite map. This produced a map with 264 distinct constructs. These were then coded into one or more of the seven themes noted above. Using the software, maps were then created for each of the themes and each map was reviewed to merge constructs, where their meaning was the same. This resulted in the total number of distinct constructs reducing to 199. A cognitive map with 199 constructs is far too complex to analyse as an undifferentiated whole. However, maps of the seven themes reveal important ideas that reflect the thinking of

the health experts. They help unravel complex ideas such as ‘engagement’ and ‘support for provider performance’ and do so in a way that allows the development of a theory of implementation that is strongly grounded in their experience and expertise. For example coding the constructs within the combined model and merging duplicate constructs resulted in 30 distinct variables within the ‘engagement map’. In drawing this ‘engagement map’ a number of clusters, i.e. constructs linked together, emerged. The map is shown in Figure 4. The cluster on the left side of the map contains factors that refer to the contracting model. The next cluster along contains factors that relate to collaborative planning and programme design, while to the right of that is a cluster relating to community involvement. The boundaries between the clusters are drawn with a dotted line to acknowledge that fact that there is overlap, with some constructs able to be included in more than one cluster. While the boundaries are permeable they do highlight the four sub-themes that the experts interviewed consider important within the theme of engagement. Furthermore, the nature of the cognitive map highlights the causal links between those clusters and how together they affect engagement in a number of different areas. Thus, the use of cognitive mapping allowed us to unravel complex ideas such as engagement and obtain a much richer understanding of what was meant by the health experts and the more detailed understanding provided a level of specificity that could support the development of a casual loop diagram (CLD) to reflect the complex interacting ideas discussed.

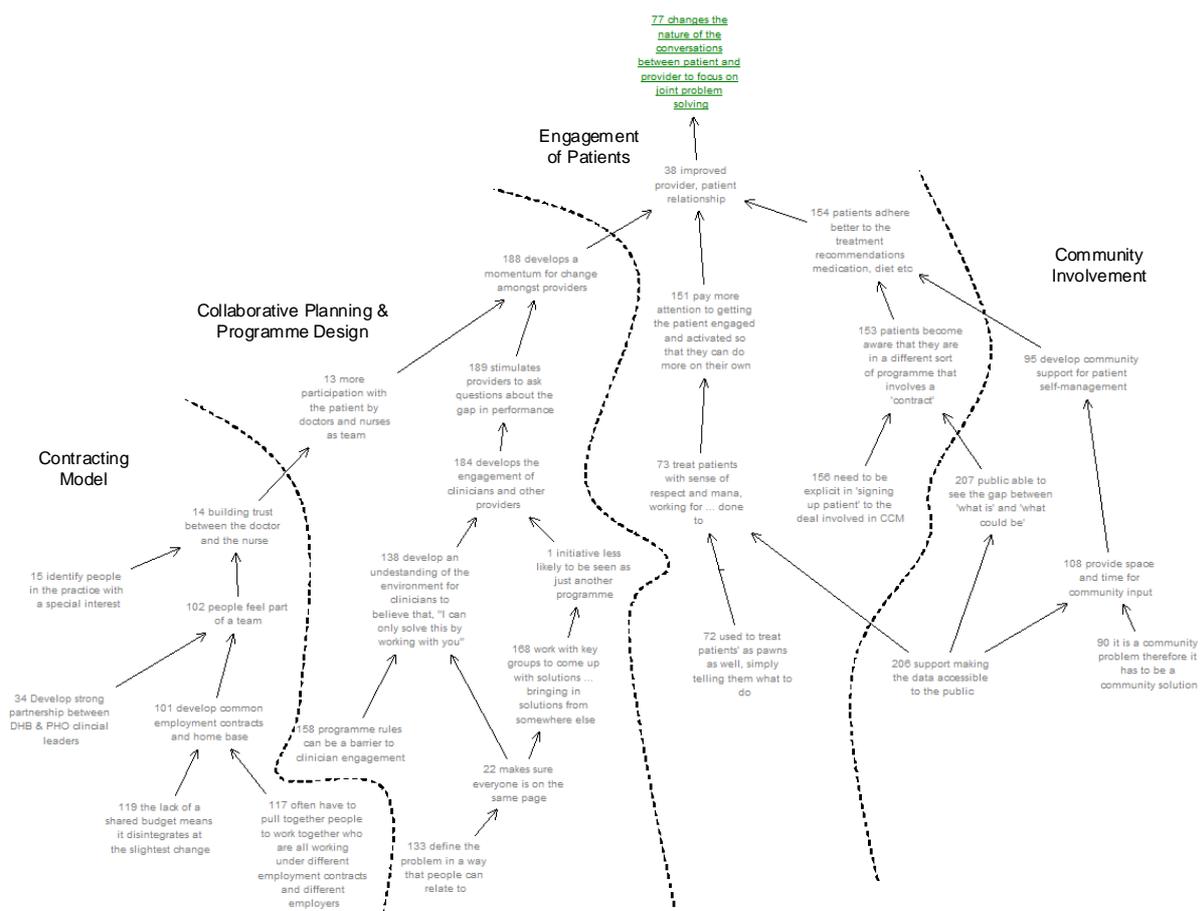


Figure 4: Composite map - Engagement

Of the seven themes that emerged out of the interviews five were used to develop the CLD. The model focused on the themes that related directly to the primary care practice rather than the broader policy and structural issues within the sector. As a consequence the themes of ‘System Change’ and ‘Organisational Leadership’ were not developed in the implementation model. This is not to say that they are unimportant but reflected a desire to focus on the themes that can be controlled, or at least influenced, by the primary care practice charged with implementing the new care models. The five themes, and the sub-themes within them, that were used to develop the initial CLD are shown in Table 2.

Table 2. Themes and sub-themes arising used to develop the initial CLD

Theme	Sub-Themes
Performance Feedback	<ul style="list-style-type: none"> – what works for practice population and what doesn’t – what processes deliver clinical outcomes – baseline data against which improvements can be assessed
Engagement	<ul style="list-style-type: none"> – clinical engagement – patient engagement – community engagement
Provider Performance	<ul style="list-style-type: none"> – support to do the right thing around the evidence – adequate resources – practice capability
Clinical Leadership	<ul style="list-style-type: none"> – clinical governance – clinical leaders working closely with practices
Models of Care	<ul style="list-style-type: none"> – self management – sustainability requires more self care – community support

A subsequent paper by the authors will describe the development of the CLD model, based on the interviews and analysis of cognitive maps. However, the full causal loop model is presented in the next section to illustrate the next stage of the process used to develop the theory of implementation in primary health care.

3. Developing the theory using CLDs

As pointed out by Schwaninger and Grösser (2008) a system dynamics model is a theoretical statement that is built upon a ‘reservoir of mental models’. In this case, the reservoir of mental models has been captured within the cognitive maps discussed above and the theoretical statement is being described using CLDs (For further details on causal loop diagrams, see, for example, Richardson & Pugh (1981), Sterman (2000) or Maani and Cavana (2007)). CLDs are a powerful tool for capturing patterns of causality and we have chosen to use this tool to bring together the ideas contained within the cognitive maps to develop a coherent theory of implementation in the context of long-term care management within the New Zealand Health sector.

The theory building starts with the ideas around clinical engagement as they were central to every single person interviewed. While some of the specific ideas around how it could be developed and how it could be undermined had different levels of emphasis the importance of clinical engagement was strong across all interviews.

The full model was developed a step-by-step process incorporating the key themes developed in the initial interviews and revealed through the development of the composite map and the thematic analysis. As the CLD model was developed the themes were also checked against the available literature to establish their validity and ‘flesh out’ the actual causal mechanisms where the interview data did not make this clear. The literature also confirmed that while each of the thematic elements has been mentioned in previous research; the combining of them into a coherent theory of implementation within the health context of long-term care management had not been done. The full theoretical model that emerged from this work is shown in Figure 5.

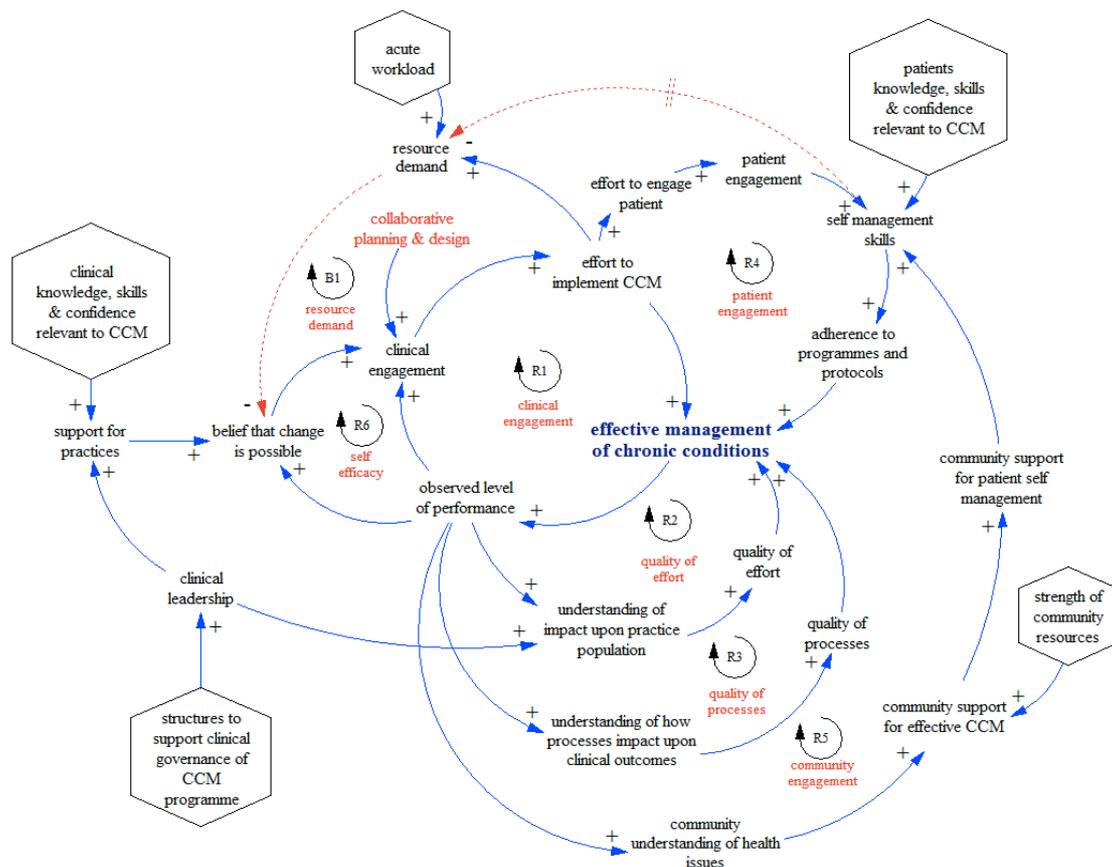


Figure 5. A theory of implementation

4. Concluding comments

The current theory, while it is yet to be validated or empirically tested, does provide a useful step in research on implementing new health innovations. Firstly, it describes a process by which expertise in the field can be captured and used to provide useful insights (ie through cognitive mapping). Secondly it provides a theory that takes into account the different contexts within which primary care practices exist (ie by the development of causal loop diagrams). Thirdly it provides an explicit, testable theory that can be used both to help in the initial planning of a new health programme as well as to evaluate programmes currently underway.

The next steps in the research are progressing in two areas. The first is to validate the theory with data from actual clinical practice. The second is to develop a fully quantified version of the theory so that the nature of the causal linkages can be explored in more detail. It is hoped that both streams of research will refine the current theory and develop a greater understanding of how it can be used to support improved practice.

5. References

- Andersen, D.F., J.M. Bryson, G.P. Richardson, F. Ackermann, C. Eden, and C.B. Finn. 2007. "Education and practice: the thinking persons' institute approach." *Journal of Public Affairs Education* **12**: 265-293.
- Carlfjord, S., M. Lindberg, P. Bendtsen, P. Nilsen, and A. Andersson. 2010. "Key factors influencing adoption of an innovation in primary health care: a qualitative study based on implementation theory." *Family Practice* **12**.
- Cavana, R.Y., and E.D. Mares. 2004. "Integrating critical thinking and systems thinking: from premises to causal loops." *System Dynamics Review* **20**: 13.
- Cavana R.Y., B.L. Delahaye, U. Sekaran. 2001. *Applied Business Research: Qualitative and Quantitative Methods*. Wiley, Brisbane.
- Connolly, M., M.-A. Boyd, T. Kenealy, A. Moffitt, N. Sheridan, and J. Kolbe. 2010. *Alleviating the Burden of Chronic Conditions in New Zealand: the ABCC NZ Study Workbook 2010*. University of Auckland, Auckland.
- Dewett, T., N.C. Whittier, and D.S. Williams. 2007. "Internal diffusion: the conceptualizing innovation implementation." *Competitive Review: An International Business Journal* **17**: 18.
- Eden, C. 1988. "Cognitive mapping." *European Journal of Operational Research* **36**: 1-13.
- Eden, C., F. Ackermann, and S. Cropper. 1992. "The analysis of cause maps." *Journal of Management Studies* **29**: 309-323.
- Eden, C. 1994. "Cognitive mapping and problem structuring for system dynamics model building." *System Dynamics Review*, **10**(2-3), 257-276.
- Eden, C., F. Ackermann, J.M. Bryson, G.P. Richardson, D.F. Andersen, and C.B. Finn. 2009. "Integrating modes of policy analysis and strategic management practice: requisite elements and dilemmas." *Journal of the Operational Research Society* **60**, 12.
- Klein, K.J., A.P. Knight. 2005. "Innovation implementation: overcoming the challenge." *Current Directions in Psychological Science* **14**, 4.
- Kopainsky, B., L.F. Luna-Reyes. 2008. "Closing the loop: promoting synergies with other theory building approaches to improve system dynamics practice." *Systems Research and Behavioral Science*. **25**: 471-486.
- Maani, K.E., R.Y. Cavana. 2007. *Systems Thinking, System Dynamics: Managing Change and Complexity* 2nd ed., Pearson Education NZ Limited, Auckland, New Zealand.
- Richardson, G.P., A.L. Pugh III. 1981. *Introduction to System Dynamics Modelling with DYNAMO*, Productivity Press, Mass.
- Schwaninger, M., S. Grosser. 2008. "System dynamics as model-based theory making." *Systems Research and Behavioral Science* **25**: 447-465.
- Sterman, J.D. 2000. *Business Dynamics. Systems Thinking and Modeling for a Complex World*, McGraw-Hill Higher Education, Boston, USA.

Cloud Computing for Operations Research

Christian Rolf
Corvid
christian@corvid.co.nz

Abstract

In this paper we present how to make high-performance solving of OR problems cost-aware. Balancing performance and cost is necessary to maximize the benefits of cloud computing.

Solving difficult OR problems requires high-performance computing, which quickly gets costly, even in the cloud. In cloud computing, where processing resources are bought and sold as commodities, the price of solving an OR model is determined by factors like bandwidth consumption and total processing time. Blindly maximizing the performance is all but guaranteed to maximize the cost.

Our work is focused on transparently and dynamically adapting the solving process to suit the current cost models of cloud computing. We accomplish this by observing the computational structure, and needs, of the solving process. Beyond the adaptation, we also find the most suitable cloud platform by matching the computational needs with the current cloud spot price.

Our preliminary results show that the cost can be reduced by changing the OR solving to suit the spot price of the current cloud computing platform. We present how to reduce this cost by migrating parts of the solving process to the platform that best suits its computational needs.

Key words: Cloud computing, Operations Research, Cost Optimization

1 Introduction

Cloud computing can be used to solve larger and more complex operations research (OR) problems than was previously thought possible. In this paper we will outline the basics of how to do this.

Cloud computing comes in three forms; Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS, like the Amazon Elastic Cloud, is simply a virtual machine that can be used to run the same programs as on a regular server. PaaS, like the Google App Engine, require the use of the platform's application programming interface (API), everything below that is hidden, i.e., you cannot run software that has not been explicitly written for the platform. SaaS, like the Salesforce Sales Cloud, provides direct access to software but little support for creating extensions. Some integer programming software is available in the cloud, for instance, Gurobi (Gurobi Optimization 2012), which is available on Amazon's cloud.

The price per processor cycle is much lower in the cloud than in, e.g., a university cluster. Moreover, the price in the cloud fluctuates depending on demand. This turns computational resources into a market where users can bid for slots. Amazon, for instance, allows bidding for different resources and clusters by simply posting a job that will be run if the price drops down to the bidding point (Amazon 2012). The current price is known as the *spot price*.

The price to run a computational task in the cloud depends on its resource demands. For instance, some tasks need a lot of data from the hard drive, while others use very small data sets. Given the pricing strategies of cloud services, different tasks will be cheaper to run on different clouds. Amazon charges per hour, but the run-time will vary depending on how powerful the computational node is. The Google App Engine is free up to a point, then charges start to apply per hour, per storage interaction, and per gigabyte of bandwidth used.

In this paper we describe how to reduce the price of solving OR problems by splitting up the computations between different nodes. By looking at the spot price, and observing the computational need, we can select the cloud platform that will be cheapest for a particular task.

2 Cloud Computing for Operations Research

When splitting up computations of OR problems, some parts of the solution space are more interesting than others. Prioritizing these parts can render a major performance improvement that increases with the level of parallelism (Rolf and Kuchcinski 2008).

In general, the computations behind solving OR models can be automatically and efficiently parallelized (Rolf 2011). Cloud computing really comes into its own when spreading computations to many nodes, allowing larger problems to be solved than on single machines.

In order to make the most of parallel computation, the most interesting parts of the solution space should be explored first. This ordering is most easily defined by the programmer, hinting to the solver the order in which the search space should be explored.

Once a priority has been established, the less interesting parts of the solution space can be split off. These jobs can then be submitted with low bids. If the spot price never drops down to that level, the jobs will remain queued until cancelled. If all interesting parts of the search space have been explored, the lower priority parts can be resubmitted at higher bids as they are likely to terminate fast. How the solution space is split is outside the scope of this paper, we refer the curious readers to (Rolf 2011).

Ideally, the splitting and prioritization of jobs should be transparent and automatic. We are currently working on this; our preliminary implementation has three levels of priority for jobs and is very likely to increase the performance while lowering the computational price.

Beyond simply improving the speed of solving, the cost should not skyrocket when using big computational nodes. Hence, the different price models of clouds has to be taken into consideration. The simplest way to do this is to place low bids for low priority jobs and target them to the smaller computational nodes. Then place higher bids for the high priority bids and send them to high-performance nodes.

As an example of the prioritization that can be performed: the most interesting 25% of the solution space is sent to high-performance nodes, at a bid of \$0.25 per hour. The least interesting 25% is sent to low-performance nodes with a price of \$0.010 per hour. Finally, the middle 50% are sent to average-performance nodes with a price of \$0.020 per hour. All prices are approximate as they fluctuate over time. The most expensive time of the year on Amazon is around Christmas.

Using only low cost nodes will not necessarily minimize the cost of solving a problem. High performance nodes will typically generate solutions faster. This, in turn, will bound the search space, reducing the total computational price.

3 Conclusions

At Corvid we leverage cloud computing to try and solve some of the hardest operations research problems in the world.

In this paper we described how cloud computing can be used to push the limits of OR solvers; allowing harder problems to be solved at an acceptable cost.

References

- Amazon. 2012, November. EC2 Spot Instances. <http://aws.amazon.com/ec2/spot-instances/>.
- Gurobi Optimization. 2012, November. Gurobi Cloud. <http://www.gurobi.com/products/gurobi-cloud>.
- Rolf, Christian. 2011, Oct. "Parallelism in Constraint Programming." Ph.D. diss., Department of Computer Science, Lund University, Sweden.
- Rolf, Christian, and Krzysztof Kuchcinski. 2008. "Load-balancing methods for parallel and distributed constraint solving." *Cluster Computing, 2008 IEEE International Conference on*, 29 2008-Oct. 1, 304–309.

Contribution Margin Optimisation at Fonterra

Kevin Ross
Chief Scientist – Optimisation Modelling
Fonterra
New Zealand
kevin.ross@fonterra.com

Michael Freimer
Chief Scientist
SignalDemand
USA
michael.freimer@signaldemand.com

Abstract

Fonterra have partnered with SignalDemand to optimise New Zealand dairy product allocation and pricing through the development of an integrated optimisation tool. The global dairy market is expected to expand rapidly in the coming decade, and this approach will allow Fonterra to focus on the highest returning product streams. The model takes the latest view of global demand, milk forecasts, yields, and production capacity to develop a 24 month optimised plan. The model is implemented as software as a service, with the optimisation model in GAMS.

Key words: Mathematical Programming, Optimisation, Supply Chain, Pricing

1 Background

Fonterra is the world's largest exporter of dairy products, selling 2.48 million tonnes of dairy products and earning almost \$20 billion NZD in revenue per year. The majority those products begin as raw milk on New Zealand farms. Fonterra manages the milk from the farms to customers around the world.

Previously, decisions of product mix and price were handled in a disjoint way, so that production, transportation, and inventory planning were managed separately from the decision of which customer demand to satisfy. In this work, Fonterra partnered with SignalDemand to develop an optimisation tool that would support the planning of demand and supply in an integrated way. The optimisation model is implemented as software as a service using GAMS, with the data automatically fed between Fonterra and SignalDemand systems.

The model developed by SignalDemand was constructed on a platform used to support similar optimisation models for other commodity processors. The Fonterra value chain has a few unique features that were incorporated into this model. Unlike most traditional supply chain situations, Fonterra is a co-operative owned by the farmers who supply the milk. The Fonterra's mandate is to maximize returns from the milk that is provided, so there is no direct payment for the raw ingredients. This eliminates the

opportunity and flexibility to purchase, delay or reject supply, and means that the plan must follow the seasonal milk production. Both milk volumes and the composition of specific milk solids within the milk are subject to seasonal variations, which must be accounted for in demand fulfilment and production planning. As a disassembly industry, production of one product often implies the production of a number of co-products, adding complexity to the product mix dynamics.

Channel mix also provides an opportunity for optimisation. Demand at Fonterra includes Fonterra Brands and a variety of products worldwide. One interesting channel for sales is the globalDairyTrade (gDT), an auction platform to sell base commodities. Fonterra must decide how much product to offer at each event, given a projection of price achievement.

Another unique feature of this problem is that production, inventory and transportation are managed around the date of departure from New Zealand ports, while the price of products is tied to the time at which a contract is agreed. Therefore a forward sales profile is used to capture the distribution of time separating a product's sale and shipment. The forward sales profiles are combined with estimates of price elasticity to model the relationships between price and demand.

For commercial reasons, the details of the model are not included here, but the overall scope is described.

2 Optimisation Model

2.1 Overview

The optimisation model balances supply and demand across the enterprise in such a way as to maximize overall returns. It makes coordinated recommendations regarding how to allocate milk from farms to production facilities, how much additional raw material (lactose) to purchase, what to make from the milk, which products to hold in inventory, which customers to sell to, what to sell them, and what price to charge.

The objective of this model is to maximize total contribution margin, while satisfying all existing commitments. The important components of margin are revenue (sales volume times price, which itself may be a function of volume) and various forms of cost which include the costs of processing, holding, transportation, and lactose procurement.

Two primary decision variables in the model are price and volume, i.e. how much to allocate to certain segments of uncommitted demand, and the prices Fonterra may expect to realize at those volume allocations. Both of these decision variables are indexed by product segment, customer segment, and time period. The two are closely tied to one another via the price elasticity; as volume increases we expect price to decrease, and vice-versa. Other decision variables correspond to the supply-side decisions that must be made: how much to produce of each product, how much to hold in inventory, and so on.

An important distinction should be made between the *supply clearing* and *product mix* aspects of the optimisation problem. Fonterra is constrained by the perishability of its raw material, so one goal of the optimisation model is to decide how to clear the available milk supply most profitably. More specifically, the model decides how best to process the raw milk and sell the finished goods within an allowable time interval. Limited production capacity poses a similar problem; capacity is perishable in the sense that unused capacity cannot be stored.

Determining the right product mix is another aspect of the optimisation problem. Fonterra can produce a broad variety of products that are sold to multiple customer segments, and the optimisation model must determine the optimal allocation among products and customers. The mix problem is particularly difficult in disassembly industries such as dairy, where the production of one product usually implies production of a number of co-products.

In either case price can be viewed as a mechanism for balancing supply with demand. Fonterra may chase additional demand when supply or capacity are long but in doing so may realize lower prices; similarly prices may rise when volume allocations decline. This dynamic is particularly evident with respect to gDT, an online platform supporting ascending price clock auctions. gDT is a relatively recent channel (July 2008) through which Fonterra offers commodity products.

Finally, two additional important aspects of the optimisation problem are determining the optimal forward sales profile and inventory profiles. The forward sales profile captures the pace at which Fonterra sells into a given time period. It is usual to have a combination of forward bookings and short term sales, and the optimisation model helps to manage the mix. The model also makes recommendations regarding the inventory profile. Although the raw material is perishable, Fonterra's dried and chilled finished materials are shelf-stable, so inventory can be used to resolve supply and demand imbalances. The optimisation model must decide the optimal mix of products in inventory and the timing of stock build and draw-down.

2.2 Model Structure

The optimisation model is based on monthly time periods. The first time period may be a partial month, and some optimisation input values are scaled appropriately. The optimisation horizon is 24 months, which is long enough to cover the remainder of the current season's production cycle plus one additional cycle.

The optimisation objective function is overall contribution margin less penalties for violating soft constraints. Contribution margin is summed over each of the time periods within the optimisation horizon. The contribution margin is made up of the revenue for optimised demand (shipped volume multiplied by price at the time the volume was ordered) minus variable costs (manufacturing, inventory, transportation and purchased ingredients). Soft constraints include minimum bounds on production and inventory.

The key decisions within the optimisation model include price, allocation, production, purchasing, inventory and transportation volumes. Most of these are indexed by time and location. The tactical nature of decisions being informed by this model does not require any integer variables to be used.

With regard to constraints, the table below describes the optimisation model's major constraint categories.

Constraint Category	Description
Material Balance	Balances supply, production, consumption, and inventory within each month. Incorporated into these constraints are any restrictions on the facility and age of material that can be used to satisfy each customer segment.
Price-Volume Relationships	Relates prices to volumes based on elasticities and forward sales profiles.
Inventory Bounds	Upper and lower bounds on inventory levels at the beginning of the time period. Includes end-of-horizon inventory targets.
Supply Purchase Bounds	Procurement must remain within the bounds defined by the supply option
Transportation Bounds	Transportation volumes must remain within minimum and maximum bounds
Allocation Bounds	Allocations must remain within minimum and maximum bounds. In particular, allocations must be less than unconstrained demand.
Resource Bounds	Resource consumption constrained to be no greater than availability
Utilization Bounds	Resource utilization must remain within minimum and maximum bounds
Production Bounds	Production must remain within minimum and maximum bounds

The backbone of the optimisation model is a collection of material balance constraints. These constraints describe how items are produced, consumed, and stored from one period to the next. There is a distinct material balance constraint for every unique combination of item, facility, and month. Items include whole milk, other liquids (e.g. cream, permeate, buttermilk, whey), and finished materials.

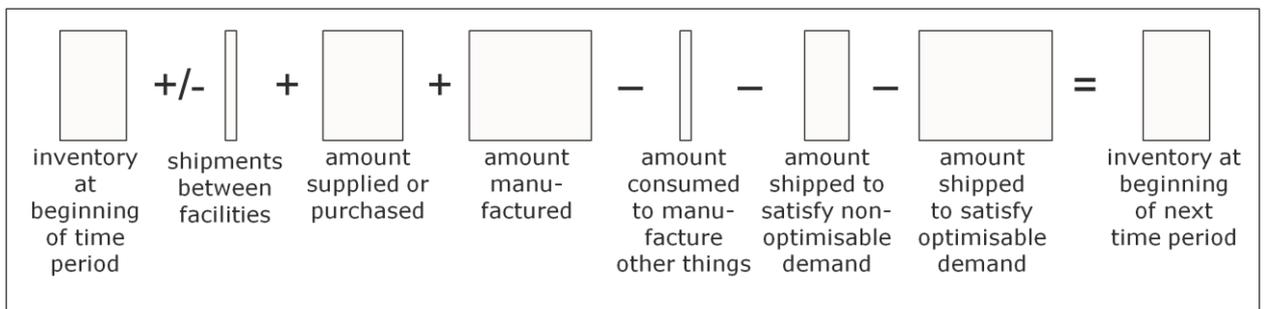


Figure 1: Generalized Material Balance Constraint

Figure 1 displays a generalized material balance constraint. The left hand side of the equation describes dynamics during the month for which the constraint is written. Starting with initial inventory of the item, we add the amount supplied or purchased and the amount manufactured. We subtract the amount consumed to make other items and the amount shipped to satisfy demand. Shipments to or from other facilities are added

or subtracted as appropriate. The result is inventory at the beginning of the following month.

2.3 Model Outputs

The primary outputs of the model are described below, and are presented in a variety of reports for Fonterra.

Production Plan	The optimisation model produces an optimal production plan consisting of recommended production volumes by material, production resource, and month.
Demand Allocation	The optimisation model produces an optimal demand allocation (a constrained demand plan) consisting of allocation recommendations.
Prices	The prices Fonterra may expect to realize given the optimized demand allocation.
globalDairyTrade	The optimisation model recommends volumes for future gDT events, split by contract.
Shadow Prices	The shadow prices inform the marginal value of additional capacity or unconstrained demand.
Other Outputs	A variety of other information is produced by the optimisation model. Examples include optimal inventory levels and optimal shipment volumes.

2.4 Modelling Issues

The supply side of the optimization model traces the flow of material from whole milk through intermediates (e.g. cream, buttermilk, whey) to finished materials (e.g. buttermilk powder, mozzarella). All of these material stages are linked by bills of material. The demand side of the model represents both unconstrained and committed demand volumes in terms of product and customer segments. (Committed demand includes existing contracts as well as most domestic sales.) The model decides the quantities of finished materials to withdraw from stock in order to satisfy demand. Supply and demand are therefore aligned at the level of product segment / customer segment / month. On the supply side, production, inventory and product mix recommendations shape supply to meet committed and prospective demand, within the bounds of production and supply constraints. On the demand side, allocation and price recommendations shape demand to match the available supply, subject to the limitations of price elasticity and demand forecasts.

Within the optimisation model, milk supply originates at roughly 300 milk cells (farm clusters) according to the milk volume forecast. In some cases milk is collected at milk transfer stations (e.g. Tuamarina, Culverden, Oamaru, Oringi, and Longburn) before being sent on to production sites, although this is not represented in the optimisation model. The model makes recommendations regarding how milk should be delivered from cells to production sites.

The composition of solids within milk is also seasonal, and Fonterra maintains projections of composition as well. Within the optimisation model this variability is captured via time-indexed bills of materials.

Production is modelled at the production line level, in monthly buckets. Since the model is designed to support aggregate planning decisions, changeovers and other details of batch processing are not represented explicitly.

A snapshot of on-hand finished materials inventories is an input to the optimisation model. From this starting point, the model makes recommendations regarding how stock levels should evolve over time while keeping track of inventory age (in months). Inventory is modelled at the regional level (two for the north island, one for the south), and separate constraints are included for dry and chilled storage capacities.

Transportation lanes are defined by item and pairs of facilities (including milk cells, production facilities, regional inventory sites and ports). Per-unit shipping costs may be associated with each transportation lane, as well as upper/lower bounds on the volume shipped in a given month. Cost and bounds may vary over time. The optimisation model makes explicit recommendations regarding how much of each item should be shipped across each lane in a given month.

Price elasticity measures the strength of the relationship between price and demand for a particular combination of product and customer segments. This relationship has been modelled explicitly for certain categories of demand, adding to the complexity of the product and channel mix problem.

3 Implementation

The optimisation model described in this paper has been fully implemented in live systems at Fonterra and SignalDemand. The model is provided through software as a service, with the model running on servers in California. This allows SignalDemand to monitor and maintain the model, while Fonterra manages their data.

As with many models of this type, two of the major challenges have been data quality and run times. Given the level of detailed data required, a major project was needed to deliver the data and maintenance processes that ensure sufficient accuracy is included.

The run time of the model itself depends on both the data processing (for implementation as a software-as-a-service model), and the model itself. Tuning of various parameters to make this more efficient has been ongoing.

4 Conclusions

The implementation of price and product mix optimisation at Fonterra has been a successful demonstration of large scale optimisation, as well as software as a service. Since the implementation is recent, measurable benefits have not yet been captured, but they are expected to be significant.

Several advances to this model are under consideration, including the incorporation of global sourcing decisions and integer constraints.

An Aggregational Approach to the Traffic Assignment Problem

Keith Ruddell and Andrea Raith
Department of Engineering Science
University of Auckland
New Zealand
krud006@aucklanduni.ac.nz

Abstract

Given a traffic network subject to congestion and given demand for travel between origins and destinations within the network, a user equilibrium is a pattern of flow that satisfies this demand and where no user can reduce her travel time by taking a different path. This can be framed as a minimisation problem with non-linear objective — the Traffic Assignment Problem (TAP). Several algorithms exist for solving TAP that start with an initial feasible solution. It is current practice to use for an initial feasible solution the so-called all-or-nothing flow allocation, where all flow simply follows the shortest path ignoring congestion. We develop the approach of solving TAP on a simplified aggregated network and then translating this optimal solution into a feasible solution on the full network. The obvious, and most effective, simplification is to group neighbouring origin and destination nodes together ignoring local traffic between them. Projecting the resulting solution onto the full network we start closer to the optimum and so require fewer expensive iterations of the full-network problem to arrive at an acceptable solution. Preliminary results of empirical tests show that, depending on the aggregation chosen, our method can be twice as fast as existing algorithms.

Key words: Traffic Assignment Problem, Graph Theory, Network equilibrium, Graph abstraction, Algorithms.

1 Introduction

The Traffic Assignment Problem (TAP) forms an important step in many transport models, most notably the four-stage model — described in Ortúzar and Willumsen (2001) — which has been standard since the 1970s. In this model traffic assignment is the final stage, after trip generation, trip distribution and mode split. It is assumed that demand for travel (in units of ‘trips’) between points in the traffic network is known, as well as the slow-down response of transport links (roads, interchanges) to increasing flows due to congestion. The congestion effects are modelled by cost-flow functions, one for every link (arc) of the network graph. They are increasing functions, so the link effectively gets longer the more travellers use it. A further

important assumption is that congestion on any link is dependent only on the flow on that link, and in particular not on its neighbours. This is no small assumption as in reality what happens at intersections is an important determinant of congestion.

The solution of TAP is a traffic equilibrium. This is meant to represent the aggregate behaviour of rational travellers seeking the ‘shortest’ path (in the sense of travel time or some combination of travel time and other costs) from origin to destination. Wardrop (1952) defines the user equilibrium as an assignment of flows such that “the journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.” We give a minimisation formulation of TAP in section 2.

In transport modelling frameworks such as the four-stage model traffic assignment is one of the more computationally demanding stages. Hence it is of practical interest to be able to quickly compute equilibrium solutions.

Many different algorithms have been developed for TAP. We describe two in section 3 which will also form sub-algorithms of our new approach presented in section 4. Our new approach, which we call centroid aggregation (CA), is really a method for initialising existing algorithms. Where most algorithms begin with ‘take a feasible set of flows’, it is traditional to use the naïve all-or-nothing (AON) solution where all the flow between a given origin and destination is assigned to the shortest path where path length is calculated at zero flow. In section 5 we give results of testing our approach on two test networks and in section 6 describe some further refinements that we have yet to implement.

2 Formulation of the Traffic Assignment Problem

Our formulation is adapted from Patriksson (1994). Let \mathcal{N} denote the set of nodes in our network, and $\mathcal{A} \subset \mathcal{N}^2$ its directed arcs. Traffic flows to and from a special set of non-traversable nodes called zone centroids, \mathcal{Z} . We write $\mathcal{C} := \mathcal{Z}^2$, for the origin-destination pairs. The demand function $d : \mathcal{C} \rightarrow \mathbb{R}$ is nonnegative and *only* depends on the origin-destination pair. The extension where demand varies with route cost reduces to this fixed demand case by the addition of virtual arcs.

The set of routes from $p \in \mathcal{Z}$ to $q \in \mathcal{Z}$ is \mathcal{R}_{pq} . We define the set of all routes $\mathcal{R} := \bigcup_{(p,q) \in \mathcal{C}} \mathcal{R}_{pq}$. The flow along any route r is h_r .

The arc-cost functions $t_a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are positive, increasing and continuous for all $a \in \mathcal{A}$. These functions may represent travel time, cost or a combination of these. In our example networks the arc-cost functions are all power functions of the form

$$t_a(f_a) := (\text{Free-flow travel time}) \cdot \left(1 + B \cdot \left(\frac{f_a}{\text{capacity}} \right)^P \right), \quad (1)$$

where B and P are constants that are calibrated to the particular network at hand. Usually, B is about 0.15 and P is a small power, between 2 and 6. Note that here ‘capacity’ is not a hard , it merely represents the point at which congestion effects begin to increase rapidly.

The variables in our formulation are the route-flows h_r . We define the vector of link flows \mathbf{f} by

$$f_a := \sum_{r \in \mathcal{R}: a \in r} h_r \quad \forall a \in \mathcal{A}. \quad (2)$$

That is, the flow along an arc is the sum of all route-flows over all routes that traverse that arc. This allows us to formulate TAP as the minimisation problem

$$\text{minimise } T(\mathbf{f}) := \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s) ds \quad (3)$$

subject to (2),

$$\sum_{r \in \mathcal{R}_{pq}} h_r = d(p, q) \quad \forall (p, q) \in \mathcal{C} \text{ and} \quad (4)$$

$$h_r \geq 0 \quad \forall r \in \mathcal{R}. \quad (5)$$

In words, the demand for travel between each origin-destination (OD) pair must be equal to the sum of route-flows over all routes connecting that OD pair, and the route-flows are all nonnegative.

The objective function requires some explanation. If we define $\pi_{pq}(\mathbf{h})$ as the length of the shortest path from p to q and $c_r(\mathbf{h})$ as the cost of route r , where both of these depend on the network flows, then we can write the Wardrop user equilibrium condition as

$$h_r(c_r(\mathbf{h}) - \pi_{pq}(\mathbf{h})) = 0 \quad \forall r \in \mathcal{R}_{pq}, \forall (p, q) \in \mathcal{C} \quad (6)$$

with h satisfying (4) and (5). But (as the reader can verify) this is just the first-order optimality condition on the Lagrangian relaxation

$$\mathcal{L}(\mathbf{f}(\mathbf{h}), \boldsymbol{\pi}) = T(\mathbf{f}(\mathbf{h})) - \sum_{(p,q) \in \mathcal{C}} \left(\pi_{pq} \left(d_{pq} - \sum_{r \in \mathcal{R}_{pq}} h_r \right) \right) \quad (7)$$

of with respect to constraints (4).

Though it is easiest, in the abstract, to present a feasible TAP solution as a set of route-flows it is common and accepted practice in traffic modelling to generate just the flow values for the links. One reason for this is that for a given set of link flows there are, in general, many possible route flow solutions. As there is no good reason to prefer one equilibrated route flow solution to another, it is sufficient to know the link flows (from which it is elementary to compute the route costs).

3 Existing Algorithms

The Frank-Wolfe algorithm 1 is one of the most commonly implemented for solving TAP. Its main feature is that it stores only the link-flow variables, not route-flow. Because of this it is quite fast to run. However, it only barely converges; approaching equilibrium, the search direction becomes perpendicular to the direction of the equilibrium point. Figure 1 shows rate of convergence getting ever-slower. See Patriksson (1994) for a fuller description.

Algorithm 1 Frank-Wolfe Algorithm

Take a feasible solution to TAP, with link-flows \mathbf{f} .
while Convergence Criterion not met **do**
 for all OD pairs **do**
 Find shortest path r , given current costs, giving AON solution \mathbf{f}^* .
 end for
 Using your favourite one-dimensional convex optimiser, find $\hat{\mathbf{f}} = (1-\lambda)\mathbf{f} + \lambda\mathbf{f}^*$
 with $\lambda \in [0, 1]$ to minimise $T[\hat{\mathbf{f}}]$.
 Update current link-flow solution $\mathbf{f} \leftarrow \hat{\mathbf{f}}$.
end while

The path equilibration algorithm 2 is representative of the path-based class of TAP algorithm. Such algorithms must use column generation to cut through the combinatorial explosion of possible routes, following (Dafermos and Sparrow 1969). It only ever stores the routes with non-zero flow (active routes) and the shortest routes at current flow-levels in each iteration.

Algorithm 2 Path Equilibration Algorithm

Take a set of active routes $\hat{\mathcal{R}} \subset \mathcal{R}$ and a feasible allocation of flows to those routes \mathbf{h} , zero outside of $\hat{\mathcal{R}}$.
while Convergence criterion not met **do**
 for all OD pairs (p, q) **do**
 Find shortest route r (in \mathcal{R}_{pq}), at current flow levels.
 Move flow from the longest route in $\hat{\mathcal{R}}_{pq}$ to r until their costs are equal.
 If the longest route reaches zero flow, remove it from the active set $\hat{\mathcal{R}}$.
 end for
end while

3.1 Relative Gap

For convergence criterion we use relative gap, defined by

$$\begin{aligned} \text{Rgap}(\mathbf{h}) &:= 1 - \frac{\boldsymbol{\pi}(\mathbf{h}) \cdot \mathbf{d}}{\mathbf{c}(\mathbf{h}) \cdot \mathbf{h}} \\ &= 1 - \frac{\boldsymbol{\pi}(\mathbf{h}) \cdot \mathbf{d}}{\mathbf{t}(\mathbf{f}) \cdot \mathbf{f}}. \end{aligned} \tag{8}$$

It measures the difference in system-wide travel cost calculated with current route costs $c_r(\mathbf{h})$ or link costs $t_a(\mathbf{f})$, and the travel cost calculated using the least cost for every OD pair, $\boldsymbol{\pi}(\mathbf{h})$. One can also think of it as measuring the flow in longer-than-minimal routes weighted by the extra length. As $c_r(\mathbf{h}) = \pi_{pq}(\mathbf{h})$ for $r \in \mathcal{R}_{pq}$ at equilibrium, the relative gap is zero at equilibrium.

As discussed in Slavin and Rabinowicz (2006), relative gap has several advantages over other possible convergence criteria. It is computable for both link-based algorithms like Frank-Wolfe and path-based algorithms. It does not depend on the difference between successive iterations, so can be used to compare different algorithms. It is scaled by the total travel time, so can be applied uniformly to different networks.

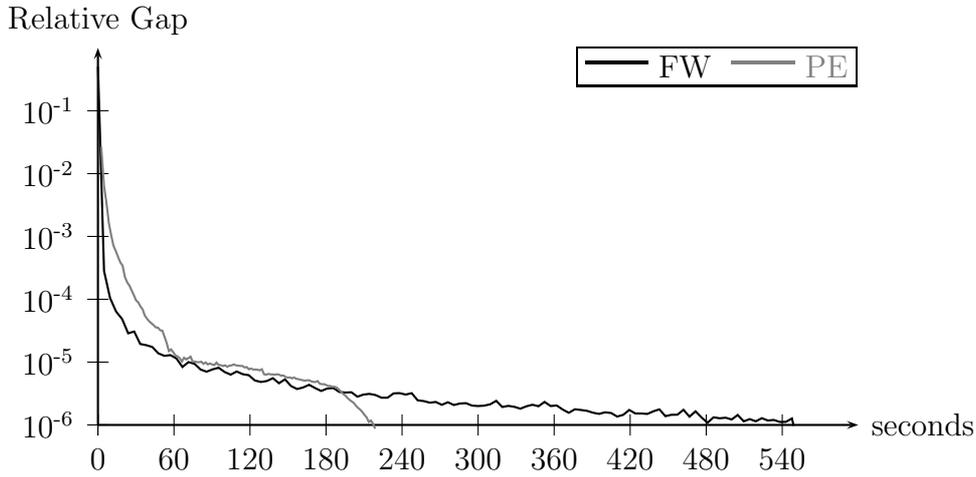


Figure 1: Relative gap vs CPU time, for Winnipeg test network

4 The Centroid Aggregation Algorithm

It is easily recognised that, in order for traffic assignment models to be tractable, we must use a simplified representation of the real transport infrastructure; we simply cannot take every piece of data into account when simulating route choice. An obvious way of simplifying is the use of zones as origins and destinations. Zone centroids stand in for the many and varied origins and destinations spread across a whole district. The number and size of zones used is largely determined by the level of detail in the data available to the modeler; having large zones means that values in the demand matrix are large enough (dozens to hundreds) that statistical methods can be used to derive them from surveys.

The central idea behind the centroid aggregation (CA) algorithm is that reducing the number of zones has a very large effect on the computational effort required to solve TAP. The solution of the simplified problem can then determine the assignment of long-distance flows in an initial solution to the full network TAP. Grouping together centroids with α centroids per group should give a simplified problem that solves in roughly $\frac{1}{\alpha^2}$ the time, having that many fewer OD pairs. The simplified problem can be solved with any TAP algorithm. For the purposes of mapping the solution back to the full network (described below), we prefer a method giving a route-flow solution. We use path-equilibration.

We define an aggregation scheme as a partition of \mathcal{Z} . It describes how the zones are grouped together into the aggregated zones.

Algorithm 3 Centroid Aggregation Algorithm

- Load network and aggregation scheme.
 - Create simplified network from full network using aggregation scheme.
 - Call AON assignment for initial solution.
 - Call route-based TA algorithm on simplified network to obtain macro-solution.
 - Map macro-solution to full network as a warm-start solution.
 - Call TA algorithm to obtain full solution.
-

Figure 2a shows a network with four zones. Each zone is represented by a special non-traversable centroid node. Centroid connectors are shown as dotted lines, they

stand in for the minor streets and paths that feed into main roads and transit routes. The other links correspond to actual roads whose traffic flows and congestion levels we are interested in modelling. For simplicity suppose that all demand is from A and B to C and D , so all flows shown will be left-to-right. In figure 2b, the zones have been aggregated under the scheme $\{\{A, B\}, \{C, D\}\}$. The active routes in equilibrium (those with positive flow) are marked in bold.

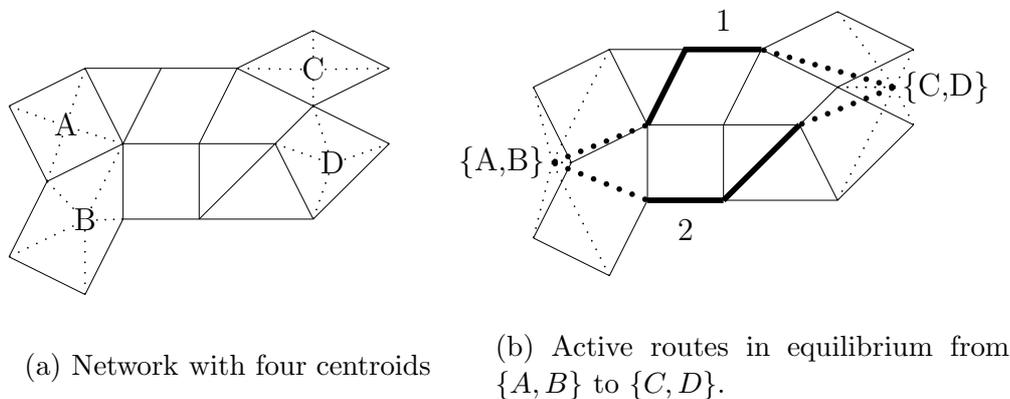
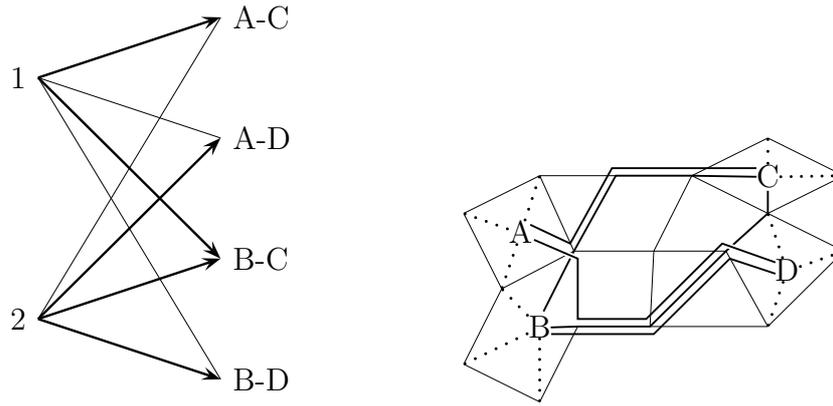


Figure 2: An example of zone aggregation

From the simplified network we must now map the equilibrated solution back to the unaggregated network. We call this mapping *untangling*. The problem of dividing the flow from one aggregated zone to another among the different origin-destination pairs can be expressed as a transportation sub-problem. The sources are the routes in the simplified solution, supplying the flow that they carry. The sinks are the origin-destination pairs with origin in the origin aggregated zone and destination in the destination aggregated zone. We define the cost of using route r to get from p to q as the distance from p to the first non-zone node along r plus the distance from the last non-zone to q . This is similar to how a human might navigate the traffic network, first choosing a major road and then solving the more local problem of how to get on and off that road at either end of the journey. Figure 3a shows an example of what an optimal solution to the sub-problem in our example network might look like. This transportation problem is solved using the network simplex method. Figure 3b shows the resulting untangled routes in the original network. Here it is easy to see that, for the search to be meaningful, zones that are grouped together should be close to each other so that they share routes to other parts of the network.

The last thing we must do before warm-starting is to assign local traffic within zonal groupings (e.g. between A and B or C and D) as this is traffic not considered at all in the simplified network. For now this is done by all-or-nothing assignment but we suspect that solving another TA sub-problem at this point may be more efficient overall see section 6 for more discussion.

From our untangled solution, which accurately models long-range flows, a standard TA algorithm should converge much quicker than from a standard all-or-nothing initial solution.



(a) Transportation subproblem optimal solution (b) Active routes between untangled OD pairs

Figure 3: Allocating flow to OD pairs and resulting routes in the full network

4.1 Generating node partitions

For small and medium sized networks it is fairly simple to group zone centroids by looking at a map or zonal distance table. However, on a large network with thousands of origins and destinations, manually grouping nodes could take hours. In an attempt to avoid unnecessary work, we have implemented a very basic grouping procedure which automatically groups nodes.

Algorithm 4 Greedy Aggregator Algorithm

```

Load traffic network, zone aggregation radius  $\rho$ .
Set  $\hat{\mathcal{Z}} = \mathcal{Z}$ , the list of unassigned zones
while  $\hat{\mathcal{Z}}$  non-empty do
  Choose a centre  $z_0 \in \hat{\mathcal{Z}}$ , place it in a new grouping  $\mathcal{Z}_0$  and remove it from  $\hat{\mathcal{Z}}$ .
  for all  $z \in \hat{\mathcal{Z}}$  do
    if  $d(z, z_0) \leq \rho$  then
      Put  $z$  in  $\mathcal{Z}_0$  and remove it from  $\hat{\mathcal{Z}}$ .
    end if
  end for
end while

```

Algorithm 4 aggregates all available zones within a specified distance of the current centre zone. The distance measure can be free-flow travel time, on-the-ground distance or number of links.

5 Results

5.1 Comparison with existing algorithms

We tested our algorithm on two medium-sized networks from the website of Bar-Gera (2012), representing Barcelona and Winnipeg. The computer used had an Intel Core2 Duo CPU E8400 @ 3.00GHz 2 processor with 4GB RAM running Ubuntu

12.04. All algorithms were implemented in C++. We were fortunate to inherit implementations of the Frank-Wolfe and path equilibration algorithms from a colleague. These have been incorporated into our new CA implementation. Aggregations were generated manually using origin-destination distance tables to identify groups of two to six nearby zones. Table 1 shows the CPU-time for the various algorithms — Frank-Wolfe (FW), path equilibration (PE) and centroid aggregation (CA) with two second-stage algorithms and two methods for the transportation subproblem: the network simplex method (NS) and uniform untangling (UU) that distributes flow uniformly across all possible path-origin-destination combinations without regard to cost.

City	FW	PE	CA			
			FW		PE	
			NS	UU	NS	UU
Barcelona	123.9	57.5	132.1	127.9	71.0	61.5
Winnipeg	552.0	218.0	560.0	566.3	95.7	171.0

Table 1: CPU-time (s) to converge to relative gap $< 10^{-6}$

These initial results are very encouraging for Winnipeg, less so for Barcelona. The Barcelona network has a several particular features, one of which may be to blame for the ineffectiveness of our new algorithm. There is a very dense concentration of centroids in one part of the network, so local traffic here may be more important for congestion than otherwise. Also, the power P in some of the link-functions (1) are very high — larger than 16 — so parts of the network respond more severely to congestion than others. We have not yet figured out how to take these features into account when constructing an aggregation scheme and there may still exist an aggregation that makes CA faster.

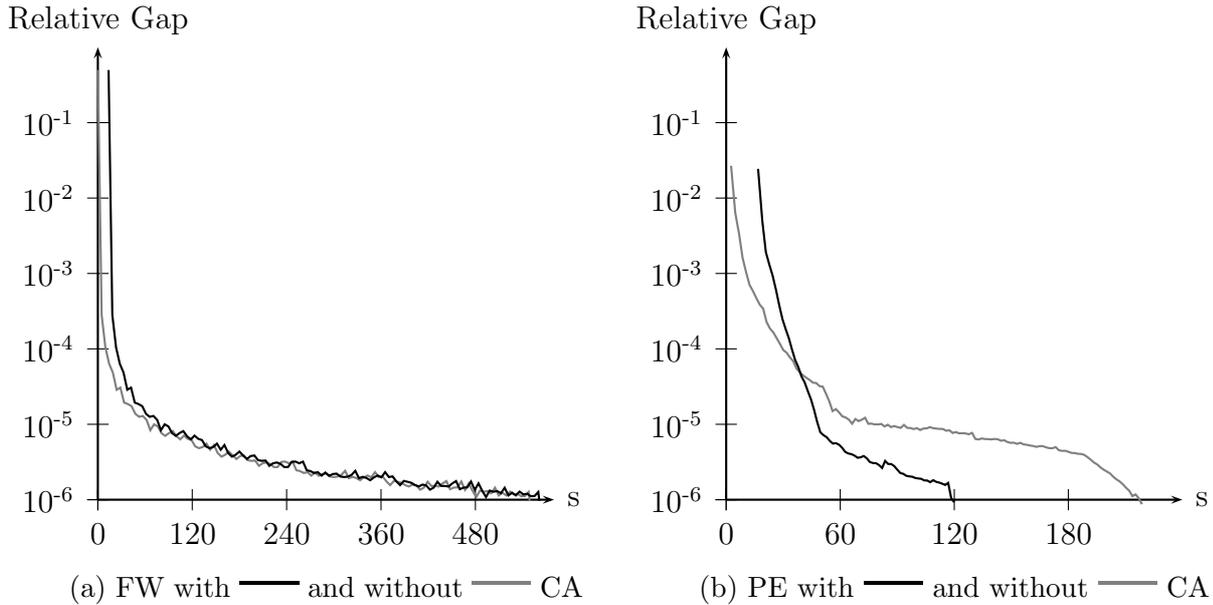
5.2 Rate of convergence

From the above it may appear that the Frank-Wolfe algorithm is hopeless, however this depends very much on how exact one wishes to calculate the equilibrium. Figure 4a shows the decrease in relative gap over time for the Frank-Wolfe algorithm, both on its own and as part of our CA algorithm. It is clear to see that below a relative gap value of 10^{-4} , every order of precision costs much more time. It is also clear that the CA algorithm gives no improvement in this case.

For the centroid aggregation algorithm using path equilibration to solve the full network, there was a substantial improvement in computation time. Figure 4b shows that using CA, the PE algorithm converges very quickly and overtakes pure PE around 45s in.

Note that the CA graph begins about 13s later than that of the other algorithms. This gap represents the time taken to solve the simplified network and initialise the full problem from that solution. The all-or-nothing assignment used in pure PE on the other hand, takes only several hundredths of a second of computer time.

Figure 4: Relative gap vs CPU time, Winnipeg test network



5.3 Automatic aggregation

We tested the basic greedy aggregator described in Algorithm 4, using number of links as unit of distance, against manually generated partitions. Results are given in Table 2, comparing pure PE with CA using both greedy aggregation and a human-devised aggregation scheme. The number of zones in the simplified network is given, as well as the total CPU-time taken, including the time to solve the simplified network and untangle that solution. Clearly there is much room for improvement in designing an aggregator that makes as effective a partition of the zones as a human can. The automatic aggregator only gives a slight (if any) improvement over pure PE. With more testing, it should become clearer what the properties of an effective partition are — for instance, should it behave differently in areas of high demand, or in central areas?

Table 2: CPU-time to converge to relative gap $< 10^{-6}$ for different aggregation radii

Algorithm	PE	CA				
		8	10	12	14	human
Zone aggregation radius	-	50	36	29	21	36
Zones after aggregation	(147)	50	36	29	21	36
CPU-time (s)	218	231	230	210	211	120

6 Further Refinements

The above results were attained from a first implementation of the CA algorithm. There are several possible improvements to the algorithm that plan to implement.

6.1 The local traffic sub-problem

The present implementation of CA only assigns local traffic (flows within the same grouping of zones) to all-or-nothing shortest paths. Of course, these flows get equilibrated by the algorithm used in the full run. However, it may be advantageous to

better equilibrate this traffic at the moment of untangling.

6.2 Obtaining route-flows from link-flows

Frank-Wolfe converges very quickly at first, in the sense that relative gap decreases at the greatest rate per second (see Figure 1. Unfortunately, our untangling procedure requires path-flows in order to enforce feasibility. If we can obtain route-flows cheaply from link-flows then this may speed up the equilibration of the simplified network in the CA algorithm. See Li and Wang (2005) for an algorithm to derive a route-flow solution from link-flows. Since solving the simplified network does not take long using the path-equilibration algorithm (typically 5 to 10s for the medium-sized networks we have tested), we would probably do better to focus our attention on attaining more equilibrated results from the untangling procedure.

7 Conclusion

What we have presented here is only the first part of our exploration of the centroid aggregation idea. We have shown that our method is implementable and is even competitive with existing algorithms for TAP. With more work we hope to refine our centroid aggregation algorithm to make it even more competitive.

Acknowledgements

Many thanks to Olga Perederieieva for the use of her computer code.

References

- Bar-Gera, Hillel. 2012. “Transportation Test Problems.” Last updated on January 12, 2011. <http://www.bgu.ac.il/~bargera/tntp/>.
- Dafermos, Stella C., and Frederick T. Sparrow. 1969. “The Traffic Assignment Problem for a General Network.” *Journal of the National Bureau of Standards - B* 738(2):91–118.
- Li, Feng, and Shuning Wang. 2005. “Determining route traffic flows for traffic assignment problem with Frank-Wolfe algorithm.” *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. 669 – 673.
- Ortúzar, Juan de Dios, and Luis G. Willumsen. 2001. *Modelling Transport*. Wiley, New York.
- Patriksson. 1994. *The Traffic Assignment Problem: models and methods*. VSP, Utrecht.
- Slavin, Howard, Jonathan Brandon, and Andres Rabinowicz. 2006. “Empirical Comparison of Alternative Equilibrium Assignment Methods.” *Proceedings of the European Transport Conference*.
- Wardrop, John Glen. 1952. “Some Theoretical Aspects of Road Traffic Research.” *Proceedings of the Institute of Civil Engineers, London, Part II* 1:325–378.

It is Time to Enjoy the Best of Both Worlds

David Ryan
Department of Engineering Science
The University of Auckland
New Zealand
d.ryan@auckland.ac.nz

Abstract

Since George Dantzig's famous discovery of the Simplex Algorithm more than 60 years ago, researchers have worked to develop and enhance algorithms to generate solutions for many practical decision problems in Operations Research. Two important classes of algorithms have evolved during this time: these can be broadly classified as mathematical optimization methods and heuristic methods. Optimization based methods have the important properties that they are guaranteed to find a feasible solution if one exists and furthermore, they are able to issue a certificate of optimality for the solution they produce. These two attractive properties come at some cost in that the computation required is often relatively large and in some cases the optimization methods can become prohibitively expensive as the problem size increases. Heuristic methods on the other hand are much less sensitive to problem size but they suffer from two disadvantages in that they are not guaranteed to find a feasible solution even if one exists and they can provide no guarantee of optimality or even how far from optimality the heuristic solution is. It is interesting to observe that the research community interested in these decision problems seems to naturally partition into two almost disjoint groups aligned with these classes of methods. In this talk we will argue that the time has come to explore the development of composite methods that exploit the benefits of heuristics to improve the performance of optimization methods or optimization methods to solve sub-problems within a heuristic framework.

Key words: Optimization, Heuristics, Composite methods.

Evaluating the Impact of Systems Thinking Workshops on Strategy Implementation in a Government Department

Rodney J. Scott
University of Queensland, Queensland, Australia
rodney.scott@gmail.com

Robert Y. Cavana
Victoria Business School, Victoria University of Wellington, New Zealand

Donald Cameron
School of Agriculture and Food Sciences, University of Queensland, Australia

Abstract

Processes for effectively implementing strategy are yet to be fully explained and explored. Strategy literature reports communication quality, insight, consensus and commitment to conclusions as success factors that predict effective strategy implementation. These have also been reported as outcomes of group model building workshops, suggesting possible applicability.

This paper presents results of a systems thinking intervention to support the implementation of an organisation strategy in a New Zealand government department. Four separate three-hour systems thinking workshops were conducted with department employees.

A range of survey and work-sample methods were used to evaluate changes in communication quality, insight, consensus and commitment to conclusions. Post-workshop survey results and work-samples showed significant increases in the outcomes measured.

This paper informs management decisions on selecting tools for strategy implementation. This paper only evaluates immediately workshop outputs – further study is planned to explore the long-term impacts of group model building.

Key words: systems thinking, strategy, evaluation.

The Application of Linear Programming to Select the Lowest Cost and Optimized Quality of Textile Wet Process

Aioporn Sophonsridsuk
Department of Industrial Technology
Faculty of Technical Education
Rajamangala University of Technology Krungthep
Thailand
dr.aioporn@gmail.com

Abstract

Operational Research can apply for many industrial processes. The main aims are to select the lowest cost, highest profit and best quality of product.

There are many researches done in the applications of operational research to textile process. They used operational research to manage textile factory. In production of mixed yarn, they used operational research to make decision to adjust the ratio of cotton fiber and polyester filament to make the best quality of cotton blended yarn at lowest price, etc.

For this research has been studied the lowest cost of bleaching process but in optimized performances of textile fabric before dyeing. In bleaching process, there are many recipes that differ in costs of chemical, water and energy usage. And we want the fabric that is suitable for next processes (dyeing, printing, finishing, etc.)

The results of this study, it shows the way to select the best way of using optimized bleaching chemical and process that make lowest cost and best quality by the knowledge of linear programming (calculate with Excel Solver).

Key words: Linear programming, Textile Wet Process, Bleaching Process.

1 Introduction

Operational research is mathematical method that helps both business and industrial sectors to make more profit and more friendly to environment. Terrazas-Moreno, Grossmann and Wassick (2012) show that chemical manufacturing sites ship finished products to customers using different modes of transportation (MOT) such as railcars, tank trucks, and pipelines. These MOT are usually loaded from or connected to storage tanks. Consequently, all finished products have to be fed from the process into the storage tanks before being shipped to customers. This type of operation imposes the need for available storage space at all times in order to avoid unnecessary shut-downs of the upstream chemical process. When these shutdowns occur, they are said to be a result of storage tanks blocking the process. If the chemical process produces several products and each one requires dedicated tanks, the product-tank assignment and the processing schedule at the finishing lines determines how efficiently the storage space is used. An inefficient assignment of tanks or processing schedule can result in blocking the production of certain products, even when there is plenty of available storage space

in tanks assigned to other products. In textile field, there are many cases that have been using linear programming to improve quality of textile product. TextileTechInfo.Com (2012) shows the application of linear programming to improve quality and minimize cost of product mix in spinning mill.

D.S. Hartley III (<http://home.comcast.net/~dshartley3/INDUSTRY/Loom.htm>) has been showing the application of linear programming in textile industry; looms are the machines used to weave cloth, which is sold for a profit. However, looms can be set up to weave various kinds of cloth, each with its own profit structure and customer demand. The heart of the problem (in the late 1970s) lay in deciding what mix of products to make with a general goal of improving profits. The caveats were what made the problem interesting. For instance, profits might be sacrificed to retain long term market share and while forecast and actual future (booked orders) demand (and prices) varied from month to month, overly radical shifts in the mix would not be permitted. Further, the software and hardware of the time were definitely not designed for interaction with managers: the glass cage and lab-coated computer servants were a reality; external terminals were available, but not ubiquitous; and the IBM linear programming software was designed to be operated only by an expert. Despite these impediments, the task of manually creating a loom plan each month was formidable and the potential for improvement looked large enough to investigate computer-based optimization. For cotton blended spinning process has state by International Business Machines Corporation (1965). The purpose of this manual is to demonstrate the application of LP in the blending of cotton. Because the cotton blending process involves complex quality control, it is particularly responsive to LP techniques. By the use of LP, the mill operator can determine the specific allocation of raw cottons required to produce a given blended yarn at minimum cost – subject to any stated restrictions on yarn quality and raw cotton availabilities. The immediate and more obvious LP results enable the mill operator to:

- Minimize the cost of cotton blends
- Minimize substandard blends
- Maintain accurate inventory records
- Purchase and sell most economically.

In bleaching process, there are many recipes from many textile chemistry companies that vary in cost of bleaching chemical. This research wants to find the best recipe (recipe is the process and amount chemicals that use in some wet process) at the lowest cost by using method of linear programming to find the optimized cost of recipes and optimized amount of textile that uses in those recipes. (Based on 1 kg of cotton fabric, if we want to operate in any size of textile chemistry machine it can expand by changing the ratio of researched fabric to be real amount of that textile chemistry machine.)

2 Linear programming

Murthy (2007) states that linear programming is one of the most versatile, powerful and useful techniques for making managerial decisions. Linear programming technique may be used for solving broad range of problems in business, government, industry, hospitals, libraries, etc. Whenever we want to allocate the available limited resources for various competing activities for achieving our desired objective, the technique that helps us is linear programming. As a decision making tool, it has demonstrated its value in various fields such as production, finance, marketing, research and development and personnel management. Determination of optimal product mix (a combination of products, which gives maximum profit), transportation schedules, assignment problem and many more.

3 Textile wet process

Textile wet process is an important process that will change the properties of textile yarn or fabric. This process begins with bleaching and scouring process to prepare yarn or fabric for suitable for next process that will be dyeing process, printing process, finishing process.

From Textile Learner (2011), in which way grey fabric is dyed is called wet process technology. Normally wet processing depends on buyer's demand. Suppose your buyer wants the more precised dyed fabric; so in this fact you should mercerize your fabric during the dyeing pre- treatment process. Basically if the buyer doesn't want that so called particular fabric there is no need to mercerize your fabric.

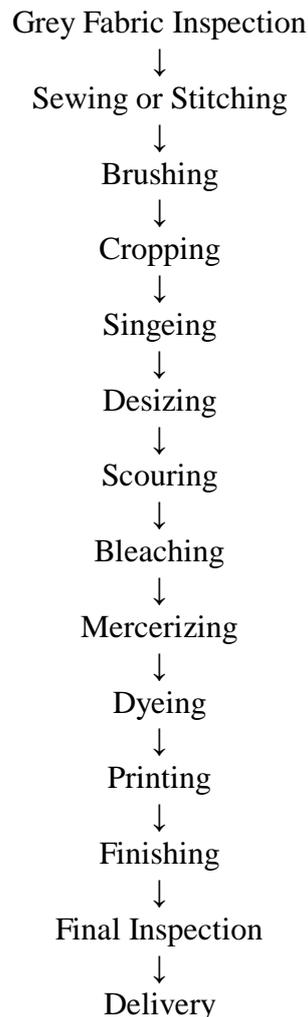


Figure 1. Flow Chart of Wet Process Technology

Grey Fabric Inspection:

After manufacturing fabric it is inspected in an inspection table. It is the process to remove neps, warp end breakage, weft end breakage, hole spot.

Stitching:

To increase the length of the fabric for making suitable for processing is called stitching. It is done by plain sewing machine.

Brushing:

To remove the dirt, dust, loose fibre & loose ends of the warp & weft threads is known as brushing.

Shearing/Cropping:

The process by which the attached ends of the warp & weft thread are removed by cutting by the knives or blades is called shearing. Shearing is done for cotton & cropping for jute. After shearing or cropping fabrics goes under singeing process.

Singeing:

The process by which the protruding/projecting fibres are removed from the fabrics by burning/heat to increase the smoothness of the fabric is called singeing. If required both sides of fabric are singed.

Desizing:

The process by which the sizing materials (starch) are removed from the fabric is known as desizing. This must be done before printing.

Scouring:

The process by which the natural impurities (oil, wax, fat etc.) and added /external/adventitious impurities (dirt, dust etc) are removed from the fabric is called scouring. It is done by strong NaOH.

Souring:

The process by which the alkalis are removed from the scoured fabric with dilute acid solution is known as souring.

Bleaching:

The process by which the natural colours (nitrogenous substance) are removed from the fabric to make the fabric pure & permanent white is known as bleaching. It is done by bleaching agent.

Mercerizing:

The process by which the cellulosic materials/substance are treated with highly conc. NaOH to impart some properties such as strength, absorbency capacity, lusture is known as mercerizing. It is optional. If the fabrics are 100% export oriented then it is done by highly conc. NaOH (48-52° Tw).

Dyeing:

A process of coloring fibers, yarns, or fabrics with either natural or synthetic dyes.

Printing:

A process for producing a pattern on yarns, warp, fabric, or carpet by any of a large number of printing methods. The color or other treating material, usually in the form of a paste, is deposited onto the fabric which is then usually treated with steam, heat, or chemicals for fixation. Then finishing treatment is done according to buyer requirements and then folding, packaging, and at last delivery.

4 Bleaching process

Georgievics (1902) states that broadly considered, any operation performed with the object of cleansing a textile fibre at any stage of manufacture or improvement may be included in this category; but, in their more restricted sense, washing and bleaching are terms confined to the cleansing of the textile fibres in their crude state in order to prepare them for the subsequent operations of dyeing, dressing, etc. Thus, in speaking of wool-washing, the first process of purification to which the wool is subjected before spinning is meant, and not the rinsing of dyed wool, for example. Corresponding operations are not always described by the same names in different branches of the textile industry; and, conversely, the same term is differently applied. Thus the term "cotton bleaching" comprises the entire set of operations employed to free the cellulose from all natural and other impurities, whereas, in the case of silk and wool, the term "bleaching" implies only the more restricted sense of the word, being confined to the operations affecting the actual decoloration of the fibre. Again, the first cleansing process applied to raw silk is not called "washing", as in the case of wool, but is termed "scouring," or "removing the bast".

5 Dyeing process

From Georgievics (1902), the two principal factors influencing the method of performing any operation of dyeing, and the behaviour of the colour when finished, are the dye-stuff and the fibre. Dye-stuffs behave variously in dyeing, the difference being dependent on their chemical composition and on the nature of the fibre to which they are applied. An examination of the classified dye-stuffs shows that the mutual relation of the members of each class depends more on a single property held by them in common than on their constitution. Thus all the numerous acid dyes, be their constitution never so divergent, behave in a perfectly analogous manner when applied to the different textile fibres, and exhibit no fundamental differences.

6 Experiments

This research is finding the way to use the lowest cost of bleaching chemical and water in the process of textile bleaching. There are methods of testing:

- 6.1 Bleaching the cotton fabrics with 12 recipes of various companies' bleaching chemicals.**
- 6.2 Test whiteness value with AATCC (American Association of Textiles Chemist and Colorists) method.**
- 6.3 Dyeing the 12 bleached fabrics in 6.2 with reactive dye.**
- 6.4 From 6.3 find %reflectant value of 12 dyeing fabric with spectronic 20 machine.**
- 6.5 Calculate the minimized cost of bleaching chemical and optimized quantity of fabric with linear programming with Excel solver and linear program solver.**

7 Results

7.1 Whiteness of 12 recipes bleaching cotton fabrics.

Table 1. 12 Recipes bleaching fabric with D, C, T parameter

	D parameter	C parameter	T parameter
Standard Value	25.48	24.44	33.34
Recipe1	65.83	65.55	68.28
Recipe2	67.14	66.89	69.36
Recipe3	35.73	34.90	33.34
Recipe4	61.83	61.46	64.87
Recipe5	61.61	61.28	64.11
Recipe6	59.03	58.60	62.34
Recipe7	63.64	63.29	66.34
Recipe8	62.26	61.82	65.08
Recipe9	60.81	60.43	63.69
Recipe10	65.34	65.01	67.68
Recipe11	42.64	41.96	48.33
Recipe12	42.42	41.73	48.17

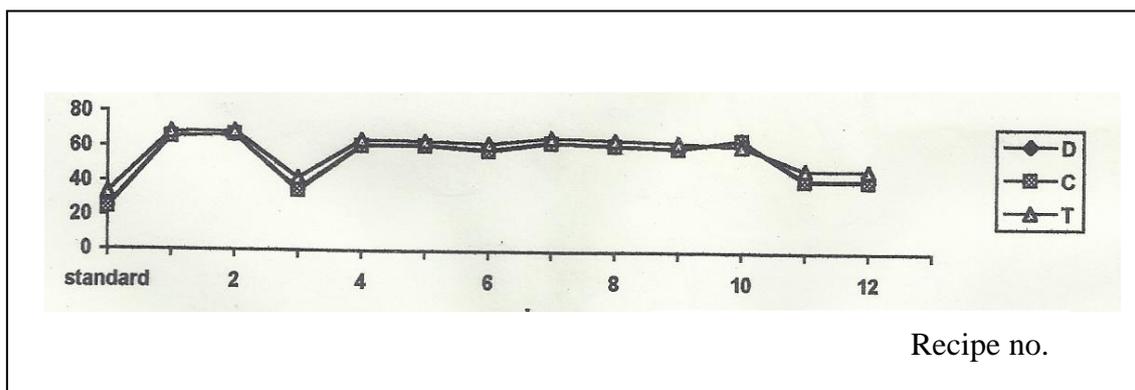


Figure 2. 12 recipes bleaching fabrics

7.2 %Reflectants of dyeing of 12 recipes bleaching cotton fabrics.

Table 2. %Reflectants of dyeing of 12 recipes bleaching cotton fabrics

	1	2	3	4	5	6	7	8	9	10	11	12
400	23.50	23.30	22.8	22.60	23.10	23.40	23.20	22.60	22.30	22.63	21.94	24.23
420	26.57	26.86	25.66	26.25	26.28	27.45	25.85	26.07	26.02	25.82	25.13	27.51
440	26.43	26.71	25.67	26.19	26.25	27.38	26.26	25.99	25.97	25.80	25.21	27.29
460	23.99	24.22	23.25	23.87	23.80	24.82	25.10	23.54	23.52	23.64	22.89	24.80
480	19.56	19.87	18.91	29.40	19.44	20.37	22.03	19.12	19.09	19.40	18.62	24.31
500	16.37	16.60	15.53	16.26	16.23	17.08	19.25	15.92	15.86	16.36	15.54	16.99
520	12.58	12.87	12.12	12.46	12.50	13.26	15.44	12.21	12.15	12.97	11.93	13.15
540	10.18	10.40	9.79	10.06	10.07	10.25	12.79	9.79	9.77	10.27	9.61	10.61
560	7.87	8.13	7.59	7.79	7.84	8.37	10.70	7.57	7.57	7.97	7.44	8.28
580	6.65	6.35	5.90	6.07	6.11	6.55	8.08	5.19	5.86	6.23	5.78	6.44
600	5.29	5.47	5.11	5.23	5.27	5.62	7.09	5.09	5.10	5.36	4.99	5.55
620	5.01	5.18	4.88	4.90	5.07	5.36	6.72	4.84	4.83	5.09	4.74	5.26
640	5.47	5.67	5.26	5.40	5.412	5.81	7.28	5.24	5.25	5.53	5.12	5.72
660	8.31	8.60	7.95	8.12	8.12	8.78	10.76	8.00	7.98	8.41	7.88	8.79
680	14.75	15.34	14.13	14.36	14.30	15.50	18.13	14.35	14.36	15.01	14.23	15.65
700	24.16	25.00	23.30	23.58	23.40	25.01	27.98	23.70	23.72	24.41	23.48	25.31

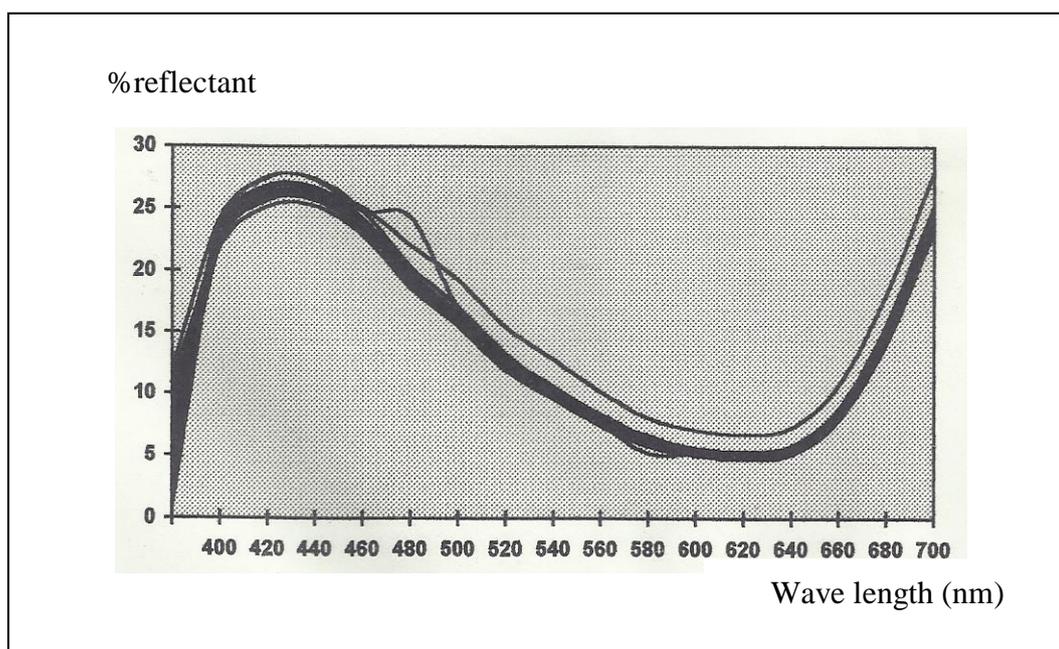


Figure 3. %Reflectants of dyeing of 12 recipes bleaching cotton fabrics

8 Finding optimized solution

8.1 Cost of Bleaching Chemical and water of 12 recipes.

Table 3. Cost of Bleaching Chemical and water of 12 recipes.

Recipe	No1	No2	No3	No4	No5	No6	No7	No8	No9	No10	No11	No12
Cost (Thai baht)	28.65	39.01	18.20	30.54	23.06	44.80	63.30	43.33	33.64	40.78	18.64	38.20

(1 Thai baht = 0.03 US Dollar)

Make optimal solution

$$\text{Min}Z = 28.65 \cdot X_1 + 39.01 \cdot X_2 + 18.2 \cdot X_3 + 30.54 \cdot X_4 + 23.06 \cdot X_5 + 44.8 \cdot X_6 + 63.3 \cdot X_7 + 43.33 \cdot X_8 + 33.64 \cdot X_9 + 40.78 \cdot X_{10} + 18.64 \cdot X_{11} + 38.2 \cdot X_{12}$$

X_i = amount (kg) of fabric that will be bleached with recipe i

8.2 Whiteness of 12 recipes bleaching cotton fabrics.

Put D, C, T value from Table 1 in to make constraint:

$$\text{D value: } 65.83 \cdot X_1 + 67.14 \cdot X_2 + 35.73 \cdot X_3 + 61.83 \cdot X_4 + 61.61 \cdot X_5 + 59.03 \cdot X_6 + 63.64 \cdot X_7 + 62.26 \cdot X_8 + 60.81 \cdot X_9 + 65.34 \cdot X_{10} + 42.64 \cdot X_{11} + 42.42 \cdot X_{12} \geq 25.48;$$

$$\text{C value: } 65.55 \cdot X_1 + 66.89 \cdot X_2 + 34.9 \cdot X_3 + 61.46 \cdot X_4 + 61.28 \cdot X_5 + 58.6 \cdot X_6 + 63.29 \cdot X_7 + 61.82 \cdot X_8 + 60.43 \cdot X_9 + 65.01 \cdot X_{10} + 41.96 \cdot X_{11} + 41.73 \cdot X_{12} \geq 24.44;$$

$$\text{T value: } 68.28 \cdot X_1 + 69.36 \cdot X_2 + 33.34 \cdot X_3 + 64.87 \cdot X_4 + 64.11 \cdot X_5 + 62.34 \cdot X_6 + 66.34 \cdot X_7 + 65.08 \cdot X_8 + 63.69 \cdot X_9 + 67.68 \cdot X_{10} + 48.33 \cdot X_{11} + 48.17 \cdot X_{12} \geq 33.34;$$

25.48, 24.44, 33.34 are standard values

8.3 %Reflectants of dyeing of 12 recipes bleaching cotton fabrics.

From Figure 3 and Table 2, we use %reflectant at 620 nm. Because at that point, it is the lowest reflectant value, it means that it is the best point of this dyeing process. Put values from Table 2 in this constraint:

$$\% \text{Reflectant: } 5.01 \cdot X_1 + 5.18 \cdot X_2 + 4.88 \cdot X_3 + 4.9 \cdot X_4 + 5.07 \cdot X_5 + 5.36 \cdot X_6 + 6.72 \cdot X_7 + 4.84 \cdot X_8 + 4.83 \cdot X_9 + 5.09 \cdot X_{10} + 4.74 \cdot X_{11} + 5.26 \cdot X_{12} \geq 4.74;$$

4.74 is norm value.

8.4 Set constraint for 1 kg of bleaching cotton fabric from all recipes.

$$X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 = 1 \text{ kg.}$$

8.5 Find the minimized cost of bleaching chemical and optimized quantity of fabric with linear programming with Excel solver and linear program solver.

```
>> Optimal solution FOUND
>> Minimum = 18.2
```

*** RESULTS - VARIABLES ***

Variable	Value	Obj. Cost	Reduced Cost
x1	0	28.56	-10.36
x2	0	39.01	-20.81
x3	1	18.2	0
x4	0	30.54	-12.34
x5	0	23.06	-4.86
x6	0	44.8	-26.6
x7	0	63.3	-45.1
x8	0	43.33	-25.13
x9	0	33.64	-15.44
x10	0	40.78	-22.58
x11	0	18.64	-0.44
x12	0	38.2	-20

*** RESULTS - CONSTRAINTS ***

Constraint	Value	RHS	Dual Price
Row1	35.73	25.48	0
Row2	34.9	24.44	0
Row3	33.34	33.34	0
Row4	4.88	4.74	0
Row5	1	1	18.2

From this, we find that the minimized cost is at 18.20 Thai baht for 1 kg of bleaching fabric. It should bleach the fabric with recipe 3.

Acknowledgments

I would like to say thank you to Rajamangala University of Technology Krungthep to give me a chance to do this research. For my mother, she is the person who gives me the full love and support.

9 References

International Business Machines Corporation. 1965. *Linear Programming-Cotton Blending and Production Allocation*, IBM, Technical Publications Department, 112 East Post Road, White Plains, N.Y. 10601.

Georgievics, G.V. 1902. *The Chemical Technology of Textile Fibres*. Scott, Greenwood & Co, London.

Hartley III, D.S. "Optimization in the textile industry."
<http://home.comcast.net/~dshartley3/INDUSTRY/Loom.htm>

Murthy, P.R. 2007. *Operations Research*. New Age International Publishers, New Delhi.

Terrazas-Moreno, S., I. E. Grossmann, and J.M. Wassick. 2012. "A mixed-integer linear programming model for optimizing the scheduling and assignment of tank farm operations." *Industrial & Engineering Chemistry Research*, **51**: 6441-6454.

Textile Learner: textile information blog for students. 2011.
<http://textilelearner.blogspot.com/2011/08/flow-chart-of-wet-processing-process.html>

TextileTechInfo.Com. 2012. <http://textiletechinfo.com/spinning/linearprogramming.htm>

On the Multicriteria Linear Bottleneck Assignment Problem

Michael Stiglmayr
Department of Mathematics and Informatics
University of Wuppertal
Germany
stiglmayr@math.uni-wuppertal.de

Abstract

We present a solution method for the multicriteria linear bottleneck assignment problem (MLBAP), which is the multicriteria analogon of the well studied linear bottleneck assignment problem. The algorithm given in this paper is an adaption and extension of the single criteria threshold algorithm. We define a residual graph by specifying a (multicriteria) threshold vector, such that all of its edges have cost dominated by the threshold. Any perfect matching in this residual graph is a candidate for an efficient solution with at least this threshold value. The computation of matchings in the residual graph is realized by an efficient update of augmenting paths.

Key words: multicriteria optimization, assignment problem, bottleneck objective, combinatorial optimization

1 Introduction and Literature Review

We consider the multicriteria bottleneck assignment problem as the multicriteria extension of the *linear bottleneck assignment problem* (LBAP), which is well known and broadly used for many applications in production planing and scheduling. LBAP was proposed first in the article Fulkerson, Glicksberg, and Gross (1953), as a model for the following problem:

Given n jobs and n machines. In a production process every job should be performed on one machine in sequence. Let $a_{i,j} \geq 0$ be the number of jobs of type j can be performed by machine i per time unit. How should the jobs assigned to the machines such that the rate of production is maximized?

Thus, one searches for a permutation π , which maximizes the bottleneck of the production rates on all machines. This results in the following max-min formulation:

$$\begin{aligned} \max \quad & \min_i \{a_{i,\pi(i)}\} \\ \text{s. t.} \quad & \pi \in \mathcal{S}(n) \end{aligned} \tag{1}$$

In Garfinkel (1971) the equivalent min-max-formulation is presented as the model for the following problem: Let the number of jobs and the number of machines be

n . As before, every machine can process every job, but in different amount of time. So let $c_{ij} \geq 0$ be the time machine i needs to complete job j . In contrast to the previous problem formulation the machines now are considered to run in parallel. Then assignment which minimizes the time required to finish all jobs can be determined by a LBAP.

$$\begin{aligned}
& \min \max_{i,j} \{c_{ij} x_{ij}\} \\
& \text{s. t. } \sum_{i=1}^n x_{ij} = 1 \quad \forall j \in \{1, \dots, n\} \\
& \quad \sum_{j=1}^n x_{ij} = 1 \quad \forall i \in \{1, \dots, n\} \\
& \quad x_{ij} \in \{0, 1\} \quad \forall i, j \in \{1, \dots, n\}
\end{aligned} \tag{2}$$

The two most commonly applied solution methods for LBAP are threshold algorithms and dual methods. Several authors contributed to the development of LBAP. The LBAP was solved in Fulkerson, Glicksberg, and Gross (1953) using a transformation to the linear sum assignment problem (LSAP), which was efficiently solvable even in those days. However, this transformation approach runs into numerical problems for large instances of LBAP, since the transformed cost coefficients are exponentially growing with the problem size.

To overcome these difficulties, the threshold algorithm was proposed in Garfinkel (1971). The threshold algorithm can be summarized as follows:

1. Select a threshold vector $\epsilon \in \mathbb{R}$
2. Build the residual graph $G[\epsilon] = (V, \{(i, j) \in E : c_{ij} \leq \epsilon\})$
3. Test if $G[\epsilon]$ contains a perfect matching of G
4. Iterate by increasing/decreasing the threshold value, depending on whether a perfect matching is found or not.

Depending on the strategy of selecting and adapting a value for ϵ , different types implementations are at hand (iterative subdivision, stepwise in- or decrease).

The dual method (Burkard, Dell'Amico, and Martello 2009) is computational quite similar to the threshold algorithm. It uses a starting solution \tilde{c} , which corresponds to the fact that the residual graph $G[\tilde{c}]$ has to be connected to contain a perfect matching of G .

$$\tilde{c} = \max_{1 \leq k \leq n} \left(\min_{1 \leq i \leq n} c_{ik}, \min_{1 \leq j \leq n} c_{ki} \right)$$

Starting from \tilde{c} a minimum vertex cover of $G[\tilde{c}]$ is determined. While the size of the cover is larger than n , \tilde{c} is increased to the smallest cost value of an edge in G , whose endpoints are not both in the cover set. Note, that one yields immediately a minimum vertex cover, when computing a maximal matching (Burkard, Dell'Amico, and Martello 2009).

2 MLBAP and its Properties

Often when multiple conflicting interests are involved, a proper solution asks for a multicriteria analysis and optimization, modeling each interest as one objective function. So, we are considering the multicriteria version of LBAP, a multicriteria

optimization problem whose objectives are all bottleneck assignment functions (but differ in the cost coefficients). For example, the components of cost coefficients can represent the production times of a job on a machine in different scenarios or production situations at a factory. For a survey on multicriteria (combinatorial) optimization, see Ehrgott (2005) and Ehrgott and Gandibleux (2000).

While the multicriteria linear sum assignment problem is well studied (Malhotra, Bhatia, and Puri 1982; Przybylski, Gandibleux, and Ehrgott 2008), the multicriteria linear bottleneck problem has—to the best of our knowledge—not been investigated in literature before. However, there are articles considering other types of combinatorial optimization problems with bottleneck objective (Ehrgott and Klamroth 1997; Gorski, Klamroth, and Ruzika 2011).

Problem 1. Let $c^1, c^2, \dots, c^q \in \mathbb{R}^{n \times n}$ be the matrices of cost coefficients of the q bottleneck objective functions be given, then the multicriteria linear bottleneck assignment problem is stated as follows:

$$\begin{aligned}
 \min \quad & \begin{pmatrix} \max \{c_{ij}^1 x_{ij}\} \\ \max \{c_{ij}^2 x_{ij}\} \\ \vdots \\ \max \{c_{ij}^q x_{ij}\} \end{pmatrix} \\
 \text{s. t.} \quad & \sum_{i=1}^n x_{ij} = 1 \quad \forall j \in \{1, \dots, n\} \\
 & \sum_{j=1}^n x_{ij} = 1 \quad \forall i \in \{1, \dots, n\} \\
 & x_{ij} \in \{0, 1\} \quad \forall i, j \in \{1, \dots, n\}
 \end{aligned} \tag{MLBAP}$$

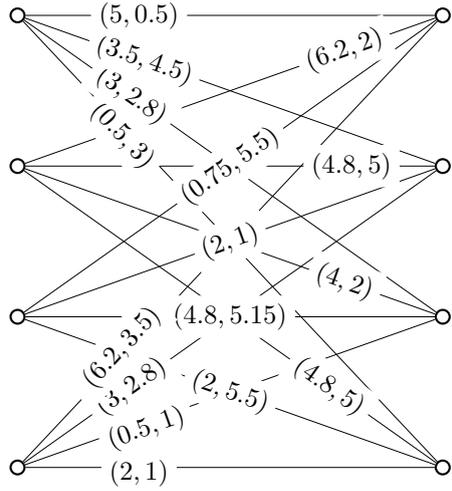
In the following we are presenting some basic properties of MLBAP, which are use to design our algorithm. The following lemma can be directly transferred from the single criteria case Burkard, Dell’Amico, and Martello (2009) to the multicriteria problem:

Lemma 2. Let $C = (c_{i,j}^k)$ be the $n \times n \times q$ -dimensional array of cost coefficients of an MLBAP. Then the following holds:

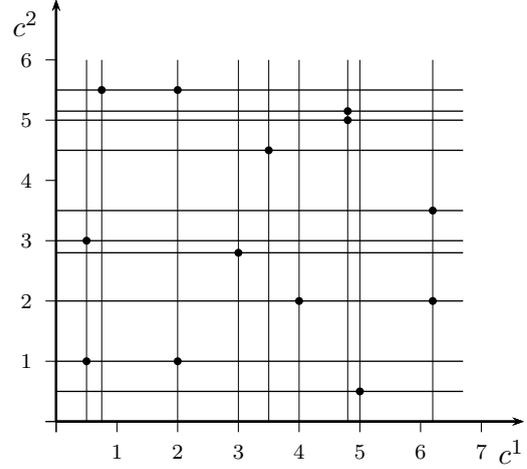
- Each component of every non-dominated solution attains a value of cost coefficients in the respective criterion. Let $\bar{f} = (\bar{f}^1, \dots, \bar{f}^q)^\top$ be a non-dominated solution, then it holds: $\bar{f}^1 \in \{c_{i,j}^1 : i, j \in I\}$, \dots , $\bar{f}^q \in \{c_{i,j}^q : i, j \in I\}$
- The set of efficient solutions depends only on the relative order of the cost coefficients in each criterion, and not on their numerical value.

Proof. Since in each component every non-dominated solution is the maximum over a finite set of coefficients, it is attained in one of these coefficients.

A necessary condition for a point y to be non-dominated is that the graph $G_y = (V, E_y)$ with $E_y = \{(i, j) : c_{ij} \leq y\}$ contains a perfect matching. The condition is sufficient, if no other point $y' \leq y$ meets this condition. These conditions depend only on the relative componentwise order of the cost coefficients and not on its numerical values. Thus, one can change the numerical values of the cost coefficients with out changing the set of efficient solutions, if the relative order of the cost coefficients remains the same. \square



(a) Bipartite graph with bicriteria cost vectors



(b) Cost vectors \mathbf{c}_{ij} (dots) and potential solutions (intersection points of the grid)

Figure 1: Grid illustration of a bicriteria bottleneck assignment problem

A consequence of this Lemma is that the non-dominated solutions of MLBAP are a subset of the intersection points of a q -dimensional, rectangular grid generated by the values of the cost coefficients \mathbf{c}_{ij} ($i, j \in \{1, \dots, n\}$). Furthermore, the grid can be considered to be uniform, since the numerical values of the cost coefficients in each criterion can be replaced by the first l_k integers (let $l_k \leq I^2$ be the number of different cost coefficients in criterion k), which does not change the dominance structure of the problem.

Definition 3. Let $G = (V, E)$ be a complete bipartite graph with $V = (V_1, V_2)$, $|V_1| = |V_2| = n$ and with multicriteria edge costs $C = (c_{ij}^k)_{i,j \in \{1, \dots, n\}, k \in \{1, \dots, q\}}$. Furthermore, let $\epsilon \in \mathbb{R}^q$ be a vector. Then the *residual graph* $G[\epsilon]$ is defined as the bipartite graph on the same nodes, which edge cost are componentwise smaller or equal to ϵ :

$$G[\epsilon] := (V, E[\epsilon]) \quad \text{with } E[\epsilon] = \{(i, j) \in E : \mathbf{c}_{ij} \leq \epsilon\}$$

Lemma 4. Let two bipartite graphs $G_1 = (V, E)$ and $G_2 = (V, E \cup \{\bar{e}\})$ be given that differ only in one edge $\bar{e} \notin E$ and let M_1 be a maximum matching in G_1 . Then for the maximum matching M_2 of G_2 holds:

$$|M_1| \leq |M_2| \leq |M_1| + 1$$

If $|M_2| = |M_1| + 1$, the new edge \bar{e} is a matching edge in M_2 ($\bar{e} \in M_2$) and there is a matching augmenting path in G_2 extending M_1 to M_2 , which contains \bar{e} .

Proof. Follows directly from the correspondence of maximum matchings and maximum network flows (Burkard, Dell'Amico, and Martello 2009). A network flow increases only with the introduction of a new edge, if there is non-zero flow on this new edge, since otherwise this edge is redundant. So, we assume the addition of edge \bar{e} increases the matching. Since every matching extension can be described by means of the symmetric distance with an augmenting path, there must be one which contains \bar{e} , since it enters the matching. \square

With this lemma we can easily check if the introduction of one new edge leads to a matching extension. Since we know that the new edge is part of the augmenting

path, we do not have to search within the complete graph for an alternating path which augments the matching. Starting from \bar{e} we construct all alternating paths until we find an augmenting path. Thus, one can easily formulate an algorithm to extend a maximum matching gradually by adding one edge at a time to a residual graph.

We start from a lexicographic extreme value of the grid. Increasing the threshold vector stepwise in one component (while the other components are fixed) we can use the much faster matching extensions whenever only one edge is entering the residual graph. As soon as the residual graph contains a perfect matching, we obtained a weakly non-dominated point and can stop at this point and adapt the threshold value in the other components. The procedure to find an augmenting path in the case that only one edge enters the residual graph is given in Algorithm 1.

Algorithm 1 Procedure: extend_matching_by_one

procedure EXTEND_MATCHING_BY_ONE($G[\epsilon] = (V, E[\epsilon])$, new edge $\bar{e} = (u, v)$, maximal matching M w. r. t. $(V, E[\epsilon] \setminus \bar{e})$)

$S = \{v\}$

while $S \neq \emptyset$ **do** ▷ forward search

 choose one node i form set S

$S = S \setminus \{j\}$

if all nodes in $\mathcal{I} := \{i : (i, j) \in G[\epsilon] \setminus M\}$ are matched **then**

$S = S \cup \{k : (i, k) \in M \wedge (i, j) \in G[\epsilon] \setminus M\}$

$predL(i) = j \quad \forall j \in \mathcal{I}$

$predR(k) = j \quad \forall k \in \{(i, k) \in M \wedge (i, j) \in G[\epsilon]\}$

else ▷ augmenting path found

 let \bar{i} be an unmatched node

$predL(\bar{i}) = i$

$S = \emptyset$

end if

end while

$P1 = (\bar{j}, predL(\bar{j}), predR(predL(\bar{j})), \dots,)$

$S = \{u\}$

while $S \neq \emptyset$ **do** ▷ backward search

 choose one node j form set S

$S = S \setminus \{j\}$

if all nodes in $\mathcal{I} := \{i : (i, j) \in G[\epsilon]\}$ are matched **then**

$S = S \cup \{k : (j, k) \in M_P \wedge (j, i) \in G[\epsilon]\}$

$predL(j) = i \quad \forall j \in \mathcal{I}$

$predR(k) = j \quad \forall k \in \{(j, k) \in M_P \wedge (j, i) \in G[\epsilon]\}$

else ▷ augmenting path found

 let \bar{i} be an unmatched node

$predR(\bar{i}) = i$

$S = \emptyset$

end if

end while

$P2 = (j, predL(j), predR(predL(j)), \dots,)$

return $M_P \ominus (P1 \cup P2)$

end procedure

3 Summary and Conclusions

We presented in this article how the multicriteria linear bottleneck assignment problem can be solved using a multicriteria threshold algorithm and utilizing computational very fast extensions of a matching if only one edge is changed. As first numerical test show, these fast matching extensions are particularly useful for increasing problem sizes. Their applicability depends highly on the diversity of the cost coefficients, which is related to the fact that only then the stepwise increase in one objective functions leads to the introduction of only one new edge in the residual graph. If our method can be adapted and efficiently applied to the case when two or three edges enter the residual graph in one step, will be subject of further research.

References

- Burkard, Rainer E., Mauro Dell’Amico, and Silvano Martello. 2009. *Assignment Problems*. SIAM.
- Ehrgott, M., and X. Gandibleux. 2000. “A survey and annotated bibliography of multiobjective combinatorial optimization.” *OR Spektrum* 22 (4): 425–460.
- Ehrgott, M., and K. Klamroth. 1997. “Connectedness of efficient solutions in multiple criteria combinatorial optimization.” *European Journal of Operational Research* 97:159–166.
- Ehrgott, Matthias. 2005. *Multicriteria Optimization*. Springer.
- Fulkerson, D.R., I. Glicksberg, and O. Gross. 1953. “A production line assignment problem.” Technical Report RM-1102, The Rand Corporation, Santa Monica, CA.
- Garfinkel, Robert S. 1971. “An Improved Algorithm for the Bottleneck Assignment Problem.” *Operations Research* 19:1747–1751.
- Gorski, Jochen, Kathrin Klamroth, and Stefan Ruzika. 2011. “Connectedness of Efficient Solutions in Multiple Objective Combinatorial Optimization.” *Journal of Optimization Theory and Applications* 150:475–497.
- Malhotra, R., H.L. Bhatia, and M.C. Puri. 1982. “Bi-criteria assignment problem.” *Opsearch* 19:84–96.
- Przybylski, A., X. Gandibleux, and M. Ehrgott. 2008. “Two phase algorithms for the bi-objective assignment problem.” *European Journal of Operational Research* 185 (2): 509–533.

A Time-Indexed Model for the Elective Surgery Scheduling Problem

Kasper Tofte and Troels Martin Range
Centre of Health Economics Research,
Department of Business and Economics,
University of Southern Denmark
kto@sam.sdu.dk

Abstract

To solve the elective surgery scheduling problem we want to allocate hospital resources to surgical cases and assign these to time slots. The resources represent both employees and technical equipment such as operating rooms or beds.

We present a time-indexed model for the elective surgery scheduling problem, formulated as a multi-mode blocking job-shop problem. The model decomposes into one problem for each day in the planning horizon. These subproblems are combined in a master problem, which takes the form of a generalized set partitioning problem. Most of the resources are only in use during surgery; others (e.g. the beds) can be needed multiple days after the surgery has finished. The first type of resources motivates the day-based decomposition, whereas the latter type of resources complicates this decomposition. To handle this, we include a subset of the resources in the master problem.

Since the number of possible schedules for each day is huge, a column generation framework is proposed to solve the model. The pricing problems of this framework consist of a job-shop problem for each day. Each of these pricing problems assigns surgical cases to time slots on the specific day.

Key words: Elective surgery scheduling, Column generation, Job-shop problem.

1 Introduction

The problems of managing operating rooms and scheduling surgery have been widely studied over the last decades, and multiple reviews introduce the literature, for example those by Blake and Carter (1997), Cardoen, Demeulemeester and Beliën (2010) or May et al. (2011). The literature covers both short and long term planning, emergency and elective patient scheduling, utilization of different hospital equipment (resources) as well as different objectives.

We focus on scheduling elective patients for surgery with a planning horizon of about a week. This problem has been modelled as a job-shop problem, for example

by Pham and Klinkert (2008), where surgery is scheduled as jobs on machines represented by surgeons or operating rooms. The jobs may be using other resources, which may lead to a resource constraint job-shop problem. In this paper we present a time-indexed model for this job-shop problem and exploit the fact that the problem naturally decomposes into smaller job-shop problems, one for each day.

Many different objectives have been studied in the literature, from utilization of resources, through put, makespan to levelling of the workload (see the review from Cardoen, Demeulemeester and Beliën (2010) for a discussion on these objectives). Our objective imposes a penalty depending on the setting and time of each surgery. This enables us to model preferences among surgeons and patients, urgency or different priorities related to patient types.

The following papers propose column generation solutions to a similar problem. That is, they deal with elective surgery scheduling and a time horizon of one week. Fei et al. (2008) present a column generation approach, where each column represents a feasible schedule for one operating room on one day. They do not consider other resources than operating rooms, and their objective is to minimize unexploited time in the operating room as well as overtime.

Cardoen, Demeulemeester and Beliën (2009) consider a problem similar to the one presented in this paper. They present a column generation approach where each column in the master problem represents a feasible schedule for a surgeon. The objective of the model is composed of six components related to patient groups and peak number of beds used in the post-anaesthesia care unit (PACU).

Velásquez and Melo (2006) present a set partitioning model where the planning horizon is discretized into time intervals, and they then consider all possible resource combinations for each possible starting time for each operation. In this paper we further develop that idea and propose a decomposition into days.

2 Problem description

The planning horizon will be discretized into a set of time intervals. We will denote the index set of the time intervals by T . The planning horizon will consist of $|D|$ days and we will let D be the set of days within the planning horizon. The subset of time intervals that makes up day d will be denoted T_d . That is, T will be made up of the intervals T_d .

A pool of patients, P , who require surgery in the near future is given. The surgery can be conducted in different settings and can begin at different times. We will denote the set of settings available for patient $p \in P$ by S_p . Besides modelling different starting times, the different settings may also model the fact that different surgeons take different times to conduct the surgery or model that a patient can wake up in the operating room instead of at the PACU, and a lot of other things. This model is very flexible, but the size of the sets S_p may be huge.

Conducting surgery consumes resources such as surgeon time, operating-room time, equipment time and so on. The set of resources is denoted by R . The availability of resource $r \in R$ at time t is denoted by m_{rt} and states how large a quantity of the resource is available in the time period. We will distinguish between resources that span a single day and multiple days. That is, R can be partitioned such that $R = R_{SD} \cup R_{MD}$, where R_{MD} is the index set for resources spanning multiple days and R_{SD} is the index set for resources spanning a single day. The multi-day re-

sources could for example be bed usage in PACU after the surgery has finished or recovery sessions. The single-day resources typically represent resources used during the surgery.

z_{pstr} will denote the resource consumption at time $t \in T$ for resource $r \in R$ when conducting surgery on patient $p \in P$ according to setting $s \in S_p$.

With each patient we associate a cost, c_{ps} , for conducting the surgery for patient $p \in P$ according to the setting $s \in S_p$. The cost can model many different objectives. For example, it can model the urgency of a patient's condition, or it may model preferences among the surgeons or prioritize different patient groups at different times of the day. The penalty might also be used to model an outsourcing cost. This cost will then be the penalty of an 'empty' setting.

2.1 The model

With decision variables x_{ps} telling whether patient $p \in P$ is to have surgery according to setting $s \in S_p$ or not, we get the following generalized set partitioning model:

$$\min \sum_{p \in P} \sum_{s \in S_p} c_{ps} x_{ps} \quad (1)$$

$$\text{s.t.} \quad \sum_{s \in S_p} x_{ps} = 1, \quad \forall p \in P \quad (2)$$

$$\sum_{p \in P} \sum_{s \in S_p} z_{pstr} x_{ps} \leq m_{rt}, \quad \forall r \in R, t \in T \quad (3)$$

$$x_{ps} \in \{0, 1\}, \quad \forall p \in P, s \in S_p \quad (4)$$

The objective minimizes the cost of all patients. Constraints (2) ensure that all patients will be assigned a setting (possibly an 'empty' setting). Constraints (3) ensure that the resource capacities are never broken.

If we arrange rows and columns of the constraint matrix, we see the structure sketched in table 1. The first $|P|$ rows are general upper bound constraints arising from constraints (2). The following rows are those of constraints (3), but divided into multi-day and single-day resources. The \mathbf{Z} matrices, however, capture the z_{pstr} from the original problem. In the general problem the \mathbf{Z} matrices may not show any structure, but we will assume that all the entries z_{pstr} are non-negative.

The table clearly shows a block angular structure arising from the single-day resources. This structure motivates the following rewriting of the model, where each column represents a surgery schedule for each day.

By K_d we denote the set of possible schedules for day $d \in D$. The modified cost coefficient c_{dk} is the aggregation of cost coefficients for the patients operated according to schedule $k \in K_d$ if performed on day $d \in D$. We also introduce the outsourcing cost of patient $p \in P$ directly, denoting it o_p .

The constraint matrix is made up of the coefficients a_{pdk} and b_{dktr} . a_{pdk} is 1 if patient $p \in P$ is part of schedule $k \in K_d$ on day $d \in D$ and 0 otherwise. b_{dktr} is the amount of resource $r \in R_{MD}$ that is in use at time $t \in T$ if schedule $k \in K_d$ is performed on day $d \in D$.

	Day 1			Day 2			...
	1	...	1	1	...	1	
P	$\mathbf{0}$	\ddots	$\mathbf{0}$	$\mathbf{0}$	\ddots	$\mathbf{0}$...
			1 ... 1			1 ... 1	
$R_{MD} \times T$		\mathbf{Z}_1			\mathbf{Z}_2		...
$R_{SD} \times T_1$		\mathbf{Z}_{T_1}		$\mathbf{0}$			$\mathbf{0}$
$R_{SD} \times T_2$		$\mathbf{0}$		\mathbf{Z}_{T_2}			$\mathbf{0}$
$R_{SD} \times T_3$		$\mathbf{0}$		$\mathbf{0}$			\ddots

Table 1: The constraint matrix of problem (1)–(4) clearly showing a block angular structure. The first rows are the general upper bound constraints (2), followed by submatrices formed by first the multi-day resources and then the single-day resources.

$$\min \sum_{d \in D} \sum_{k \in K_d} c_{dk} y_{dk} + \sum_{p \in P} o_p s_p \quad (5)$$

$$\text{s.t.} \quad \sum_{k \in K_d} y_{dk} = 1, \quad \forall d \in D \quad (6)$$

$$\sum_{d \in D} \sum_{k \in K_d} a_{pdk} y_{dk} + s_p = 1, \quad \forall p \in P \quad (7)$$

$$\sum_{d \in D} \sum_{k \in K_d} b_{dktr} y_{dk} \leq m_{rt}, \quad \forall r \in R_{MD}, t \in T \quad (8)$$

$$s_p \geq 0, \quad \forall p \in P \quad (9)$$

$$y_{dk} \in \{0, 1\}, \quad \forall d \in D, k \in K_d \quad (10)$$

2.2 The pricing problems

Since the sets K_d will be huge we will rely on column generation techniques. We will introduce a pricing problem for each day. That is, given a day $d \in D$ the pricing problem becomes

$$\min_{k \in K_d} c_{dk} - \sum_{p \in P} a_{pdk} \mu_p - \sum_{r \in R_{MD}} \sum_{t \in T} b_{dktr} \tau_{rt} - \pi_d \quad (11)$$

where $\mu_p \in \mathbb{R}$ are the duals associated with (7), $\tau_{rt} \leq 0$ are the duals associated with (8) and $\pi_d \in \mathbb{R}$ are the duals associated with (6).

The problem is to find a minimum cost schedule for operations to perform on day $d \in D$. The problem can be formulated with almost the same model we presented initially. The resource constraints are slightly remodelled and the objective needs to take the duals into account.

In the objective \hat{c}_{ps} is an adjusted cost coefficient such that the dual μ_p is incorporated; that is

$$\hat{c}_{ps} = c_{ps} - \mu_p$$

We keep the decision variables x_{ps} but restrict the settings to those on day $d \in D$ with the notation that S_{pd} is the set of legal settings for patient $p \in P$ on day $d \in D$. The model also introduces the variable b_{rt} , which is the amount of resource $r \in R_{MD}$ this schedule consumes in time period $t \in T$. The rest of the notation is as introduced earlier.

$$\min \sum_{p \in P} \sum_{s \in S_{pd}} \hat{c}_{ps} x_{ps} - \sum_{r \in R_{MD}} \sum_{t \in T} b_{rt} \tau_{rt} \quad (12)$$

$$\text{s.t. } \sum_{s \in S_{pd}} x_{ps} \leq 1, \quad \forall p \in P \quad (13)$$

$$\sum_{p \in P} \sum_{s \in S_{pd}} z_{pstr} x_{ps} \leq m_{rt}, \quad \forall r \in R, t \in T_d \quad (14)$$

$$\sum_{p \in P} \sum_{s \in S_{pd}} z_{pstr} x_{ps} - b_{rt} = 0, \quad \forall r \in R_{MD}, t \in T \quad (15)$$

$$0 \leq b_{rt} \leq m_{rt}, \quad \forall r \in R_{MD}, t \in T \quad (16)$$

$$x_{ps} \in \{0, 1\}, \quad \forall p \in P, s \in S_p \quad (17)$$

With the adjusted cost, the objective (12) is equivalent to (11) except for the dual π_d , but it is constant when the day is fixed. Constraints (13) ensure that no patient has surgery more than once. Constraints (14) and (15) ensure that the resource limit is observed and stores the resource consumption of the multi-day resources in b_{rt} . Note that (14) only deals with the resources used on the given day, where (15) deals with the whole time horizon.

3 Solution methods

An a priori enumeration of the columns for problem (1)–(4) might be possible, and so we will try to solve that problem by a general purpose MIP solver. If the general purpose MIP solver fails to solve the problem, we can apply constraint branching, which was introduced in Ryan and Foster (1981) and applied to the elective surgery scheduling problem in Velásquez and Melo (2006). In our case we will branch on constraints (2) and (3); that is, whether patient $p \in P$ is using resource $r \in R$ at time $t \in T$ or not.

Even though we might be able to solve the problem directly, we do believe that we can benefit from the decomposition described in section 2. This solution approach is an attempt to solve a smaller and relaxed version of problem (5)–(10) where we have restricted the problem to only a subset of each K_d and relaxed the integrality constraint (10). We will then rely on the pricing problem described in section 2.2 to generate new columns when needed.

In order to impose integrality on the relaxed solution we will again apply constraint branching. Two branching strategies have been considered. First, one can branch on constraints (6) and (7). That is, in one branch we will enforce that patient $p \in P$ is to have surgery on day $d \in D$ and in the other branch that patient p is to

have surgery on another day than d . This branching rule can easily be enforced in the pricing problem by adjusting constraints (13). An alternative branching strategy could be to branch on the amount of a specific resource at a specific time that a schedule from one day is allowed to use. That corresponds to putting bounds on the b_{dktr} parameter of constraints (8). In the pricing problem, this will be handled by adjusting the bounds on b_{rt} in constraints (16).

One problem that remains is how to enumerate the columns of the pricing problem as well as the number of columns. The latter point can be addressed by two observations. First, if the bounds on b_{rt} imposed by the branching suggested above are tight enough, we might be able to eliminate the columns that have too high z_{pstr} value. Secondly, since the z_{pstr} values are all non-negative, only columns having $\hat{c}_{ps} < 0$ will contribute to a negative reduced cost column. That is, we can eliminate all columns having $\hat{c}_{ps} \geq 0$ from the pricing problem.

4 Future work

In this paper we have presented a flexible model for the elective surgery scheduling problem. In our future work we will, based on cooperation with a day surgery center in Denmark, develop the model in greater detail and suggest methods for efficiently enumerating columns both for model (1)–(4) and for the pricing problems.

Our future work will also include an implementation of the presented model. It will be tested on available datasets as well as on the problem arising at the day surgery center in Denmark.

Acknowledgements

The research presented in this paper has been conducted while the first author was a visiting scholar at the Department of Engineering Science at the University of Auckland, and some of the ideas have been developed through fruitful discussions with Andrew Mason.

References

- Blake, J.T., and M.W. Carter. 1997. “Surgical process scheduling: a structured review.” *Journal of the Society for Health Systems* 5 (3): 17–30.
- Cardoen, Brecht, Erik Demeulemeester, and Jeroen Beliën. 2009. “Sequencing surgical cases in a day-care environment: An exact branch-and-price approach.” *Computers & Operations Research* 36 (9): 2660–2669.
- Cardoen, Brecht, Erik Demeulemeester, and Jeroen Beliën. 2010. “Operating room planning and scheduling: A literature review.” *European Journal of Operational Research* 201 (3): 921 – 932.
- Fei, H., C. Chu, N. Meskens, and A. Artiba. 2008. “Solving surgical cases assignment problem by a branch-and-price approach.” *International Journal of Production Economics* 112 (1): 96–108. Special Section on Recent Developments in the Design, Control, Planning and Scheduling of Productive Systems.

- May, Jerrold H., William E. Spangler, David P. Strum, and Luis G. Vargas. 2011. "The Surgical Scheduling Problem: Current Research and Future Opportunities." *Production and Operations Management* 20 (3): 392–405.
- Pham, Dinh-Nguyen, and Andreas Klinkert. 2008. "Surgical case scheduling as a generalized job shop scheduling problem." *European Journal of Operational Research* 185 (3): 1011–1025.
- Ryan, D.M., and B.A. Foster. 1981. "An integer programming approach to scheduling." In *Computer scheduling of public transport urban passenger vehicle and crew scheduling*, edited by A. Wren, 269–280. North Holland, Amsterdam, The Netherlands.
- Velásquez, R., and M.T. Melo. 2006. "A Set Packing Approach for Scheduling Elective Surgical Procedures." In *Operations Research Proceedings 2005*, edited by Hans-Dietrich Haasis, Herbert Kopfer, and Jörn Schönberger, Volume 2005 of *Operations Research Proceedings*, 425–430. Springer Berlin Heidelberg.

Mediated Modelling to Support Spatial Planning: Population Change, Inequality and City Attractiveness in Wellington

Marjan van Den Belt
Ecological Economics Research New Zealand
Massey University
New Zealand
M.vandenBelt@massey.ac.nz

Abstract

A Mediated Modelling (MM) approach will be presented, which was used in a Wellington urban sustainability context to explore the integration between the four aspects of well-being: social, economic, ecology and culture. MM refers to “model building with rather than for stakeholders”. This project is part of an action research programme (Sustainable Pathways 2), spanning six years. Population growth, or in Wellington the concern about population decline was the starting point. Inequality and overall attractiveness of the city were the main indicators. The MM approach emphasized integration based on the perception of 15 high-level stakeholders representing different interests from local and regional government as well as the education and health sector. None of the participants were model builders. All indicated that a more integrated approach would be desirable before the workshops. The results were an evaluation based on interviews. Storytelling, based on a causal loop diagram made up of model sectors, surfaced as a useful tool that could enhance communication with other non-participants. The model was used to simulate two scenarios: a) What if population changes by 10% (up or down) by 2020 then what happens to “inequality”? and b) What if the relative income per person in the Wellington region changes by 10% (up or down)? How does this impact on the relative attractiveness of the Wellington region? Sixty percent of the participants who were interviewed before the workshops were also interviewed afterwards. All but two expressed an interest in on-going participation in subsequent MM workshops, following the recommendation that a Dynamic Genuine Process Indicator be the topic. This article presents the progression of the MM workshops in Wellington and highlights the role of causal loop diagramming, model building, storytelling and simulation; what worked well and what didn't.

Modeling Repairs of Systems with a Bathtub-Shaped Failure Rate Function

Sima Varnosafaderani

School of Mathematics, Statistics and Operations Research
Victoria University of Wellington
New Zealand
sima@msor.vuw.ac.nz

Stefanka Chukova

School of Mathematics, Statistics and Operations Research
Victoria University of Wellington
New Zealand
stefanka@msor.vuw.ac.nz

Abstract

Most of the reliability literature on modeling the effect of repairs on systems assumes the failure rate functions are monotonically increasing. For systems with non-monotonic failure rate functions, most models deal with minimal repairs (which do not affect the working condition of the system) or replacements (which return the working condition to that of a new and identical system). We explore a new approach to model repairs of a system with a non-monotonic failure rate function; in particular, we consider systems with a bathtub-shaped failure rate function. We propose a repair model specified in terms of modifications to the virtual age function of the system, while preserving the usual definitions of the types of repair (minimal, imperfect and perfect repairs) and distinguishing between perfect repair and replacement. In addition, we provide a numerical illustration of the proposed repair model.

1 Introduction

Most engineered systems – defined as an arrangement of components that together perform an identified (and predefined) set of functions – are susceptible to failures, and require some form of rectification in order to return to a functioning condition. Most rectifications have an effect on the probability and number of future failures of the system over a given period of time.

The sequence of numbers of failures of the system in time, i.e. the failure process, is modeled as a stochastic counting process, assuming that there can be at

most one failure in an infinitesimally small interval of time. When the rectification action following each failure is immediate and instantaneous, this process can also be described as the sequence of times to failure (or consecutive system's lifetimes).

A counting process is completely described by its conditional intensity function, and therefore, rectifications are usually defined in terms of their effect on the conditional intensity function of the failure process. The initial conditional intensity function is the failure rate function of the original lifetime (time to first failure of the system), which is often a continuous function of time, and is classified as constant, monotonic increasing or decreasing, or a combination of these. Beyond the first failure, the conditional intensity function is altered in accordance with the rectifications performed following the first and all consequent failures.

Not all rectifications have the same effect on the system, and based on their effect, they are categorized as either replacements or repairs with varying degrees of effectiveness. In some cases, replacements can be viewed as extreme repairs.

In this article, we suggest an approach to model the effect of rectifications (here, repairs) for a system having a non-monotonic (here, bathtub-shaped) failure rate function. We define repairs in terms of their effect on the virtual age of the system.

The article is arranged as follows. In Section 2 we discuss the concepts mentioned above in more detail, and provide a brief review of existing models relevant to our study. In Section 3, we describe the repair model and provide model formulation. In Section 4, we provide a numerical illustration of the proposed model. Finally, in Section 5, we conclude with a discussion of the proposed model and some directions for future research.

2 Background and Definitions

Let $\lambda_c(t)$ denote the conditional intensity function of the failure process denoted by $\{N(t); t \geq 0\}$. Then

$$\lambda_c(t) = \lim_{dt \rightarrow 0} \frac{P\{N(t+dt) - N(t) = 1 \mid \mathcal{F}_t\}}{dt},$$

where $N(t+dt) - N(t)$ is the number of failures in the interval $(t, t+dt]$, and $\mathcal{F}_t = \{N(s); 0 \leq s < t\}$ is the history of the process before time t . The initial conditional intensity (or baseline intensity), denoted by $\lambda_0(t)$, is the failure rate of the time to first failure, which is

$$\lambda_0(t) = r(t) = \lim_{dt \rightarrow 0} \frac{P\{N(t+dt) - N(t) = 1 \mid N(t) = 0\}}{dt}.$$

A failure rate or the corresponding distribution is categorized as: constant failure rate (CFR) when $r(t)$ is constant over t ; increasing failure rate (IFR) when $r(t)$ is increasing in t ; decreasing failure rate (DFR) when $r(t)$ is decreasing in t ; or some combination of these. For instance, the bathtub-shaped failure rate (BFR) function which is initially decreasing, then constant and finally increasing:

$$r(t) = \begin{cases} r_1(t) : r'_1(t) < 0, & t \leq a_1 \\ r_2(t) : r'_2(t) = 0, & a_1 < t \leq a_2 \\ r_3(t) : r'_3(t) > 0, & t > a_2, \end{cases} \quad (1)$$

where a_1 and a_2 are the change points (points at which the the derivative of the failure rate function changes sign) of the BFR function. The BFR function is a generalization of the above categories; setting $a_1 = a_2 = 0$ ($a_1 = a_2 = \infty$ or $a_1 = 0$ and $a_2 = \infty$), it becomes an increasing (decreasing or constant) function. Setting $a_1 = a_2 = a$, we get a U-shaped failure rate (UFR) function (Lai and Xie 2006); see Figure 1.

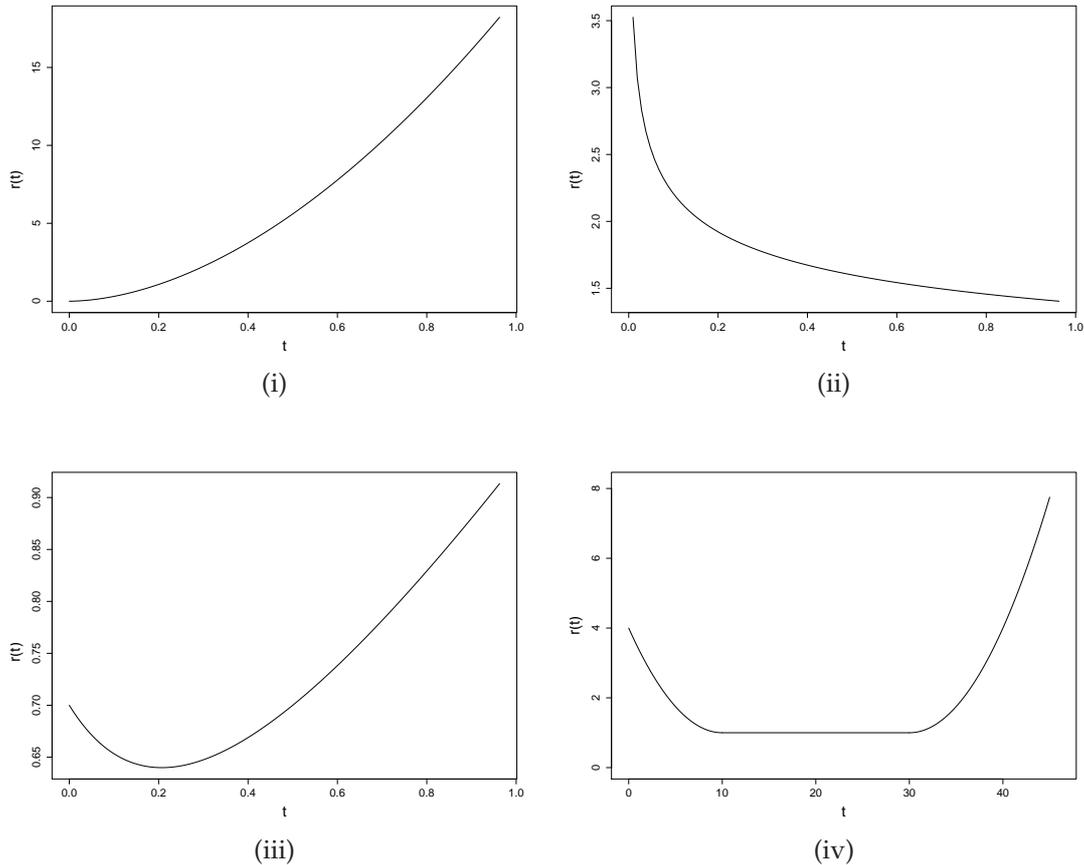


Figure 1: Failure rate functions: (i) IFR; (ii) DFR; (iii) UFR; (iv) BFR.

The quantity $r(t) dt$ is the approximate probability that the system will fail for the first time in $(t, t + dt]$. The distribution of the time to first failure, denoted by T_1 , is defined by the failure rate $r(t)$, but subsequent failure times are affected by the type of rectification performed after a failure.

Rectifications are broadly classified into repair and replacement. With replacements, the system is replaced by a new, identical system upon failure. The condition of the system following a replacement is therefore identical to a new system. In this case, the failure process is modeled as a renewal process with conditional intensity function $\lambda_c(t) = r(t - T_{N(t)})$ (where $T_{N(t)}$ is the time of the last replacement (perfect repair)), and the expected number of replacements is given by the renewal function (Hokstad 1997; Hunter 1974; Aven and Jensen 1998). Replacements apply when the system is beyond repair (or non-repairable) or when replacing the system is more feasible than repairing it.

Repairs are characterized by their effectiveness, which is often referred to as the *degree of repair* – the degree to which the functioning condition of the system

is restored following the repair. Based on this degree, repairs are typically categorized as minimal, imperfect or perfect repair.

Minimal repairs have no effect on the failure rate function; i.e., the system immediately before and after the failure is in the same functioning condition. Therefore, when all repairs are minimal, the failure process is a non-stationary Poisson process with conditional intensity function $\lambda_c(t) = r(t)$. The expected number of failures over an interval $[0, t)$ is then given by $E[N(t)] = \int_0^t \lambda_c(s) ds$.

Perfect repairs, for IFR functions, leave the repaired system in an as-good-as-new working condition, which implies that a perfect repair is equivalent to a replacement and can be modeled as one. This definition works for systems having an IFR function, since the system at the start of its lifetime has the lowest failure rate, i.e., it is at its best. However, if a system has a failure rate function that is initially decreasing (e.g. BFR), this definition does not hold, because a repair that leaves the system in an as-good-as-new condition is actually worsening the system, since the failure rate of the system at the start of its lifetime is higher than its failure rate when it is working at its best. Therefore, we distinguish between replacement and perfect repair, and describe a perfect repair as the best form of repair (not taking into account improvements/upgrades). In other words, a perfect repair is one that restores the functioning condition of the system to its condition when it is performing at its best. This point of ideal performance is at the start of the system's lifetime for a system with an IFR function, but not for a system having an initially decreasing failure rate function.

Imperfect repair, sometimes referred to as general repair, is any repair that leaves the system in a functioning condition that is between the functioning conditions following minimal and perfect repairs. For systems with a IFR function, the definition of an imperfect repair has included the extremes minimal repair and replacement (aka perfect repair). Here, imperfect repair includes as its extremes minimal and perfect repairs, but not replacements. Therefore, we distinguish between repair and replacement.

In most settings, the degree of repair is a variable with range $[0, 1]$, where a degree of zero corresponds to a minimal repair, a degree of one corresponds to a *perfect* repair, and a degree between these extremes corresponds to an imperfect repair. Therefore, the higher the degree of repair, the bigger the improvement in the functioning condition of the system.

Here, we do not consider repairs that can worsen the system or upgrades (or improvements).

Many repair models have been suggested for systems having IFR functions. Some common models are the virtual age models discussed in Kijima (1989), Varnosafaderani and Chukova (2012a) and Doyen and Gaudoin (2004); and the intensity reduction models discussed in Lindqvist (1998), Varnosafaderani and Chukova (2012b) and Doyen and Gaudoin (2004). Models of repair in the case of BFR functions assume that rectifications are either minimal or replacements. The virtual age models for IFR functions have been applied to BFR functions; see (Dijoux 2009), but due to the failure rate being initially decreasing, repairs of degree greater than zero actually worsen the product.

In this article, we propose a new approach to modeling imperfect repairs for systems having BFR functions which better suit the definitions of the types of

repair. The effects of repairs are described in terms of modifications to the virtual age function of the system.

3 Modeling the Effect of Repairs

Let T_i denote the time of the i th failure (also repair, since repairs are immediate and instantaneous), and let δ_i denote the degree of the i th repair. Also, let $A(t)$ denote the virtual age of the system at time t .

Based on the virtual age function of the system, we propose the following repair model. The virtual age of the system at time t , is given by

$$A(t) = \begin{cases} t + \sum_{i=1}^{N(t^-)} \delta_i [a_1 - A(T_i)] , & t \leq a_1 \\ t - \sum_{i=N(a_1^+)}^{N(t^-)} \delta_i [A(T_i) - a_1] , & t > a_1 \end{cases} \quad (2)$$

where $A(T_i)$ is the virtual age of the system at the time of its i th failure. Before the first failure, when $N(t^-) = 0$, the virtual age is simply $A(t) = t$; and immediately after a_1 and before the first failure in the useful life period, the virtual age of the system is again $A(t) = t$. See Figure 2 for an illustration of the virtual age function for three failures occurring at times t_1, t_2 and t_3 .

With a failure rate function that is initially decreasing, the point of best performance is not the start of the system's lifetime, but the point at which the failure rate function is at its lowest. This point for a BFR function is the first change point a_1 .

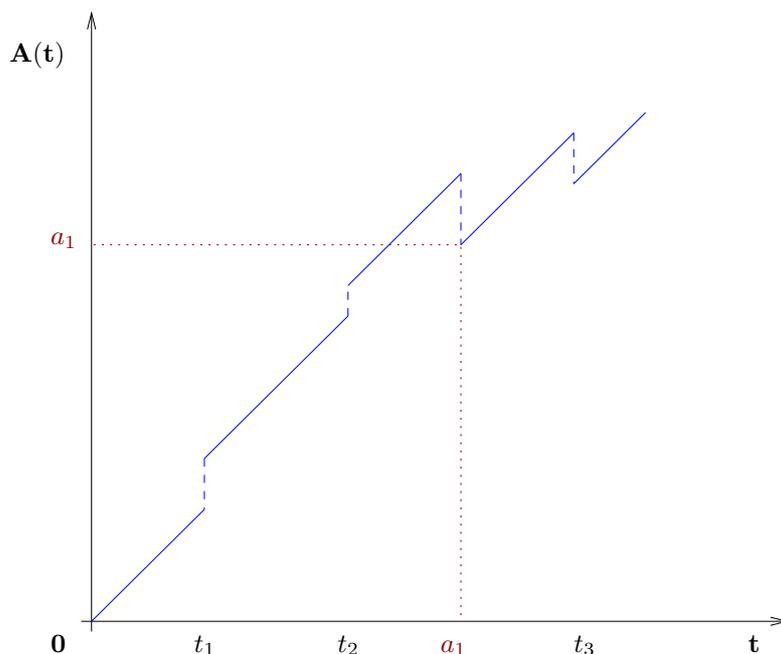


Figure 2: Virtual age function following imperfect repairs of varying degree.

Therefore, according to this model, when the virtual age of the system at the time of the i th failure is less than the first change point a_1 , the effect of a repair is modeled as an increase in the virtual age of the system, such that, a perfect repair

results in the virtual age being a_1 . At a_1 , the virtual age of the system is set to its calendar age, i.e. $A(a_1) = a_1$. This extends the useful life period of the system, which will decrease the probability of future failures. When the age of the system is greater than the first change point a_1 , then the effect of a repair is a decrease in the virtual age of the system, such that, a perfect repair results in the virtual age being a_1 . The point a_1 is the point of ideal performance, because it is the start of the useful life of the system, and the failure rate of the system at this point is at its lowest.

The conditional intensity function of the failure process is given by

$$\lambda_c(t) = \begin{cases} \lambda_0(t) , & t \leq T_1 \\ \lambda_0(A(t)) , & t > T_1 \end{cases}$$

where $\lambda_0(\cdot)$ is the baseline intensity (or failure rate) function. See Figure 3 for an illustration of this function following repairs of varying degree.

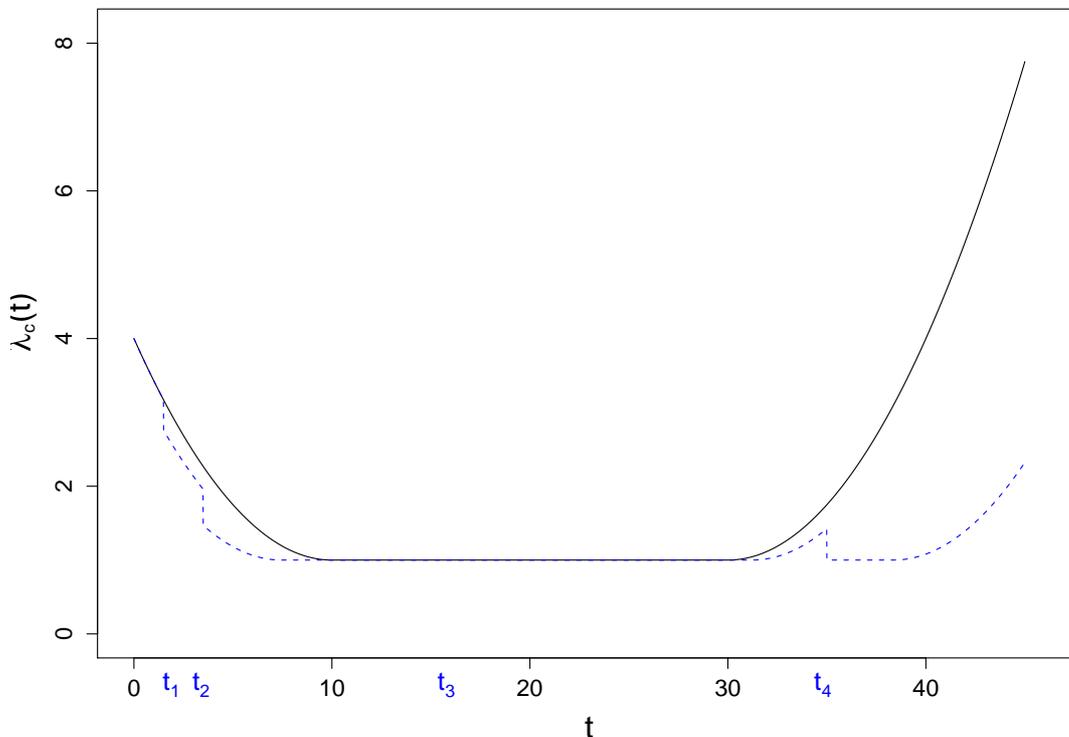


Figure 3: Conditional intensity function following: four imperfect repairs of varying degree (dashed line); minimal repairs (solid line).

The repair model stays true to the definitions of the types of repair. A perfect repair is the best form of repair, and should result in the system performing at its best, which is in this case at a_1 . A minimal repair, should by definition leave the system in the same condition that it was prior to failure, and here, the virtual age does not change following a minimal repair. The effect of an imperfect repair should be between those of the minimal and perfect repairs, and effectiveness of the repair should increase with its degree. Here, as the degree of repair increases, so does the effectiveness of the repair (which is reflected in the decrease in the conditional intensity function of the process).

The assumption for this model is that the useful life period $(a_1, a_2]$ of the system is at least as long as the DFR period $(0, a_1]$, i.e. $a_2 - a_1 \geq a_1$.

4 Numerical Illustration

In this section, we provide a simple example that illustrates the proposed repair model.

The baseline intensity function used in this example is

$$\lambda_0(t) = \begin{cases} \lambda + \alpha_1 (a_1 - t)^{\beta_1} , & t \leq a_1 \\ \lambda , & a_1 < t \leq a_2 \\ \lambda + \alpha_2 (t - a_2)^{\beta_2} , & t > a_2 , \end{cases} \quad (3)$$

where $\lambda > 0$, $\beta_1, \beta_2 > 0$, $\beta_1 \geq \beta_2$, and $\alpha_1, \alpha_2 > 0$. The parameter values are chosen to be $\lambda = 1$, $\alpha_1 = 0.6$, $\alpha_2 = 0.5$, $\beta_1 = 2.5$, and $\beta_2 = 2.8$, and the change points are chosen to be $a_1 = 4$ and $a_2 = 8$.

Since virtual age models for IFR functions have been frequently examined and the effect of repairs in this case is known, we limit our illustration to exploring the effect of repairs based on our virtual age model in the DFR phase. To do so, we select an arbitrary mission time τ , and applying repairs of varying degree in the interval $[0, a_1)$, we compute the expected number of failures in $(0, \tau]$. Here, the mission time is chosen to be $\tau = 10$.

The repairs performed are chosen according to the following strategy: the first repair in the interval $(0, a_1]$ is imperfect, and all other repairs are minimal.

Let T_1 denote the time of the first failure. The density function of T_1 in terms of the baseline intensity function is given by

$$f_1(t) = \lambda_0(t) e^{-\int_0^t \lambda_0(s) ds} . \quad (4)$$

The expected number of failures in the interval $[0, \tau)$, is then derived as follows:

$$E[N(\tau)] = \int_0^{a_1} \left[1 + \int_{t_1}^{a_1} \lambda_0(s + \delta_1(a_1 - t_1)) ds \right] f_1(t_1) dt_1 + \int_{a_1}^{\tau} \lambda_0(s) ds ,$$

where δ_1 is the degree of the imperfect repair performed in $(0, a_1]$.

Tabulated in Table 1 are the expected numbers of failures $E[N(10)]$ for degrees of repair $\delta_1 \in \{0.1, 0.2, \dots, 1.0\}$.

Table 1: Expected number of failures in the interval $[0, \tau)$ for various degrees of repair

δ_1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$E[N(10)]$	33.78	27.3	22.4	18.81	16.29	14.64	13.63	13.09	12.86	12.79	12.78

Note that, according to the repair model, as the degree of repair increases, the expected number of failures decreases; also see Figure 4.

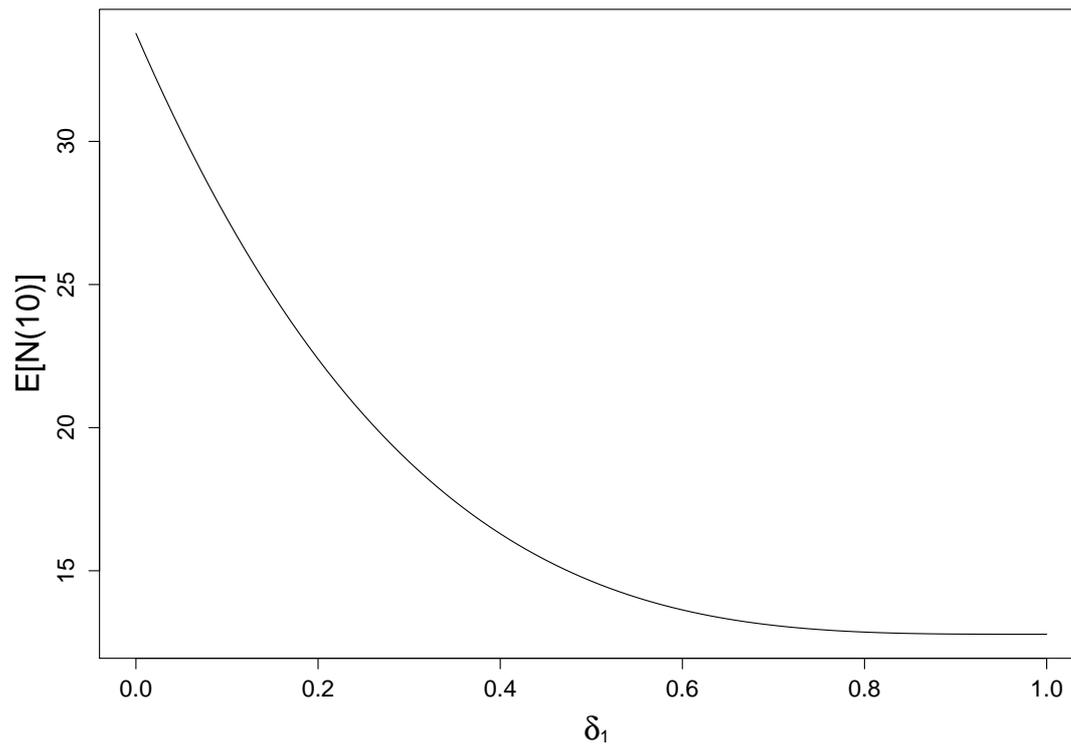


Figure 4: Expected number of failures $E[N(10)]$ for $\delta_1 \in [0, 1]$.

5 Conclusion

In this article, we proposed a new repair model for systems having a BFR function. The effect of repairs was modeled as a modification in the virtual age of the system following the repairs.

According to the proposed model (illustrated in Section 4), as the degree of any given repair increases (while others remain fixed), the expected number of failures decreases, since the reliability of the system is improved.

Some possible future research directions are deriving virtual age models for systems with more than two change points and extension of these models to two dimensions.

References

- Aven, Terje, and Uwe Jensen. 1998. *Stochastic Models in Reliability*. Springer-Verlag.
- Dijoux, Yann. 2009. "A virtual age model based on a bathtub shaped initial intensity." *Reliability Engineering & System Safety* 94 (5): 982–989.
- Doyen, Laurent, and Olivier Gaudoin. 2004. "Classes of imperfect repair models based on reduction of failure intensity or virtual age." *Reliability Engineering & System Safety* 84 (1): 45–56.
- Hokstad, Per. 1997. "The failure intensity process and the formulation of reli-

- ability and maintenance models." *Reliability Engineering & System Safety* 58 (1): 69–82.
- Hunter, Jeffery. 1974. "Renewal Theory in Two Dimensions: Asymptotic Results." *Advances in Applied Probability* 6 (3): 546–562.
- Kijima, Maasaki. 1989. "Some Results for Repairable Systems with General Repair." *Journal of Applied Probability* 26 (1): 89–102.
- Lai, C D, and M Xie. 2006. *Stochastic Ageing and Dependence for Reliability*. Springer-Verlag.
- Lindqvist, Bo Henry. 1998. "Statistical Modeling and Analysis of Repairable Systems." *Limnios (Eds.); Statistical and probabilistic models in reliability*, Birkhauser. 3–25.
- Varnosafaderani, Sima, and Stefanka Chukova. 2012a. "An imperfect repair strategy for two-dimensional warranty." *Journal of the Operational Research Society* 63 (6): 846–859.
- . 2012b. "A two-dimensional warranty servicing strategy based on reduction in product failure intensity." *Computers and Mathematics with Applications* 63 (1): 201–213.

Intervention as Language Games

Jorge Velez-Castiblanco
Management School
Universidad EAFIT
Colombia
velez.castiblanco@gmail.com

Abstract

Intervening to improve the conditions of a situation can make use of many approaches. It can use systems methods and methodologies, models and techniques. They can be from different paradigms, and they can be non systems thinking too. What is more, for intervening, devices such as jokes, anecdotes, or comforting pats on the back can also be drawn. From systems thinking to resources more akin to ethnographical accounts, all of them can be use. This paper proposes the concept of language games in Wittgenstein philosophy as a perspective able to encompass all this diversity. The dynamics of the situation in itself is a language game and everything that we use to affect the situation can be thought of as language games. The overall view of an intervention process can be seen as the “overlapping of many fibres.” It is argued that this perspective allows for a flexible way to adjust and combine distinct approaches without forgetting actor’s central role, and its influence on the process. Considering the intervention as language games, aims to understand the possibilities and effects that the uses of the different tools have on the “activities” and “forms of life” of the engaged actors.

Key words: Multimethodology, Intervention, Wittgenstein, Philosophy of Language, Language Games, Critical Systems Thinking.

1. Intervention Tools: An Extensive “Family”

Systems Thinking and Operational Research, provide a big range of tools aimed to guide / conduct intervention in organizational contexts. Following Jackson (2000) it is possible to see tools underpinned by Funtionalist, Interpretive, Critical and Postmodern paradigms. These paradigms correspond to strategies for representing and optimizing a situation, achieving inter-subjective understanding, achieving emancipation and uncovering conflict and marginalization.

Although dissimilar, all of them represent what Keys (1997) calls the theory driven approach. They encode knowledge in methods, methodologies, models and techniques. This helps to act, reflect and learn about interventions. However, Keys (2002, p.212) points out that “a main disadvantage [of the approach] is that they simplify the work, often to an unrealistic degree, and do not formally acknowledge its social aspects”.

Overcoming the limitation of the social, leads to what Keys (1997) identify as the practice driven approach. This derives from social science, and it focuses on concrete

cases of intervention. It “examines in detail particular pieces of MS/OR work and seeks to develop an understanding of how experienced practitioners carry out their work” (Keys, 2002, p.212). This approach pays attention to the social and political as well to the creative aspects of intervening. This approach is one that can give importance to pats on the back.

Merging together both approaches enhance greatly the range of possible actions in an intervention setting. Actions can be previously encoded as in the theory driven, but they can also be emergent from the socio political context. This presents us with what Keys identifies as an emergent third approach, one that looks to work across the other two circuits of analysis, one that enables to work with a very extensive family of tools.

This paper contributes to this third approach with the proposition that philosophy of language, especially Wittgenstein’s work on Language Games is a suitable foundation that let us consider the whole of the situation, and the whole of strategies and tools used for intervene.

2. Wittgenstein’s Two Ways to Approach Language

Wittgenstein is one of the most important philosophers of the twentieth century. His philosophy centres on the problem of language, and on that problem, he articulated two powerful and influential views. First, he constructed the idea of language as “the mirror of the world” (Wittgenstein, 1922), arguing that the structure of sentences and ideas was showing the logical structure of the facts in the world. For instance, these ideas were use by the Vienna circle to support their project of achieving the unity of science (Nodoushani, 1999). This implied expressing the whole of knowledge in a single logical standard language.

However, if it is considered that a methodology must mirror something about the world, or that it must agree with a single logical standard language, then having the effect of an actor in their use will cause a deviation from the “real” image of the world and the logic of a standard language. Under this rigid view, the task of the actor is simply to assess the ‘facts’ of the situation and then select the appropriate methodology. (see, for instance, Jackson and Keys (1984)).

According to Garfinkel (1981), conceptual frameworks guide the kind of questions and explorations that we can make. Consequently, in order to take into account actors, it requires a conceptual framework capable of embracing multiple factors and situations in interventions. It requires abandoning the idea that there is an intrinsic or “true” nature in the tools that have to be expressed in particular ways. It requires a philosophical position where the actor can have a place in understanding such use.

Wittgenstein’s (1958) later view on language meets this purpose. Here the ‘reality’ is not out there, so there is no need to mirror it. The metaphor is now the one of the tool. Language allows us to do things in the world. Knowledge is created by social interactions or, in Wittgenstein’s terms, ‘language games’. Knowledge is dependent on of the actors involved. Here it is possible to find a place for actors.

This view is presented in ‘Philosophical Investigations’ a book that some consider the most important book in 20th century philosophy (Stern, 2004). The views in this book are associated with what is now called social constructionism, an influential view not

only in organization studies but also in broader humanities and social sciences (Gergen & Leach, 2001; Gergen & Gergen, 2003; Schwandt, 2000).

The influence of Wittgenstein's view on MS/ST is indirect and not so visible. Authors such as White and Taket (1997) and Jackson (2000) allude to him in support of their ideas but without making his ideas central to their claims. One possible exception is Hassard's (1990) proposal of mediating incommensurable paradigms using a meta-language game. However, it seems that this proposal has not been developed or further commented on the literature. Yet, there are ideas on second order cybernetics where the concept of socially constructed knowledge in language finds resonances (see, for example, Varela (1979), Maturana (1988), Von Foerster (1989), and Von Glasersfeld (1996)). Using Wittgenstein's ideas to inform MS/ST is one of the contributions that this paper looks to make to the field.

3. Language Games

Through a series of thought experiments, Wittgenstein (1958, §43) not only shows that meanings are affected by use, but provides the crucial idea that "the meaning of a word is its use in the language". This implies that meaning does not stem from intrinsic characteristics in the word. Meaning arises from what it is possible to do with such a word. In consequence, words are being seen as tools that can be applied to affect interactions in the context. For example, following Winograd and Flores (1986), a speech act of declaration, can pronounce a couple married, or declare somebody as CEO in a company.

This idea of meaning as use derives from perhaps the most famous concept in Wittgenstein philosophy: Language Games. Accordingly, to Wittgenstein (1958, §7), language can be understood in terms of language games. These are defined as "the whole, consisting of language and the actions into which it is woven" also "the term 'language-game' is meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life" (Wittgenstein, 1958, §23).

However, it is difficult to apply the concept of Language Games, mainly because Wittgenstein explicitly avoids elaborating the concept. In his own words:

"For someone might object against me: "You take the easy way out! You talk about all sorts of language-games, but have nowhere said what the essence of a language-game, and hence of language, is: what is common to all these activities, and what makes them into language or parts of language. So you let yourself off the very part of the investigation that once gave you yourself most headache, the part about the general form of propositions and of language.

And this is true. – Instead of producing, something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all". (Wittgenstein, 1958, §65)

Wittgenstein poses as an example a list of different games: board games, card games, Olympic Games, ball games, ring-a-ring-a-roses, and bouncing a ball against the wall. Next he invites us to find out if those games have something in common consider aspects such as how they stand against luck, skill, losing, winning, amusing, or patience.

Wittgenstein's conclusion is that you can find family resemblances between one or other game, but nothing that runs throughout all of them.

Although there is not a central concept that gives strength to the structure of language, "the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres" (Wittgenstein, 1958, §67). Games constitute families, and there are family resemblances between them giving some similar characteristics between some of the games, but none that runs in all of them.

Nevertheless, Wittgenstein's thought experiments about "all sorts of language-games" help us to infer some implications. For instance, social reality is socially constructed and rooted in the "forms of life" of the actors involved. This derives from the idea that because in a language game "it is not possible to obey a rule 'privately'" (Wittgenstein, 1958, §202), the rules have to be socially constructed.

Also, Language Games can be iterated. When people are engaged in interactions, successful co-ordinations enable the actors to establish common ground, like, for instance, when a group of practitioners develop a method in Kotarbiński's (1966) terms. This common ground can be used as the base for future co-ordinations enabling the apparition of more complex structures such as dialects or structures of meaning particular to the participants (Moldoveanu, 2002). In the MS/ST context this could represent communities of practitioners using and developing a methodology that becomes a communication medium to facilitate, share and develop further experiences.

What is appealing from this approach for this research is that meaning came from the language and actions that constitute the use. Use came from actors and this give us a place to think about actors. Additionally everything on language is a tool. So there is the possibility to consider in an intervention all sorts of tools, word, phrases, body language (because actions are covered in the Language Game). Perhaps it can extent event to methodologies, point that I argue in the next section.

4. Methodologies use as Language Games

"Language is an instrument. Its concepts are instruments. Now perhaps one thinks that it can make no great difference which concepts we employ ... the difference is merely one of convenience." (Wittgenstein, 2001, §569)

Here, I will use Wittgenstein philosophy to understand methodology. I am looking to this for three reasons. First, in MS/ST, they are an important part of how interventions are approached. Second, I need a way to look at methodologies in which actors have a say in how they are used. Third, there is a huge variety of methodologies. Their philosophical and theoretical underpinnings in many cases are not necessarily compatible. Under those circumstances, it is needed a philosophical perspective that will be able to refer to all of them to argue how intention plays in intervention regardless the tools employed.

The idea of using language games for understanding methodology is possible because Wittgenstein treats language as a tool for acting in the world. Examples of what we can do with it are giving and obeying orders, describing objects, reporting, speculating, forming and testing a hypothesis, translating, and asking. (Wittgenstein, 1958).

Wittgenstein also poses the idea that there are innumerable numbers of language games. This opens the opportunity to propose more “tools” in terms of language games.

Mauws and Phillips (1995, p.327) argue that the concept is powerful enough to enable an understanding of organization science and managerial practice in terms of collections of diverse language games or “flexible networks of language games”. In this logic, it is also possible to consider interactions, discourses, practices and interventions (a part of managerial practice) in organisational contexts in terms of Language Games.

The proposal here is a middle ground between Wittgenstein and Mauws and Phillips. A methodology is something not as simple as giving an order (although in the process of applying one, orders can be given). Furthermore, a methodology is not so big that it can encompass all managerial practice, (something more likely for MS/ST practice or a complex intervention process). A methodology can be seen as part of this network of language games; it is a game among others.

Methodologies in particular are specifically designed language games. Inventing methodologies is like “invent[ing] a language” that “could mean to invent an instrument for a particular purpose” (Wittgenstein, 2001, §492). For instance, consider the following instruments with their purposes and their different underpinnings: System dynamics “Explore the operation of a complex real-world system to aid understanding and control”; Soft systems methodology “Learn about and improve a problematic situation by gaining agreement on feasible and desirable changes”; Critical systems heuristics “Provide support for planners and citizens to raise, explore and critique the normative implications of plans and designs” (Mingers, 2003, p.563–564).

However, it is important to clarify that what can be seen as a language game is not properly the methodology. Here I follow Kay and Halpin (1999) when they suggest that a methodology is not the written advice, principles or stages. Methodologies appear when they are put into action by actors in a context. Methodologies can be considered as language games when they consist of the “language and the actions into which they are woven”.

Perhaps the power and flexibility of the concept derive from Wittgenstein using the notion of language games as “objects of comparison which are meant to throw light on the facts of our language” (Wittgenstein, 2001, §130). Using his idea as a postulate, other similarities between language games and methodologies can be proposed:

- If language games are “objects of comparison” it follows that they can be used to learn and compare against methodology use.
- The use of methodologies just as language games involves a “whole, consisting of language and the actions into which it is woven”.
- Just as it happens with games, MS/ST methodologies seem to share family resemblances between some of them but not a feature present in all. You could argue a family resemblance among the methods in classic OR, or the ones in soft approaches. However, it is difficult to see the family resemblance between mathematical programming whose models rely on mathematical equations, and Critical Systems Heuristics in which non experts challenge experts through critical questions.
- Methodologies, like language, evolve, change and grow over time. “if you want to say that this shows them to be incomplete, ask yourself whether our language is

complete; – whether it was so before the symbolism of chemistry and the notation of the infinitesimal calculus were incorporated in it” (Wittgenstein, 2001, §18).

5. Conclusions

The Language Games’ framework advocated in this paper, let me treat a whole set of seemingly dissimilar elements under the same framework. This is so despite differences such as level of elaboration (consider a word and a methodology), or philosophical underpinnings (consider mathematical modelling and storytelling). When used all of them are threads of language and action. Consequently, the whole of the intervention process with all the possible tools that it can encompass can be framed as multiple overlapping Language Games. This means that methods, methodologies, jokes, stories and even the pats on the back can be seen as overlapping fibers giving strength to the process. The possibility to see all of these elements as Language Games also implies that their meaning is not fixed. It is given by the actors through the way in which they use them. Actors are important, and they have a saying in the intervention process.

In addition, there are some interesting ‘side effects’ from working with the notion of language games applied to methodologies. First, for Wittgenstein, rules in language games cannot be private, so the understanding of a methodology use needs to be seen as a social construction. Even in the case of a single use by single individual, the concepts and understanding from which s/he will draw from have their origin in social interactions.

Secondly, apart from methodology use, the interactions, languages, activities and “forms of life” in the intervention context can also be considered in terms of language games. So when we are intervening, what we are trying to do using methodologies (as a language game) is to affect the language games already in place, which is to say, we are using a tool to modify the tools that people in that context had developed in order to interact. What is more, because tools modify tools, it is also likely that the language games in place will modify the methodology in use.

Consequently, considering the use of methodologies as Language Games is not directed to shed light on the underpinnings of each methodology. It will not help, for example, to improve the mathematics behind a methodology that relies on that kind of knowledge. Considering their use as language games is aimed at understanding the possibilities and effects that the use of those tools has on the “activities” or “forms of life” of the actors engage in the process.

References

- Garfinkel, A. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory*. New Haven: Yale University Press.
- Gergen, K. J., & Leach, C. W. 2001. “Introduction: The Challenge of Reconstruction”. *Political Psychology*, 22(2): 227–232.
- Gergen, M. M., & Gergen, K. J. 2003. *Social Construction: A Reader*. London: SAGE.
- Hassard, J. 1990. “An alternative to paradigm incommensurability in organization theory”. In J. Hassard, & D. Pym (Eds.), *The Theory and Philosophy of*

- Organizations: Critical Issues and New Perspectives*: 219–230. London: Routledge.
- Jackson, M. C. 2000. *Systems Approaches to Management*. New York: Kluwer/Plenum.
- Jackson, M. C., & Keys, P. 1984. “Towards a System of Systems Methodologies”. *The Journal of the Operational Research Society*, 35(6): 473–486.
- Kay, R., & Halpin, D. 1999. “Redefining the role of the practitioner in critical systems methodologies”. *Systems Research and Behavioral Science*, 16(3): 273.
- Keys, P. 1997. “Approaches to understanding the process of OR: Review, critique and extension”. *Omega*, 25(1): 1–13.
- Keys, P. 2002. “Process of MS/OR”. In H. G. Daellenbach, & R. L. Flood (Eds.), *The Informed Student Guide to Management Science*: 211–212. London: Thomson.
- Kotarbiński, T. 1966. *Gnosiology: The Scientific Approach to the Theory of Knowledge* (O. Wojtasiewicz, Trans.). Oxford: Pergamon.
- Maturana, H. R. 1988. “Reality: the search for objectivity or the quest for a compelling argument”. *Irish Journal of Psychology*, 9(1): 25–82.
- Mauws, M. K., & Phillips, N. 1995. “Understanding Language Games”. *Organization Science*, 6(3): 322–334.
- Mingers, J. 2003. “A classification of the philosophical assumptions of management science methods”. *The Journal of the Operational Research Society*, 54(6): 559.
- Moldoveanu, M. 2002. “Language, games and language games”. *Journal of Socio - Economics*, 31(3): 233.
- Nodoushani, O. 1999. “Systems Thinking and Management Epistemology”. *Systemic Practice and Action Research*, 12(6): 557–571.
- Schwandt, T. A. 2000. “Three Epistemological Stances For Qualitative Inquiry: Interpretivism, Hermeneutics, and Social Constructionism”. In N. K. Denzin, & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research*, 2nd ed.: 189–213. Thousand Oaks: SAGE.
- Stern, D. G. 2004. *Wittgenstein's Philosophical Investigations: An Introduction*. Cambridge, UK: Cambridge University Press.
- Varela, F. J. 1979. *Principles of Biological Autonomy*. Limerick: Elsevier.
- Von Foerster, H. 1989. “The Need of Perception for the Perception of Needs”. *Leonardo*, 22(2): 223–226.
- Von Glasersfeld, E. 1996. “Farawell to Objectivity”. *Systems Research*, 13(3): 279–286.
- White, L., & Taket, A. 1997. “Critiquing Multimethodology as metametodology: working towards pragmatic pluralism”. In J. Mingers, & A. Gill (Eds.), *Multimethodology: The Theory and Practice of Combining Management Science Methodologies*: 379–405. Chichester: Wiley.
- Winograd, T., & Flores, F. 1986. *Understanding computers and cognition: a new foundation for design*. Norwood, N.J: Ablex.
- Wittgenstein, L. 1922. *Tractatus logico-philosophicus* (C. K. Ogden, Trans.). London: Kegan Paul, Trench, Trubner & Co.
- Wittgenstein, L. 1958. *Philosophical Investigations* (3rd ed.). Oxford: Basil Blackwell.
- Wittgenstein, L. 2001. *Philosophical Investigations* (G. E. M. Ascombe, Trans.) (3rd ed.). Oxford: Blackwell Publishing.

Bi-Linear Reductions for the Multiprocessor Scheduling Problem With Communication Delays Using Integer Linear Programming

Sarad Venugopalan and Oliver Sinnen
Department of Electrical and Computer Engineering
University of Auckland
New Zealand
sven251@aucklanduni.ac.nz, o.sinnen@auckland.ac.nz

Abstract

With computer processors running at speeds closer to their theoretical limit, the recent focus has turned to the use of parallelism in hardware by the use of multi-core processors for speedup. However, duplicating processors do not automatically translate to faster task execution. The tasks are to be carefully assigned and scheduled so that their total execution time on the multiple processors is minimal. We propose an optimal Integer Linear Programming formulation for the Multiprocessor Scheduling Problem with Communication Delays (MSPCD). The formulations use an effective method to linearise the bi-linear forms arising out of communication delays and introduce new overlap constraints to ensure that no two tasks running on the same processor overlap in time. The proposed formulation is compared with known ILP formulations that solve the MSPCD problem.

1 Introduction

In the past, engineering and science have strongly benefited from an exponential growth in processor performance. Yet, due to the reached physical limits of processor technology the improvements are fading out (Olukotun and Hammond 2005) and manufacturers have moved to multi(core)processors. With multiple processors, however, the performance growth is not automatic (Grama et al. 2003) and can only be achieved when the processors are efficiently employed in parallel. Existing scheduling algorithms are therefore heuristics that try to produce good rather than optimal schedules, e.g. Löwe and Zimmermann (1999), Palmer and Sinnen (2008), Hagraas and Janecek (2005), Radulescu and van Gemund (2002), Sinnen (2007), Yang and Gerasoulis (1993), Zomaya, Ward, and Macey (1999), Coffman Jr. and Graham (1972), Hwang et al. (1989). However, having optimal schedules can make a fundamental difference, e.g. for time critical systems or to enable the precise evaluation of scheduling heuristics. Given the NP-hardness of the processor scheduling

problem (Sarkar 1989), finding an optimal solution requires an exhaustive search of the entire solution space. For scheduling, this solution space is spanned by all possible processor assignments combined with all possible task ordering. Clearly this search space grows exponentially with the number of tasks, thus it becomes difficult already for very small task graphs. However, recent improvements in the Integer Linear Programming (ILP) formulation for the Multiprocessor Scheduling Problem with Communication Delays (MSPCD) in Davidović et al. (2007), Venugopalan and Sinnen (2012) allows larger instances of task graphs to be solved in a shorter time. The contribution of this work is to model the ILP based on the task overlap variable defined in Davare et al. (2006) and compare it with the results in (Davidović et al. 2007) and (Venugopalan and Sinnen 2012).

2 Task Scheduling Model

A fully connected network of homogeneous multiprocessors $P = \{1, \dots, |P|\}$ with identical communication links is assumed. Each processor may execute several tasks but each task has to be assigned to exactly one processor, in which it is entirely executed without pre-emption. Further, no multitasking or parallelism is permitted within a task. The tasks to be scheduled are represented by a directed acyclic graph (DAG) defined by a 4-tuple $G = (V, E, C, L)$ where $i \in V$ denotes the set of tasks and $(i, j) \in E$ represents the set of edges. The set E defines precedence relation between tasks. A task cannot be executed unless all of its predecessors have completed their execution and all relevant data is available. The set $C = \{\gamma_{ij} : i, j \in V\}$ denotes the set of edge communication time. If tasks i and j are executed on different processors $h, k \in P, h \neq k$, they incur a communication time penalty γ_{ij} . If both tasks are scheduled to the same processor the communication time is zero. For a graph with n tasks, the set $L = \{L_1, \dots, L_n\}$ represents the task computation times (execution time length). Let $\delta^-(j)$ be the set of precedents of task i , that is $\delta^-(j) = \{i \in V | (i, j) \in E\}$.

3 Related Work

Different approaches have been proposed to solve the MSPCD. One popular approach to the MSPCD makes use of Linear Programming (Davidović et al. 2007). This involves linearisation of the bilinear forms resulting from communication delays. The work in Davidović et al. (2007) discusses a classic formulation and a packing formulation of the MSPCD. Their results indicate that the packing formulation is about 5000 times faster than the classic formulation. The work in Venugopalan and Sinnen (2012) further improves the ILP formulations in Davidović et al. (2007) and introduces two ILP formulations, one which runs faster when scheduled over a smaller number of processors and the other when scheduled over a larger number of processors. Another popular approach to solve the MSPCD is to use the A* search algorithm and is discussed in Kwok and Ahmad (2005), Shahul, Zaki, and Sinnen (2010), Dechter and Pearl (1985), Russell and Norvig (2010). A* is a best-first search technique and also a popular Artificial Intelligence algorithm guided by a problem specific cost function $f(s)$ for each solution state s . The main drawback of A* is that it keeps all the nodes in memory and it usually runs out of memory long

before it runs out of time making it unusable for medium and large sized problem instances.

4 Bi-Linear Reductions

Let t_i be the start time of task i and t_j the start time of task j . Let L_i be the execution time of task i and γ_{ij} be the communication time between tasks i and j . Further, let $P = |P|$ be the total number of processors available for scheduling. Define

$$x_{ih} = \begin{cases} 1 & \text{task } i \text{ runs on processor } h \\ 0 & \text{otherwise.} \end{cases}$$

If any two tasks i and j incur a communication cost, then

$$\forall j \in V : i \in \delta^-(j) \quad t_i + L_i + \sum_{h,k \in P} \gamma_{ij}(x_{ih} \cdot x_{jk}) \leq t_j \quad (1)$$

By definition, x_{ih} and x_{jk} are Boolean variables and their multiplication needs to be linearised. The previous best linearisation in Davidović et al. (2007) uses two linearisation approaches. All though, the communication model in Davidović et al. (2007) has no restriction on the number of communication links, the model presented here assumes a fully connected network. Their linearisation variable z_{ij}^{hk} wherein task i runs on processor h and task j runs on processor k , is defined as

$$\forall j \in V : i \in \delta^-(j), h, k \in P \quad (z_{ij}^{hk} = x_{ih} \cdot x_{jk})$$

Using this definition, the multiplication of the Boolean variables $x_{ih} \cdot x_{jk}$ is replaced by the linearisation variable z_{ij}^{hk} in (2).

$$\forall j \in V : i \in \delta^-(j) \quad t_i + L_i + \sum_{h,k \in P} \gamma_{ij} \cdot z_{ij}^{hk} \leq t_j \quad (2)$$

By (2), the number of constraints produced is $|E|$ and the number of variables per constraint in terms of the processor combinations over z_{ij}^{hk} is $O(|P|^2)$. The first linearisation uses constraints (2) and (3) - (5).

$$\forall j \in V, i \in \delta^-(j), h, k \in P \quad x_{ih} \geq z_{ij}^{hk} \quad (3)$$

$$\forall j \in V, i \in \delta^-(j), h, k \in P \quad x_{jk} \geq z_{ij}^{hk} \quad (4)$$

$$\forall j \in V, i \in \delta^-(j), h, k \in P \quad x_{ih} + x_{jk} - 1 \leq z_{ij}^{hk} \quad (5)$$

By (3) - (5), the number of constraints produced is $|E||P|^2$ and the number of variables per constraint is $O(1)$. Hence, the complexity of the first linearisation by (2), (3) - (5) in terms of number of constraints is $O(|E||P|^2)$.

The second linearisation uses constraints (2) and (6) - (7).

$$\forall i \neq j \in V, k \in P \quad \sum_{h \in P} z_{ij}^{hk} = x_{jk} \quad (6)$$

$$\forall i \neq j \in V, h, k \in P \quad z_{ij}^{hk} = z_{ji}^{kh} \quad (7)$$

By (6), the number of constraints generated is $O(|V|^2|P|)$ and the number of variables per constraint is $O(|P|)$. So, the complexity of the second linearisation by (2) and (6) in terms of number of constraints is $O(|E| + |V|^2|P|) = O(|V|^2|P|)$.

5 Proposed Formulation

In this section a new ILP formulation (ILP-DELTA) is proposed and compared with the Packing formulation in (Davidović et al. 2007) and ILP-TC in (Venugopalan and Sinnen 2012). The ILP formulations, ILP-DELTA AND ILP-TC differ in the definition of the task overlap variable. They eliminate the use of the variable z for the linearisation of the bi-linear forms. This frees up to $|V|^2|P|^2$, z variables in the ILP formulation and speeds up the solution time.

5.1 ILP-DELTA

For each task $i \in V$ let $t_i \in \mathbf{R}$ be the start execution time and $p_i \in \mathbf{N}$ be the ID of the processor on which task i is to be executed. Let W be the total makespan and $|P|$ the number of processors available. The task overlap variable Δ_{ij} is modeled similar to the definition of the task overlap variable o_{ij} in Davare et al. (2006). If any two tasks i and j have a serial ordering in time, one of Δ_{ij} or Δ_{ji} is set to 1. Both Δ_{ij} and Δ_{ji} are set to 1 if the two tasks overlap in time. The variables Δ_{ij} and ϵ_{ij} are defined as follows:

$$\forall i, j \in V \quad \Delta_{ij} = \begin{cases} 1 & \text{task } j \text{ finishes after task } i \text{ starts} \\ 0 & \text{otherwise} \end{cases}$$

$$\forall i, j \in V \quad \epsilon_{ij} = \begin{cases} 1 & \text{the processor index of task } i \text{ is strictly less than task } j \\ 0 & \text{otherwise} \end{cases}$$

$$\min \quad W \quad (11)$$

$$\forall i \in V \quad t_i + L_i \leq W \quad (12)$$

$$\forall i \neq j \in V \quad \Delta_{ij} + \Delta_{ji} \geq 1 \quad (13)$$

$$\forall i \neq j \in V \quad \epsilon_{ij} + \epsilon_{ji} \leq 1 \quad (14)$$

$$\forall i \neq j \neq k \in V \quad \epsilon_{ij} + \epsilon_{jk} \geq \epsilon_{ik} \quad (15)$$

$$\forall i \neq j \in V \quad \Delta_{ij} + \Delta_{ji} + \epsilon_{ij} + \epsilon_{ji} \geq 1 \quad (16)$$

$$\forall i \neq j \in V \quad \Delta_{ij} + \Delta_{ji} - 1 \leq \epsilon_{ij} + \epsilon_{ji} \quad (17)$$

$$\forall i \neq j \in V \quad p_j - p_i - 1 - (\epsilon_{ij} - 1)|P| \geq 0 \quad (18)$$

$$\forall j \in V : i \in \delta^-(j) \quad t_i + L_i + \gamma_{ij}(\epsilon_{ij} + \epsilon_{ji}) \leq t_j \quad (19)$$

$$\forall i \neq j \in V \quad t_j - t_i - L_i - (\Delta_{ij} - \Delta_{ji} - 1)W_{max} \geq 0 \quad (20)$$

$$\forall j \in V : i \in \delta^-(j) \quad \Delta_{ij} = 1 \quad (21)$$

$$W \geq 0 \quad (22)$$

$$\forall i \in V \quad t_i \geq 0 \quad (23)$$

$$\forall i, j \in V \quad \Delta_{ij}, \epsilon_{ij} \in \{0, 1\} \quad (24)$$

$$\forall i \in V \quad p_i \in \{1, \dots, |P|\} \quad (25)$$

The upper bound on the makespan W is given by W_{max}

$$W_{max} = \sum_{i \in V} L_i + \sum_{i, j \in V} c_{ij} \quad (26)$$

Constraints (11) and (12) are min-max constraints and minimises the maximum start task execution times. Constraint (12) specifies that the sum of task start time and its execution time is to be less than or equal to the makespan W . Constraints (13)-(17) are overlap constraints. Together, they ensure that no two tasks overlap in time and space. I.e. if two tasks have overlapping execution times, then they must run on different processors. The variable Δ defines a serial ordering of tasks in time if Δ_{ij} or Δ_{ji} is set to 1. If the execution of two tasks i and j overlap in

time, both Δ_{ij} and Δ_{ji} are simultaneously set to 1. By constraint (13), at least one of Δ_{ij} or Δ_{ji} is set to 1. Constraint (14) mandates that the sum of ϵ_{ij} and ϵ_{ji} be less than or equal to 1. If two tasks i and j run on the same processor, then both ϵ_{ij} and ϵ_{ji} are set to 0. If the two tasks run on different processors, then one of ϵ_{ij} or ϵ_{ji} is set to 1, depending on which of the two task is assigned to a higher processor index. Both ϵ_{ij} and ϵ_{ji} cannot be simultaneously set to 1, as it is not possible to assign a task to a higher and lower processor index at the same time. In constraint (15), the ϵ variables enforces a partial ordering of the processor indices with the help of an additional transitivity clause. Constraint (16) prevents task executions from overlapping in time on the same processor by setting either one of the Δ or ϵ variables to 1. By constraint (17), if any two tasks i and j overlap in time (i.e. both Δ_{ij} and Δ_{ji} are set to 1), then either ϵ_{ij} or ϵ_{ji} is set to 1. Constraint (18) sets the processor constraint. It is used to enforces the condition, $p_j > p_i + 1$, if $\epsilon_{ij} = 1$. This ensures that if $\epsilon_{ij} = 1$, then the processor index of task j is higher than task i . Constraint (19) and (20) are timing constraints. Constraint (19) models the communication between tasks with edges. If any two tasks i and j run on different processors, then ϵ_{ij} or ϵ_{ji} is set to 1. This implies that a communication cost is incurred. If tasks i and j run on the same processor, then ϵ_{ij} and ϵ_{ji} are both set to 0. In this case, constraint (19) reduces to $t_i + L_i \leq t_j$. For all tasks, constraint (20) enforces the condition $t_j \geq t_i + L_i$ if and only if $\Delta_{ij} = 1$ and $\Delta_{ji} = 0$. I.e. only when the tasks have a serial ordering in time. Constraint (21) is an edge constraint. All variables Δ_{ij} for which there is an edge from task i to task j in the graph is set to 1. Constraints (22) to (26) are the bounds on the ILP formulation. In constraint (26), the upper bound on the makespan is computed as the sum of all task execution costs and edge communication costs in the graph.

5.2 ILP-TC

In this formulation (Venugopalan and Sinnen 2012) the variable σ defines a serial task execution order in time if one of σ_{ij} or σ_{ji} is set to 1. If the tasks overlap in time, both σ_{ij} and σ_{ji} are set to 0. The variables σ_{ij} and ϵ_{ij} are defined as follows:

$$\forall i, j \in V \quad \sigma_{ij} = \begin{cases} 1 & \text{task } i \text{ finishes before task } j \text{ starts} \\ 0 & \text{otherwise} \end{cases}$$

$$\forall i, j \in V \quad \epsilon_{ij} = \begin{cases} 1 & \text{the processor index of task } i \text{ is strictly less than task } j \\ 0 & \text{otherwise} \end{cases}$$

In ILP-TC, constraint (17) is removed since it is no longer required by the definition of σ_{ij} . Constraints (13),(16) and (20) are replaced by constraints (31),(32) and (33) respectively. All other constraints remain unchanged.

$$\forall i \neq j \in V \quad \sigma_{ij} + \sigma_{ji} \leq 1 \quad (31)$$

$$\forall i \neq j \in V \quad \sigma_{ij} + \sigma_{ji} + \epsilon_{ij} + \epsilon_{ji} \geq 1 \quad (32)$$

$$\forall i \neq j \in V \quad t_j \geq t_i + L_i + (\sigma_{ij} - 1)W_{max} \quad (33)$$

By constraint (31), the sum of σ_{ij} and σ_{ji} is utmost one. If there is a serial ordering of the tasks executed, either σ_{ij} or σ_{ji} is set to 1 depending on which task finishes its execution before the other starts. If the two tasks overlap in time, both

σ_{ij} and σ_{ji} are set to 0. By constraint (32), at least one of the 4 variables ϵ_{ij} , ϵ_{ji} , σ_{ij} or σ_{ji} must be set to 1. Constraint (32) ensures that no two tasks i and j run on the same processor if their execution overlaps in time. By constraint (33), if $\sigma_{ij} = 1$ then $t_j \geq t_i + L_i$.

Both ILP-DELTA and ILP-TC differ by 3 constraints (31),(32) and (33). A simpler definition of the task overlap variable in ILP-TC allows (17) to be dropped. ILP-DELTA and ILP-TC have a constraint complexity of $O(|V|^3)$ due to (15). However, both these formulations have only $O(|V|^2)$ variables ($\sigma_{ij}, \epsilon_{ij}$) to assign a value to. Computational results indicate that though these formulations are faster than the Packing formulation, there is no clear winner between ILP-DELTA and ILP-TC as they exhibit a similar run time.

6 Computational Results

In this section, we compare the run times of the proposed formulation (ILP-DELTA) with the Packing formulation (Davidović et al. 2007) and ILP-TC (Venugopalan and Sinnen 2012). The computations are carried out using CPLEX 11.0.0 (ILOG 2007) on an Intel Core i3 processor 330M, 2.13 GHZ CPU and 2 GB RAM running with no parallel mode and on a single thread on Windows 7.

6.1 Experimental Setup and Result Table

All experiments are run for a fully connected processor network with identical bandwidth capacity. The input graphs for this comparison are taken from those proposed and used in Davidović et al. (2007), Davidović and Crainic (2006). The graph files with a name starting with `ogra_` are suffixed with the number of tasks in that file followed by its edge density in terms of a percentage of the maximum possible number of edges (I.e. $|V|(|V| - 1)/2$). According to Davidović et al. (2007), they have a special graph structure that makes it hard to find the task ordering which yields the optimal solution when the number of mutually independent tasks is large. The graph file with a name starting with `t_` were generated randomly and are suffixed with the number of tasks in that file followed by its edge density and the index used to distinguish graphs of the same characteristics. The Stencil graph is suffixed by the number of tasks followed by the Computational cost to Communication Ratio (CCR) value.

The experiments are run on small to medium sized instances of the graphs on 8 processors. ILP-TC (Davidović et al. 2007) is designed to work well for tasks scheduled on to a larger number of processors. Since ILP-DELTA is modelled similar to ILP-TC, the proposed ILP also works well for tasks assigned to a larger number of processors. For tasks assigned to a smaller number of processors (e.g. 2 or 4), ILP-RBL (Davidović et al. 2007) is well tailored and suitable for the purpose. A 24 hour time out is set for all the task graphs solved. If the execution exceeds 24 hours, the execution is terminated and the *gap* recorded. The *gap* gives a guaranteed lower bound on the optimal schedule length. For e.g. if the gap is 0.69% at 24 hours, it implies the optimal schedule length is within 0.69% of the schedule length returned by the ILP solver at 24 hours. The usual timing convention h:m:s is used to denote hours:minutes:seconds.

Table 1 compares the solution time of ILP-DELTA with the Packing Formulation

Table 1: Solution Time Comparison of ILP-DELTA with Packing and ILP-TC

Graph	p	n	Packing	ILP- DELTA	ILP- TC
Ogra20_75	8	20	51m:28s	3m:30s	2m:37s
t20_90	8	20	2m:24s	2s	7s
t30_30_2	8	30	24h, 0.69% gap	5h:15m:11s	4h:36m:18s
t30_60_1	8	30	7h:22m:19s	1h:39m:10s	2h:6m:54s
Stencil15_CCR_1	8	15	14m	35s	16s

and ILP-TC. For these instances the solution time of ILP-DELTA or ILP-TC is found to be 5 to 20 times or upward faster than the best version of the Packing formulation. The result table indicates that changing the definition of the task overlap variable does not result in a significant difference in the run time between ILP-DELTA and ILP-TC.

7 Conclusion

An ILP formulation for the MSPCD was proposed and modeled based on the task overlap variable defined in Davare et al. (2006) and compared with known ILP formulations in (Davidović et al. 2007) and (Venugopalan and Sinnen 2012). The ILP formulations in ILP-DELTA and ILP-TC eliminated the use of the variable z for the linearisation of the bi-linear forms, hence speeding up the solution time. It was found that the proposed formulation easily outperforms the packing formulation in Davidović et al. (2007) but had a similar run time as compared to ILP-TC, when scheduled on a larger number of processors. Although ILP-DELTA and ILP-TC use different task overlap variables, their run time were similar when formulated in a concise form.

Acknowledgement

We gratefully acknowledge that this work is supported by the Marsden Fund Council from Government funding, Grant 9073-3624767, administered by the Royal Society of New Zealand.

References

- Coffman Jr., E. G., and R. L. Graham. 1972. “Optimal scheduling for two-processor systems.” *Acta Informatica* 1:200–213.
- Davare, Abhijit, Jike Chong, Qi Zhu, Douglas Michael Densmore, and Alberto L. Sangiovanni-Vincentelli. 2006, Dec. “Classification, Customization, and Characterization: Using MILP for Task Allocation and Scheduling.” Technical Report UCB/EECS-2006-166, EECS Department, University of California, Berkeley.
- Davidović, T., L. Liberti, N. Maculan, and N. Mladenovic. 2007. “Towards the Optimal solution of the Multiprocessor Scheduling Problem with Communication Delays.” *3rd Multidisciplinary International Conference on Scheduling: Theory and Application*. 128–135.

- Davidović, Tatjana, and Teodor Gabriel Crainic. 2006. “Benchmark-problem instances for static scheduling of task graphs with communication delays on homogeneous multiprocessor systems.” *Computers and Operations Research* 33 (August): 2155–2177.
- Dechter, Rina, and Judea Pearl. 1985. “Generalized best-first search strategies and the optimality of A*.” *Journal of ACM* 32 (3): 505–536 (July).
- Grama, A., A. Gupta, G. Karypis, and V. Kumar. 2003. *Introduction to Parallel Computing*. Second. Pearson, Addison Wesley.
- Hagras, T., and J. Janecek. 2005. “A high performance, low complexity algorithm for compile-time task scheduling in heterogeneous systems.” *Parallel Computing* 31 (7): 653 – 670.
- Hwang, Jing-Jang, Yuan-Chieh Chow, Frank D. Anger, and Chung-Yee Lee. 1989. “Scheduling Precedence Graphs in Systems with Interprocessor Communication Times.” *SIAM Journal on Computing* 18 (2): 244–257.
- ILOG. 2007. “ILOG CPLEX 11.0 User’s Manual.” ILOG S.A., Gentilly, France.
- Kwok, Yu-Kwong, and Ishfaq Ahmad. 2005. “On multiprocessor task scheduling using efficient state space search approaches.” *Journal of Parallel and Distributed Computing* 65 (12): 1515 – 1532.
- Löwe, Welf, and Wolf Zimmermann. 1999. “Scheduling Iterative Programs onto LogP-Machine.” Euro Par 99 Parallel Processing, Springer, Lecture Notes in Computer Science.
- Olukotun, K., and L. Hammond. 2005., September. *The future of microprocessors*. Volume 3. Queue - Multiprocessors.
- Palmer, A., and O. Sinnen. 2008. “Scheduling Algorithm Based on Force Directed Clustering.” dec., 311 –318. Parallel and Distributed Computing, Applications and Technologies, 2008. PDCAT 2008. Ninth International Conference on.
- Radulescu, A., and A.J.C. van Gemund. 2002. “Low-cost task scheduling for distributed-memory machines.” *Parallel and Distributed Systems, IEEE Transactions on* vol 13 (6): 648 –658 (June).
- Russell, S.J., and P. Norvig. 2010. *Artificial intelligence: a modern approach*. Prentice hall.
- Sarkar, V. 1989. *Partitioning and scheduling parallel programs for multiprocessors*. MIT press.
- Shahul, S., A. Zaki, and O. Sinnen. 2010. “Scheduling task graphs optimally with A*.” *Journal of Supercomputing* 51 (3): 310–332 (March).
- Sinnen, Oliver. 2007. *Task Scheduling for Parallel Systems (Wiley Series on Parallel and Distributed Computing)*. Wiley-Interscience.
- Venugopalan, Sarad, and Oliver Sinnen. 2012. “Optimal Linear Programming Solutions for Multiprocessor Scheduling with Communication Delays.” *ICA3PP (1)*. 129–138.
- Yang, Tao, and Apostolos Gerasoulis. 1993. “List scheduling with and without communication delays.” *Parallel Computing* 19 (12): 1321 – 1344.

Zomaya, A.Y., C. Ward, and B. Macey. 1999. "Genetic scheduling for parallel processor systems: comparative studies and performance issues." *IEEE Transactions on Parallel and Distributed Systems* 10 (8): 795–812 (August).

Organizational Goals, Feedback Effects, and Performance

Miles M. Yang
Michael Shayne Gary
Philip W. Yetton
Australian School of Business
University of New South Wales
Australia
miles2yang@gmail.com

Abstract

The effects of stretch goals on organizational performance are investigated in an experimental study using the well-known People Express flight simulator. 106 managers, with 15 years of work experience on average, participated in the experiment. The results show that stretch goals for growing profits lead to higher variance in performance compared with the effects of moderate goals, but do not increase the level of performance of the median firm. Surprisingly, stretch goals do not lead to higher bankruptcy rates compared with the effects of moderate goals. In addition, supplementary analysis shows that stretch goals do not improve decision makers' mental models as implied by the literature in organizational learning. Instead, our data suggests that decision makers in stretch and moderate goal conditions use similar information cues in making strategic decisions for aircraft orders, fare, target service scope, and employee hiring. This research extends the theory on organizational goals to explain how different goal levels impact performance outcomes in a dynamic decision making task. The findings also raise a warning that adopting stretch goals may lead to increased variance in financial performance without increasing the expected value of performance outcomes for the typical firm.

Key Words: organizational goals, mental models, aspirations, performance variance

1 Introduction

The role of organizational goals in guiding decision-making and the search for alternative courses of action has a rich history in organization research (Cyert & March, 1963). Goals, or aspiration-level reference points, are central to modern theories of individual and organizational choice (March & Shapira, 1992). Performance targets impact how managers interpret their experience and frame their responses (Denrell & March, 2001; Lant, 1992; Lant & Shapira, 2008; March & Shapira, 1987).

Current business norms suggest that the board of directors should set difficult or even impossible organizational goals for managers because stretch goals stimulate exploration rather than exploiting current routines. The exploration of unknown possibilities may lead to outstanding performance. This normative advice has been widely accepted by a broad range of organizations around the world. The use of stretch goals is fairly common in business practice including in companies such as TOYOTA and General Electric (Collins & Porras, 2002; Oxnard, 2004). However, there is limited

empirical research examining the effects of adopting stretch goals for organizational performance. In addition, more research is needed to understand the mechanisms through which stretch goals impact performance (Sitkin, See, Miller, Lawless, & Carton, 2011).

Evaluating organizational performance relative to goals informs managerial decisions to allocate and mobilize resources in efforts to achieve those goals (Fiegenbaum, Hart, & Schendel, 1996; Greve, 1998). When performance is above aspirations, firms are expected to continue current activities and routines, avoid actions that might result in performance below goals, and to strive for slightly higher performance (Bromiley, 1991; Bromiley, Miller, & Rau, 2001). When performance falls short of aspirations, organizations increase search activity and change by selecting new strategies and courses of action in attempts to increase performance (Greve, 1998; Lant, 1992; March & Shapira, 1992).

While it is widely agreed that organizational goals influence strategic behavior and performance outcomes (Greve, 2008), the effects of different goal levels on organizational performance is still not well understood. Organizations frequently set higher goals for profit and growth in order to motivate managers to achieve these goals (Lant, 1992). However, there is very little empirical evidence about the impact of very aggressive, stretch organizational goals on either the levels or the variance in performance outcomes in comparison to less aggressive goals.

We need to better understand how organizational goals for overall financial performance impact managerial decisions and strategies and, ultimately, the implications of different goal levels on both the levels of and variance in organization performance.

Here, we report the results of an experimental study examining the effects of stretch organizational profit growth goals on the median levels of and variance in performance outcomes. The results show that stretch goals lead to higher variance in performance outcomes compared with moderate goals. In addition, the results of supplementary analysis show that firms with stretch and moderate goals have similar mental models. Without appropriate mental models, decision makers managing firms with stretch goals experience compensating feedbacks, pushing for an aggressive asset growth strategy, but only a few of them achieve outstanding performance. Most decision makers assigned stretch goals remain far below from the assigned goal.

2 Theory and Hypotheses

The example performance distributions shown in Figure 1 illustrate how different goal levels could cause different means and variances in performance outcomes. In the left example, the mean goal effect is constant across all firms that were originally in the low goal condition, and the variance or standard deviation is equal across the two goal conditions. These are the implicit assumptions, or the central tendency, in much of the research on goal setting theory (Locke & Latham, 2002) is that adopting a stretch goal shifts the entire performance distribution to a higher mean (increasing performance).

In contrast, the example distribution in the right half of Figure 1 provides another possibility. If we relax the assumption of constant mean goal effects across all firms, research on variable risk preferences suggest one reason to expect variable goal effects across firms when moving from lower to higher (stretch) organizational goals (March & Shapira, 1992; Miller & Chen, 2004). In this case, firms with stretch goals take more risks and therefore the variance in performance also increases. The example is drawn such that there is no significant difference in mean goal effects, but that there is a significant difference in the variance in performance outcomes between the two goal levels.

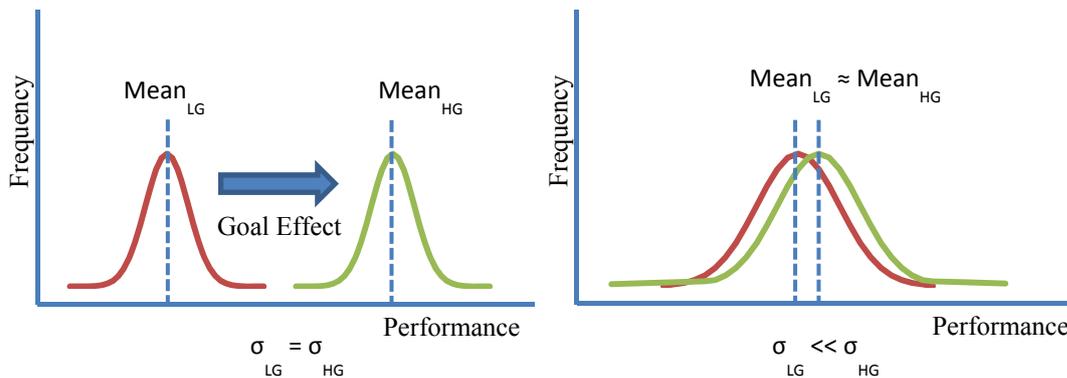


Figure 1 Two possible example performance distributions for moderate and stretch goals effects

Another possible reason to expect variable goal effects across firms faced with stretch goals is that there are differential learning rates that lead to higher diversity in the mental models governing strategic choices when firms are faced with stretch goals (Gary & Wood, 2011). Under stretch goals, firms engage in greater problem driven search attempting to discover strategies for achieving the targets. A small number of firms identify high quality strategies and achieve extraordinarily high performance. However, the vast majority of firms with stretch goals continue to perform far below the potential achievable levels because improving mental models and discovering high quality strategies for complex organizations in dynamic environments is difficult.

Hypothesis 1: Stretch organizational profit growth goals will lead to higher variance in firm performance outcomes compared with moderate profit growth goals.

Goal setting theory provides extensive evidence that specific, stretch goals increase mean performance across a wide range of tasks (for a reviews of this extensive research area see: Locke & Latham, 2002; Locke, Shaw, Saari, & Latham, 1981). However, a number of prior studies have shown that specific, stretch goals have mixed effects on performance outcomes for complex tasks (Wood, Mento, & Locke, 1987). A meta-analysis showed that the goal effect on mean performance decreases as task complexity increases (Wood et al., 1987). Managing a complex organization in a dynamic competitive environment certainly qualifies as a very complex task. Overall, research has shown that difficult goals sometimes improve performance on complex tasks, but they also can result in lower performance (Larrick, Heath, & Wu, 2009).

Hypothesis 2a: Stretch organizational profit growth goals will lead to higher average firm performance outcomes compared with moderate profit growth goals.

Hypothesis 2b: Stretch organizational profit growth goals will lead to lower average firm performance outcomes compared with moderate profit growth goals.

Stretch goals by definition are targets that only few or even none firms can achieve (Sitkin et al., 2011). The discrepancy gap between actual performance and stretch goal will typically be much greater than the gap between actual performance and moderate goals. The desire to close the discrepancy gap leads decision makers generally to increase risk taking (Greve, 1998; Lee, 1997; March & Shapira, 1987; Singh, 1986). In an empirical study, Larrick et al. (2009) show that goal difficulty increases risky behavior in a negotiation decision making task. However, risky strategies do not always pay off. Decision makers with poor mental models are less likely to perform well when taking risks (Sterman, 1989a). Decision makers assigned stretch goals try high risk

strategies, for example doubling aircraft capacity in a short period of time, and may get into trouble quickly. Higher risk may be associated with more performance failure in which decision makers fall into vicious spiral and then go bankrupt.

Hypothesis 3: Stretch compared with moderate organizational goals lead to higher bankruptcy rate.

3 Experimental Design

To test Hypotheses 1-3, we designed an experimental study using organizational profit goals. We used two different goal levels—one stretch goal and one moderate (more easily achieved) goal.

Participants. The participants were 106 managers enrolled in an Executive MBA course who took part in the simulation as a class exercise. The participants were 36 years of age and had 15 years of work experience on average. All participants were randomly assigned to a team and each team was randomly assigned to manage either a firm with stretch profit growth goals ($n = 25$) or a firm with moderate profit growth goals ($n = 25$).

Task. The class exercise was an interactive, computer-based simulation of an airline operating in a competitive market. The management simulation has been utilized in previous research and captures many well-established features of competition between new entrants and incumbents in the airline industry (Bakken, Gould, & Kim, 1992). Participants take on the role of the top management team of an airline and make quarterly decisions for aircraft orders, employee hiring, average fare (price), marketing spend and service scope. Their goal is to deliver on a stretch or moderate cumulative profit goal over a forty-quarter simulation.

The business environment changes as a consequence of participants' decisions. It includes a large number of interdependent variables with multiple feedback effects, time delays, nonlinear relationships, and stock accumulations (Graham, Morecroft, Senge, & Sterman, 1992). These features of the management simulation also characterize the sort of complex environments that senior managers typically operate in while making strategic decisions.

Procedures. In both goal conditions, teams were given a specific financial target. This was either a moderate or stretch profit goal measured in cumulative profits over 40 quarters. Teams in the stretch (moderate) goal group were told, "The Board of Directors has set your Cumulative Net Income target equal to \$315 (\$60) million by the end of 10 years. This long term growth in profit will deliver the financial results that our shareholders expect." The moderate goal represents a 12 percent compound annual growth in cumulative profit. The stretch goal represents a 32 percent compound annual growth. These levels of growth are achievable in the simulation with a range of different strategies. The stretch goal was based on the 90th percentile performance levels achieved in pilot tests in which decision makers were instructed to, "Do Your Best to maximize Cumulative Net income".

Cumulative net income (profit) was adopted as the most suitable organizational performance measure. Recent research findings show that senior managers use a wide variety of primarily internal referents to assess performance (Short & Palmer, 2003).

Most teams completed three simulation rounds of 40 decision trials each. After each decision trial, participants received outcome feedback in both table and graphical format on their results for that trial plus their cumulative performance. After each simulation round of 40 quarters, the simulation was reset to the same initial values and the next simulation round began. The simulated outcomes could be, and were, very different from one simulation round to the next because different decisions result in different simulated responses by competitors and customers.

4 Results

Table 1 presents the descriptive statistics for the moderate and stretch goal groups. Figure 2 illustrates the performance distributions for firms with moderate and stretch goals. Normality tests showed that the performance distribution was highly non-normal (Kolmogorov–Smirnov’s $D[39] = 0.20, p < .001$).

Table 1. Descriptive statistics

Variables	MODERATE GOAL				STRETCH GOAL			
	Mean	Median	Std Dev	Range	Mean	Median	Std Dev	Range
Performance SR1	66.02	60.2	90.05	-21.09–360.33	76.43	28.25	133.18	-32.27–583.62
Performance SR2	104.16	75.24	175.73	-27.93–840.63	140.5	26.15	236.21	-93.38–895.16
Performance SR3	112.65	120.07	113.45	-21.26–352.97	144.41	6.41	219.20	-167.52–685.98
Survival Rate	0.67	1.00	0.48	0–1	0.57	1.00	0.51	0–1

Note: N=50 for Simulation rounds 1 and 2 (25 in Moderate goal and 25 in Stretch goal); N=39 in Simulation round 3 (18 in Moderate goal and 21 in Stretch goal)

We used the Levene test to assess the equality of variances between the stretch and moderate goal groups. The Levene test does not require normality of the underlying data. Consistent with Hypothesis 1, there is a significant difference in the variance of performance outcomes between the two goal conditions. Firms with stretch profit growth goals exhibited significantly higher performance variance than firms with moderate profit growth goals ($L[1, 37] = 8.29, p < .01$). Stretch goals result in both higher and lower performance outcomes than do moderate profit growth goals.

We used the Mann-Whitney nonparametric test to assess whether the performance outcomes of firms with stretch goals were significantly different from the performance outcomes of firms with moderate goals. As shown in Table 1, firms with stretch goals achieved median cumulative profits of \$6.41 million and firms with moderate goals achieved median cumulative profits of \$120.07 million on simulation round 3. The difference in cumulative profits between the two group is not significant (Mann-Whitney’s $U=180.00, z=-0.25, p=0.81$). These results do not support Hypothesis 2a and 2b. Stretch organization profit growth goals do not result in higher or lower performance outcomes.

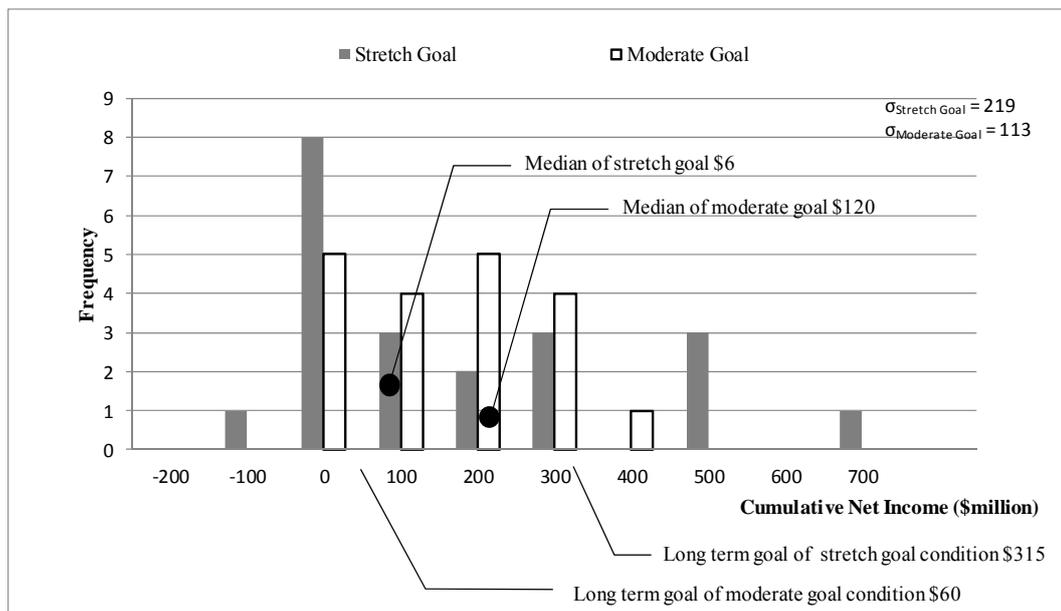


Figure 2 Performance distribution at the end of Year 10 for Stretch and Moderate Goal Condition

We used Chi-square test to assess whether the bankruptcy rate of firms with stretch goals were significantly higher than the bankruptcy rate of firms with moderate goals. As shown in Table 1, 43% of the firms with stretch goals went bankrupt and 53% of the firms with moderate goals went bankrupt. The difference in bankruptcy rate between two groups is not significant ($\chi^2_{0.05}(1) = 0.78$). These results do not support Hypothesis 3. Stretch organizational goals do not result in higher bankruptcy rate.

5 Supplementary Analysis

In order to better understand the decision making process, we recruited an additional 59 undergraduates from a big university to collect data on the information cues used in making quarterly decisions for People Express. 29 of them were assigned stretch goals and 30 of them were assigned moderate goals. After completing 3 simulation rounds on People Express, participants completed an open-ended questionnaire identifying the information cues (pieces of information) they used in making each of the five decisions (fare, aircraft orders, hiring, service of scope and marketing fraction). The question for each decision variable was: what pieces of information did you use to make your decisions (e.g., fare, aircraft orders, hiring, service of scope and marketing fraction)? We counted how the frequency each information cue was mentioned for each decision variable.

The results of the post-hoc survey shows the information cues decision makers used most frequently. By linking these information cues and observations of the decision making process/patterns from participants, the simplified mental model that decision makers were using in the simulation was identified. This simplified mental model in a feedback loop structure (shown in grey lines in Figure 3), contains 3 balancing loops and 2 reinforcing loops.

The first loop is “Growing to Achieve the Goal”. In line with prior research in goal seeking dynamics (Barlas & Yasarcan, 2006), decision maker perceive the discrepancy gap coming from the difference between the goal and actual performance and this discrepancy gap generates pressure to close the gap. A typical reaction to increased pressure is to grow the business. Buying more aircraft to expand routes and the potential markets increases the number of passengers who choose to fly People Express as customers. As the number of customer increases, revenue goes up, and profits increase. The expectation is that this growth will reduce discrepancy gap, closing the balancing loop “1. Growing to Achieve the Goal”.

When the number of planes and routes increase, this increases the available passenger miles. This leads to an increase in projected load factor and then decision makers could place more aircraft orders, and this closes the reinforcing loop “2. Fulfill Load Factor”. In addition, after some delay as customers of PE increase, this in turn increases passenger miles and the projected load factor. And then followed by a similar process mentioned above from the projected load factor to PE customers, it closes a reinforcing loop “3. Passenger Adoption”.

Decision makers learn the potential troubling side effect of expanding capacity is the loss of service quality (i.e., they used service quality as an important information cue for aircraft order decision). When passenger miles increase, the firm requires more service capacity to maintain its service quality. Such requirement can enlarge the service capacity gap. A service capacity gap is the difference between required service capacity and actual service capacity. Once the service capacity gap increases, the service quality diminishes. This, in turn, has a negative impact on the attractiveness of PE flights, leads to a decrease of adoption rate (i.e., passenger win-lose rate) and closes a balancing loop “4. Required Service Capacity & Service Quality”.

Service quality problems due to firm growth need to be fixed through an enlargement of service capacity. That is also why decision makers used service quality

as one of the important information cues in their hiring decision. When buying aircraft, many decision makers have a number of desired employees per plane in mind. Hiring increases the number of staff and enhances service capacity by alleviating the service capacity gap, and in turn improves service quality. This closes another balancing loop “5. New Staff on Increasing Service Capacity”.

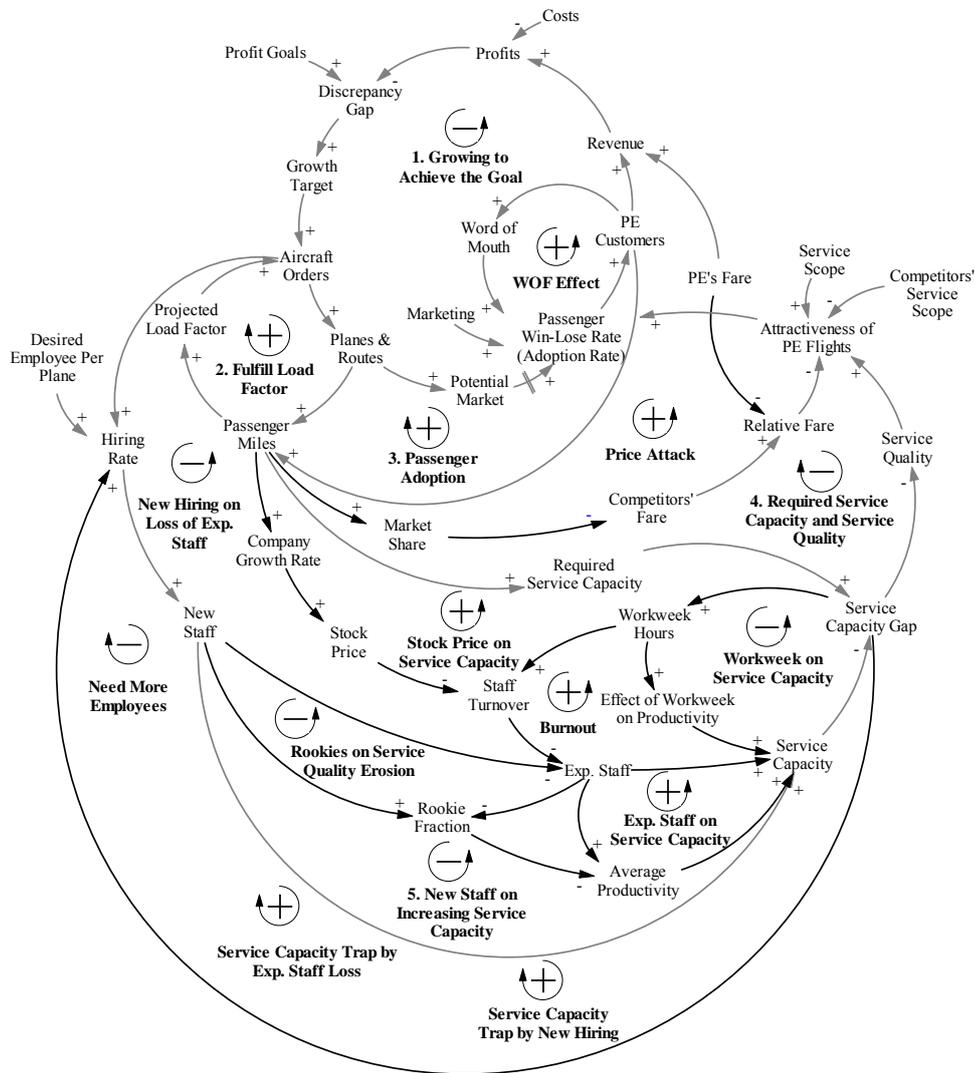


Figure 3 Feedback loops of simplified mental models and People Express

In addition to the five feedback loops mentioned above, there are some simple causal relationships that decision makers bear in mind when making decisions. For example, to promote attractiveness of PE flights, decision makers would modify PE's fare compared with competitors' fare or modify service scope compared with competitors' service scope. To improve revenue, decision makers would charge a high fare (i.e., dollars per seat mile) to their customers.

The feedback loops and simple causal relationships mentioned above show the simplified mental model that decision makers have when they make quarterly decisions.

The organizational learning literature suggests that failure to achieve the goal increases search for discovering new strategies to improve performance (Sitkin, 1992). In a novel task, decision makers lack a base of cause-and-effect knowledge and experimentation generates information that cannot be acquired by other ways (Mcgrath, 2001). Search is expected to lead to greater knowledge, resulting in the right strategic choices and superior performance. Hence, this line of research implies that decision

makers assigned stretch goals should have better mental models from the search and learning process than those assigned moderate goals.

We examine the top three cues for each decision by goal condition (moderate goal vs. stretch goal). Surprisingly, the results show that decision makers in both conditions used similar information cues in aircraft acquisition, hiring, marketing fraction and service scope decisions. Among the top three most frequently used cues, 11 out of 15 (73.3%) of the cues are the same in both goal conditions. This suggests decision makers in both goal conditions share similar mental models.

Beyond the feedback loops identified for the simplified mental models, several additional feedback loops have been identified for People Express flight simulator (Graham et al., 1992; Senge, 1990; Sterman, 1988). These additional feedback loops play an important role in determining performance. The black lines in Figure 3 show these additional feedback loops. These feedback loops are neglected in participants' decision making process. This includes "Word of Mouth Effect", "Price Attack" – competitors launch price war due to a loss of market share, "Workweek on Service Capacity" – trying to close the service capacity gap by increasing workweek hours, "New Hiring to Loss of Exp. Staff" – new hiring leads to a loss of experienced staff, "Rookies on Service Capacity Erosion" – new staff harms the average productivity, "Need more employees" – the shortage of service capacity leads to hiring more new employees, "Burnout" – the increase of workweek hours leads to staff turnover, "Stock Price on Service Capacity" – an increase of stock price reduces turnover rate, "Experienced Staff on Service Capacity" – an opposite effect of "Rookies on Service Capacity Erosion", "Service Capacity Trap by Experienced Staff Loss", and "Service Capacity Trap by New Hiring".

Compared with the grey and black lines in Figure 3, there are significant differences between participants' simplified mental model and the expanded feedback structure of the simulation. First, the number of causal relationships is underestimated by decision makers. Second, the feedback effect between the company's growth and service quality are often neglected by decision makers.

These two differences have great impact on the quality of strategy and firm performance. When decision makers realize a discrepancy gap, they start to acquire aircraft and hire employees to expand firm's capacity. Their simplified mental model suggests that a simple asset growth strategy should pay off and close the performance gap. However, the quicker they expand the fleet, the faster they are trapped by the service quality problem. That is, the expansion of firm's capacity does not lead to profit growth unless they maintain the delicate balance between fleet size and service capacity. There is significant time delay and complex dynamics between capacity growth and profit growth. To close the performance gap, decision makers must identify a strategy to navigate the effects of all the feedback effects.

6 General Discussion

This paper examines the impact of different organizational goal levels on variation in performance outcomes. Hypothesis 1 is supported. Firms provided with stretch profit growth goals exhibit higher performance variance than firms provided with moderate profit growth goals. Hypothesis 2a and 2b are not supported; median firm performance is not significantly higher or lower, under stretch versus moderate profit growth goals. The results do not support Hypothesis 3. Firms with stretch goals do not have a greater bankruptcy rate relative to firms with moderate goals.

Our study is consistent with prior research in misperceptions of feedback. The results of supplementary analysis show that decision makers in both groups end up with similar mental models (i.e., the supplementary analysis shows that 73.3% of the most frequently used information cues are the same between two groups). Prior research

suggests that decision makers have great difficulty managing dynamically complex tasks and continually make systematic and costly errors (Gary & Wood, 2011; Paich & Sterman, 1993; Sterman, 1989b). Also, decision makers' understanding of the complex feedback structure is poor (Diehl & Sterman, 1995). The experimental task, the People Express simulator, like other complex social systems, is dramatically complex. In such situations learning is difficult and inefficient. Decision makers lack a systematic strategy to explore the problem space. This makes the process of accumulating knowledge more difficult.

7 Conclusions

Organizations frequently set higher goals for profit and growth, rather than being satisfied with having met previous targets, in order to motivate managers to achieve these goals (Lant, 1992). Over the last decade, senior managers of publicly listed companies have increasingly embraced the practice of announcing ambitious financial performance goals for the next year or years. These stretch performance goals often take the form of compound annual or quarterly growth in earnings, stock price, revenue, market share, or some other performance metric. This continuous escalation of performance goals has been reinforced by pressure from investment banking analysts who urge companies to reach for ever higher growth targets every quarter.

However, there is increasing concern that such ambitious performance goals ultimately lead management to adopt actions and strategies that damage organizations' long-term health (Fuller & Jensen, 2010). The results presented here raise a warning that stretch profit growth goals may lead to increasing variance in financial performance without increasing the expected value of performance outcomes.

References

- Bakken, B., J. Gould, and D. Kim, 1992. "Experimentation in learning organizations: A management flight simulator approach." *European Journal of Operational Research*, 59(1), 167–182.
- Barlas, Y., H. Yasarcan, 2006. "Goal setting, evaluation, learning and revision: A dynamic modeling approach." *Evaluation and Program Planning*, 29(1), 79–87.
- Bromiley, P., 1991. "Testing a causal model of corporate risk taking and performance." *Academy of Management Journal*, 33(3), 520–533.
- Bromiley, P., K.M. Miller, and D. Rau, 2001. "Risk in strategic management research". *The Blackwell handbook of strategic management*, 259–288.
- Collins, J., J.I. Porras, 2002. *Built to Last: Successful Habits of Visionary Companies*. *Successful Habits of Visionary Companies*. HarperBusiness.
- Cyert, R.M., J.G. March, 1963. *A Behavioral Theory of The Firm*, 2nd edition, Englewood Cliffs, NJ: Prentice-Hall.
- Denrell, J., J.G. March, 2001. "Adaptation as information restriction: The hot stove effect." *Organization Science*, 12(5), 523–538.
- Diehl, E., J.D. Sterman, 1995. "Effects of feedback complexity on dynamic decision making." *Organizational Behavior and Human Decision Processes*, 62(2), 198–215.
- Fiegenbaum, A., S. Hart, and D. Schendel, 1996. "Strategic Reference Point Theory." *Strategic Management Journal*, 17(3), 219–235.
- Fuller, J., M.C. Jensen, 2010. "Just say no to Wall Street: Putting a stop to the earning game." *Journal of Applied Corporate Finance*, 22(1), 59–63.
- Gary, M.S., R.E. Wood, 2011. "Mental models, decision rules, and performance heterogeneity." *Strategic Management Journal*, 32, 569–594.
- Graham, A.K., J.D.W. Morecroft, P.M. Senge, and J.D. Sterman, 1992. "Model-supported case studies for management education." *European Journal of Operational Research*, 59(1), 151–166.
- Greve, H.R., 1998. "Performance, aspirations, and risky organizational change." *Administrative Science Quarterly*, 43(1), 58–86.

- Greve, H.R., 2008. "A behavioral theory of firm growth: Sequential attention to size and performance goals." *Academy of Management Journal*, 51(3), 476–494.
- Lant, T.K., 1992. "Aspiration level adaptation: An empirical exploration." *Management Science*, 38(5), 623–644.
- Lant, T.K., Z. Shapira., 2008. "Managerial reasoning about aspirations and expectations." *Journal of Economic Behavior & Organization*, 66, 60–73.
- Larrick, R.P., C. Heath., and G. Wu., 2009. "Goal-induced risk taking in negotiation and decision making." *Social Cognition*, 27(3), 342–364.
- Lee, D.Y., 1997. "The impact of poor performance on risk-taking attitudes: A longitudinal study with a PLS causal modeling approach." *Decision Sciences*, 28(1), 59–80.
- Locke, E.A., G.P. Latham., 2002. "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey." *American Psychologist*, 57(9), 705–717.
- Locke, E.A., K.N. Shaw., L.M. Saari., and G.P. Latham., 1981. "Goal setting and task performance: 1969-1980." *Psychological Bulletin*, 90(1), 125–152.
- March, J.G., Z. Shapira., 1987. "Managerial Perspectives on Risk and Risk Taking." *Management Science*, 33(11), 1404–1418.
- March, J.G., Z. Shapira., 1992. "Variable risk preferences and the focus of attention." *Psychological Review*, 99(1), 172–183.
- Mcgrath, R.G., 2001. "Exploratory learning, innovative capacity, and managerial oversight." *Academy of Management Journal*, 44(1), 118–131.
- Miller, K.D., W.R. Chen., 2004. "Variable organizational risk preferences: Tests of the March-Shapira model." *Academy of Management Journal*, 47(1), 105–115.
- Oxnard, T., 2004. "Stretch! How Toyota reaches for big goals." *Supply Chain Management Review*, 8(2), 28–35.
- Paich, M., J.D. Sterman., 1993. "Boom, Bust, and Failures to Learn in Experimental Markets." *Management Science*, 39(12), 1439–1458.
- Senge, P. M., 1990. *The Fifth Discipline: The Art & Practice of The Learning Organization*. Doubleday/Currency.
- Short, J., T.B. Palmer., 2003. "Organizational performance referents: An empirical examination of their content and influences." *Organizational Behavior and Human Decision Processes*, 90(2), 209–224.
- Singh, J.V., 1986. "Performance, slack, and risk taking in organizational decision making." *Academy of Management Journal*, 29(3), 562–585.
- Sitkin, S.B., 1992. "Learning through failure: The strategy of small losses." *Research in Organizational Behavior*, 14, 231–266.
- Sitkin, S.B., K.E. See., C.C. Miller., M.W. Lawless., and A.M. Carton., 2011. "The paradox of stretch goals: Organizations in pursuit of the seemingly impossible." *Academy of Management Review*, 36(3), 1–60.
- Sterman, J.D., 1988. "People express management flight simulator." *School of Management, MIT*.
- Sterman, J.D., 1989a. "Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment." *Management Science*, 35(3), 321–339.
- Sterman, J.D., 1989b. "Misperceptions of feedback in dynamic decision making." *Organizational Behavior and Human Decision Processes*, 43, 301–335.
- Wood, R.E., A.J. Mento., and E.A. Locke., 1987. "Task complexity as a moderator of goal effects: A meta-analysis." *Journal of Applied Psychology*, 72(3), 416–425.

A Pragmatic Approach to Optimization of Water Distribution Networks

Martin Young
Derceto Limited
myoung@derceto.com

Abstract

Derceto Ltd. is a New Zealand company that provides a purpose-designed software solution to optimise water production and distribution at lowest operational cost. The product called Aquadapt is an “off the shelf” product with foundational operations research design. Aquadapt is customised for each client in order to integrate with the existing SCADA systems and to implement specific constraints. Here lie the challenges of applying computational solutions to real world applications.

We have learnt through our experience of installing Aquadapt systems around the world that a pragmatic approach is required when applying operations research based software solutions. In this presentation, the complexities of practical application of operations research will be discussed with particular focus on examples that have provided challenges to implementing optimised solutions in the commercial environment.

Simulation Optimisation for Ambulance Redeployment

Lei (Oddo) Zhang
The Optima Corporation

Andrew Mason

Andy Philpott

Department of Engineering Science

University of Auckland

New Zealand

o.zhang@theoptimacorporation.com

Extended Abstract

The study of ambulance redeployment, also known as move-up or system status management, is the practice of dynamically deciding stand-by locations for free ambulances in attempt to achieve quick response times.

Traditionally, each ambulance is assigned to a pre-determined stand-by location (home base). Whenever an ambulance completes its service for a call, it returns to its home base. This type of policy is known as a static policy. More recently, emergency medical services (EMS) providers have started to employ move-up to manage their operations. However, most move-up policies used in practice are constructed on an ad-hoc basis, e.g, testing a set of randomly generated candidate policies, leading to limited performance gains and crew frustrations at what are perceived as pointless moves.

In this talk, we present two move-up models using simulation optimisation in attempt to construct high-performance move-up policies in a systematic manner.

The first move-up model constructs an optimised nested compliance-table move-up policy for a given system. A compliance table describes, for each number of free ambulances, a unique configuration, i.e., a set of stand-by locations. A nested compliance table means that if a stand-by location is used in the configuration associated with n free ambulances, it is also used in the configuration associated with $n + 1$ free ambulances.

Whenever a move-up decision is required, we solve an assignment problem to minimise total travel times in order to achieve the target configuration.

To find an optimised compliance-table move-up policy for a given system, we propose a simulation-based next-descent local search algorithm. The algorithm performs operations of `-move` and `-swap` on a list of stand-by locations to construct neighbour solutions. The quality of a given solution is measured via simulation on a common training dataset.

The second move-up model, which is formulated as a linear integer programming (IP) model, is a generalisation of the first move-up model. The key difference between the first move-up model and the second move-up model is that the former forces n free

ambulances into a unique configuration, while the latter does not. The IP model considers both the benefits of a new move-up configuration and travel costs for achieving the configuration. Parameters associated with the benefits and travel costs are tuned by a simulation-based numerical optimisation algorithm. In other words, another optimisation problem which seeks the best model parameters to maximise the long-term performance is solved. Each set of candidate model parameters is evaluated via simulation on a common training dataset.

We report experimental results using artificial data based on Auckland road networks, the local population distribution and the local ambulance base locations. The results suggest that the compliance-table move-up model and the IP move-up model produce statistically equivalent policies with respect to response times. However, the policies based on the IP move-up model are more cost-effective.

Key words: Ambulance redeployment, simulation optimisation, integer programming.

Developing a Rotation Scheme to Reduce Expiration for the Medical Reserve Supply

Quan Zhou, Tava Olsen
Department of Information Systems and Operations Management
University of Auckland
New Zealand
q.zhou@auckland.ac.nz
t.olsen@auckland.ac.nz

Abstract

The New Zealand government maintains large quantities of medical supplies, known as the “National Reserve Supply”, to protect the public in case of unexpected pandemics and emergencies. These medical supplies have a limited shelf life, but the demand for medical supplies from emergencies is highly uncertain. This leads to a serious expiration problem in the reserve supply in New Zealand. Our study recognise that an alternative to reduce expiration is to rotate the reserve to hospitals’ operational use. The approach is to use old items in the reserve in hospitals during normal days, and to replenish the reserve with new items, so we can reduce expiration in the reserve. In this paper, we propose a rotation scheme to use the old reserve items in hospitals. We consider two rotation policies: decentralised and centralised, and compare the resulting expiration and overall cost under these two policies with those under non-rotation policy. We demonstrate that the rotation scheme can effectively reduce expiration in the reserve, and also can improve the overall cost-effectiveness of the medical inventory system.

Key words: Medical reserve supply, Expiration, Stock rotation.

1 Introduction

Medical supplies are among the most critical necessities in the disaster relief process. Governments need to hold sufficient stocks to ensure continuous access to essential health-related products after emergencies. This stock is called the “National Reserve Supply” in New Zealand.

Currently, the Ministry of Health (MOH) manages this national reserve supply, and local District Health Boards (DHBs) are responsible for the storage. MOH determines a minimum stock level of the reserve based on the population in each area. According to an interview with staff from Health Benefit Limited, some critical medicines stored in the reserve are enough to cover all the population in the area

for 3 days after an emergency. This indicates the huge size of the reserve stockpile. Besides this large amount of national reserve, local DHBs have to hold their own operational stocks to fulfil the demand for regular hospital-use (Ministry of Health 2009). However, medical supplies have a limited shelf life that is usually a couple of years, while the likelihood of a large-scale public health emergency is relatively low during that time period. After years of storage in the warehouses, many medical stocks in the reserve expire before being used. Expiration in emergency medical supplies causes substantial waste in the process of disposing of and replacing expired items (Whybark 2007). Moreover, with a large portion of expiring inventories, there may not be enough effective supplies when a disaster occurs. Such a medical inventory system is vulnerable to disasters.

A useful alternative to reduce expiration is to rotate the national reserve to hospitals' operational use. That is, to transfer old items in the reserve to hospitals and consume them in hospitals. In such a way, items can be used before their expiry date, so as to avoid disposing of and replacing expired stocks (Shen, Dessouky, and Ordóñez 2010; Dhankhar et al. 2010).

In this paper, we propose a rotation policy to transfer old reserve items, and develop models to compare expiration and system costs under different rotation approaches. In such a way, we illustrate the relationship between rotation size and expiration, demonstrate how the rotation can affect expiration and overall inventory cost, and analyse the conditions for appropriate rotation approaches.

2 Problem Description

2.1 The Reserve Rotation Policy

The logic of rotation is that we can rotate old reserve items to operational stock, and at the same time, replenish the reserve with new stocks, so as to keep the minimum required stock level in the reserve. The rotation flow can be shown in Fig.1.

Essential decisions for a rotation scheme would be when and how to implement rotation, and we cannot determine them arbitrarily. There is a trade-off between reduced expiration and increased logistics costs. Each rotation will incur considerable inspection, transferring, and management cost, and we need to balance these costs with the benefits from reduced expiration. Furthermore, rotation should depend on the demand in hospitals. Specifically, if the rotation is infrequent or in a small scale, it would not be sufficient to reduce expiration; if too many items are frequently rotated, it is possible that the hospitals cannot consume such a large quantity with its limited demand, and so rotated items would finally expire in hospitals. Therefore, we need a balanced rotation policy.

As shown in Fig.1, rotation can be made with several hospitals. Here, for simplicity, we model it as one reserve and one operational stock. Actually, for deterministic models in this paper, modelling as one hospital makes no difference from modelling as several hospitals, and the demand from the single operational stock can be seen as an aggregation of demands from different hospitals. So, in the rest of this paper, we illustrate our models with one reserve and one operational stock.

We propose a rotation policy like this: each time the hospital places an order, the reserve transfers its oldest items to the operational stock, and then, the external supplier sends new items directly to the reserve, so that the reserve can maintain

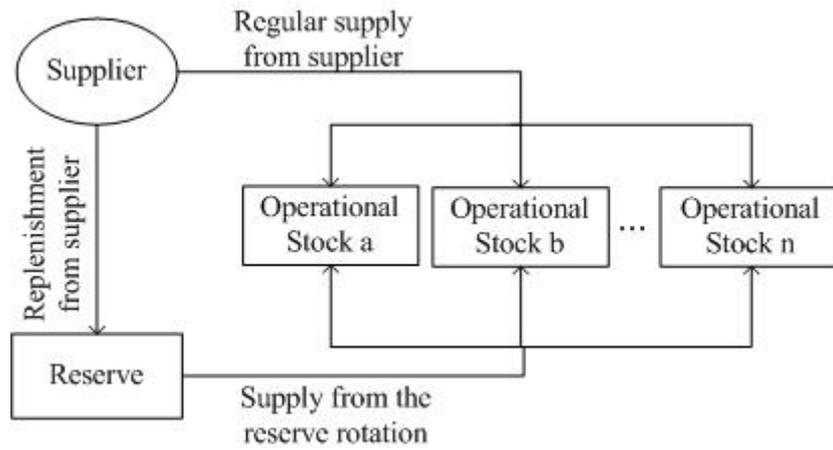


Figure 1: General rotation flow.

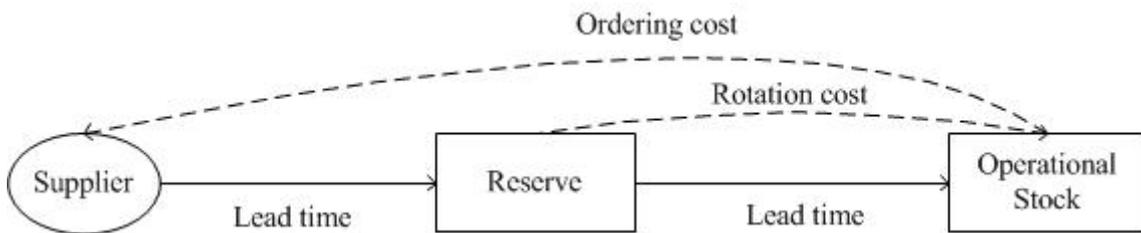


Figure 2: The process of rotation policy.

its stock level. This rotation process is illustrated in Fig.2. Further in Section 3.2, we show there are two different rotation approaches: decentralised and centralised. Under rotation, the demand for operational stock is a drive for the whole rotation system, and the order size and timing from the hospital determine the stock level and age distribution in the reserve, and then further influence expiration and the system cost. Rather, when there is no rotation, the reserve and the operational stock are operated separately: the hospital orders from external supplier, the reserve sits and expires at the end of shelf life, and the decision of hospital orders has no influence on the reserve. We refer it as the non-rotation policy, and the process is illustrated in Fig.3.

Comparing these two process flows, we can see that while rotation can increase the turnover rate and so reduce expiration in the reserve, it also incurs additional rotation and transferring cost, and so the total cost of the reserve and the operational stock may increase as a result of rotation. The goal of rotation should be to reduce expiration and the overall system cost. We will compare the rotation policy with the non-rotation policy, in order to see their influence on expiration and the system cost.

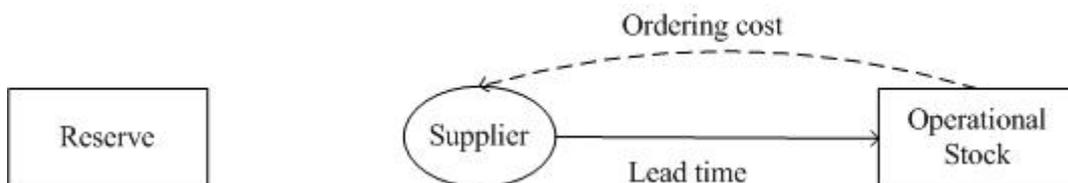


Figure 3: The process of non-rotation policy.

2.2 Assumptions and Notation

Assumptions and notation made to build models in this paper are as follows.

Assumptions

1. There is only one reserve stockpile and one operational stockpile.
2. All items in the reserve have the same fixed shelf life, and all are fresh at the beginning.
3. Hospital sources all its orders from the external supplier when there is no rotation, and gets all replenishment from the reserve under rotation.
4. The reserve issues items following FIFO (First In, First Out). That is, the oldest items are rotated first.
5. Items in the operational stock will not expire. That is, all items in the operational stock can be consumed before use-by date.
6. Under rotation, the reserve's stock level is slightly higher than the required minimum stock level, but cannot be smaller than the required level.
7. The hospital is facing patients' demand with constant rate.
8. Replenishment lead time is negligible.
9. Unit holding cost is the same for both the reserve and the operational stock.

Justification Assumption 6 is because the reserve needs to keep up to its minimum stock level as well as timely satisfying orders from the hospital. Assumption 8 is because the lead time is shorter than one review period (Nahmias 2011).

Notation

τ	Shelf life of items in the reserve, described as unit time periods (days)
P	The required minimum stock level in the reserve
μ	Unit time (per day) demand for operational stock in hospital
c_h	Unit holding cost per unit time (day)
K_1	Fixed ordering cost when ordering from external supplier
K_2	Fixed ordering cost when transferring from the reserve under rotation
p	The unit price when purchasing from external supplier
c_r	Unit replacement cost, and it makes sense that $c_r \geq p$

3 Formulation of The Model

In this section, we formulate the decision under different situations, non-rotation and decentralised and centralised rotation approaches.

3.1 Non-rotation Model

When there is no rotation, it is a traditional deterministic inventory model for the hospital's decision. The objective of the hospital is to minimise the average unit time cost of operational stock:

$$\min ETC_1 = \frac{\mu}{Q}K_1 + c_h \frac{Q}{2} + p\mu.$$

So, the optimal order quantity under non-rotation is

$$Q_1 = \sqrt{\frac{2\mu K_1}{c_h}}. \quad (1)$$

The expected length of a reorder cycle is

$$T_1 = \frac{Q_1}{\mu}. \quad (2)$$

Under non-rotation policy, all the items in the reserve expire and should be replaced at the end of shelf life τ , so a replacement cost of $c_r P$ will be incurred. Therefore, during each τ , the total system cost of the reserve and the operational stock under non-rotation is:

$$\begin{aligned} TC_1 &= ETC_1 * \tau + c_r * P \\ &= \left(\frac{\mu}{Q_1} K_1 + c_h \frac{Q_1}{2} + p\mu \right) \tau + c_r P. \end{aligned} \quad (3)$$

3.2 Rotation Models

There are two possible approaches to implement rotation. One is that the hospital makes the ordering decision according to the costs incurred to the hospital itself, and the reserve, as a supplier, has to accept the order size determined by the hospital. Under the second situation, the ordering policy is determined with consideration of the costs of both the reserve and the operational stock, like there is a centralised coordinator making the ordering decision for the hospital. We refer the first one as decentralised rotation, and the second one as the centralised rotation.

No matter which rotation approach is used, the size of rotation is determined by the orders from hospital, and ultimately determined by the actual demand from patients. So, the number of items rotated during each shelf life period τ is equal to the demand in this period, which is $\mu\tau$. Therefore, if the required minimum stock level $P < \mu\tau$, there will be no expiration in the reserve under such rotation; but if $P \geq \mu\tau$, there will still be expiration even under rotation. In practice, for some items, the required reserve size is huge and larger than the regular demand of the same period (Shen, Dessouky, and Ordóñez 2010). Moreover, $P \geq \mu\tau$ is a more general situation where expiration still exists under rotation, so the overall cost under $P \geq \mu\tau$ would be greater than that under $P < \mu\tau$. Therefore, in the following discussion, we first consider the situation where $P \geq \mu\tau$, and then we briefly present the analysis when $P < \mu\tau$.

3.2.1 Rotation When $P \geq \mu\tau$

Decentralised Rotation

Under decentralised rotation policy, the hospital determines the order quantity based on its cost, so the order size and cycle time are the same as non-rotation. Therefore, the order quantity and the expected length of a reorder cycle are respectively:

$$Q_2 = Q_1, \text{ and} \quad (4)$$

$$T_2 = T_1. \quad (5)$$

Though the hospital cost under decentralised rotation is the same as non-rotation, the overall system cost is different. Rotation generates additional logistics and holding cost. Firstly, each rotation calls for a fixed setup cost K_2 for the reserve. The average unit time cost of this fixed transferring cost is $K_2/T_2 = K_2\mu/Q_2$. Secondly, the reserve needs to hold more stocks on hand to timely complete the rotation. Even

though we assume the lead time is negligible, it just means the lead time is less than one cycle time period, there is still some gap between the time when the items are transferred out and the time when new replenishment comes in. Since the order quantity and the lead time are fixed, it would make sense to assume that the reserve needs to hold Q_2 more items, so the total stocks on hold in the reserve would be $Q_2 + P$, and the unit time holding cost of the reserve increases by $c_h Q_2$.

Also, rotation can reduce expiration in the reserve. Every T_2 , Q_2 items are rotated from the $P + Q_2$ items held in the reserve. Since the beginning items are fresh, expiration occurs after the first τ , and after that, the age distribution of the reserve will be a cyclic process which repeats itself with expiration cycle time τ . Let $n = \lceil \tau/T_2 \rceil$, $(n - 1)Q_2$ items are rotated during each τ . So the number of expired items is $P + Q_2 - (n - 1)Q_2 = P + 2Q_2 - nQ_2$, and nQ_2 can be approximated by $\mu\tau$, the demand from hospital during that time period. Therefore, the average number of expired items in each τ period is $P + 2Q_2 - \mu\tau$.

Therefore, during each τ , the total system cost of the reserve and the operational stock under decentralised rotation is:

$$TC_2 = \left(\frac{\mu}{Q_2} K_1 + c_h \frac{Q_2}{2} + p\mu + c_h Q_2 + \frac{\mu}{Q_2} K_2 \right) \tau + c_r (P + 2Q_2 - \mu\tau). \quad (6)$$

Comparing the cost of decentralised rotation and that of non-rotation, we show that under certain condition, decentralised rotation is more cost-effective than non-rotation. Let $m = K_2/K_1$, we give the condition Proposition 3.1.

Proposition 3.1. *Compared to non-rotation, decentralised rotation policy can always reduce expiration, but the overall cost-effectiveness of decentralised rotation policy is dependent on m and c_r .*

When Condition (7) holds, the overall cost of decentralised rotation policy is less than non-rotation policy:

$$\frac{c_r}{m/2 + 1} \geq \frac{c_h \tau}{\mu\tau/Q_1 - 2}. \quad (7)$$

Proof. It is clear that $P + 2Q_2 - \mu\tau < P$, because $2Q_2 < nQ_2 \leq \mu\tau$. So, the decentralised rotation policy leads to less expiration than the non-rotation policy. For the overall cost, from Equation (3), (4), and (6), we can get the difference between the costs of decentralised rotation and non-rotation:

$$\begin{aligned} TC_2 - TC_1 &= \left(c_h Q_1 + \frac{\mu}{Q_1} K_2 \right) \tau + c_r (2Q_2 - \mu\tau) \\ &= \left(c_h Q_1 + \frac{\mu}{Q_1} K_1 * \frac{K_2}{K_1} \right) \tau + c_r (2Q_1 - \mu\tau) \\ &= \sqrt{2\mu c_h K_1} \tau \left(\frac{m}{2} + 1 \right) + c_r (2Q_1 - \mu\tau). \end{aligned} \quad (8)$$

If decentralised rotation leads to overall cost reduction, it means $TC_2 - TC_1 \leq 0$, so we have:

$$\begin{aligned} TC_2 - TC_1 \leq 0 &\Leftrightarrow c_r (\mu\tau - 2Q_1) \geq \sqrt{2\mu c_h K_1} \tau \left(\frac{m}{2} + 1 \right) \\ &\Leftrightarrow \frac{c_r}{m/2 + 1} \geq \frac{ETC_1 \tau}{\mu\tau - 2Q_1} \\ &\Leftrightarrow \frac{c_r}{m/2 + 1} \geq \frac{c_h Q_1 \tau}{\mu\tau - 2Q_1} \\ &\Leftrightarrow \frac{c_r}{m/2 + 1} \geq \frac{c_h \tau}{\mu\tau/Q_1 - 2}. \end{aligned}$$

This completes the proof of Proposition 3.1. \square

Centralised Rotation

Under the centralised rotation policy, the ordering decision is made based on the cost of the whole system, rather than the hospital alone deciding the order size. Similar to decentralised rotation, let Q be the order size from the hospital, the reserve will hold Q more items to ensure timely rotation, and also, the average quantity of expiration under rotation is $P + 2Q - \mu\tau$. In this case, the objective is to minimise the average unit time cost of the reserve and the hospital, which includes the ordering and logistics cost, and also expiration cost in the reserve. Therefore, the average unit time cost of both the reserve and the operational stock is:

$$\min ETC_3 = \frac{\mu}{Q} * (K_1 + K_2) + c_h \left(\frac{Q}{2} + Q \right) + p\mu + \frac{c_r}{\tau} (P + 2Q - \mu\tau).$$

So, the order quantity under centralised rotation is

$$Q_3 = \sqrt{\frac{2\mu(K_1 + K_2)}{3c_h + 4c_r/\tau}}. \quad (9)$$

The expected length of a reorder cycle is

$$T_3 = \frac{Q_3}{\mu}. \quad (10)$$

Therefore, during each τ , the total system cost of the reserve and the operational stock under centralised rotation is:

$$\begin{aligned} TC_3 &= ETC_3 * \tau \\ &= \left[\frac{\mu}{Q_3} (K_1 + K_2) + c_h \frac{3Q_3}{2} + p\mu \right] \tau + c_r (P + 2Q_3 - \mu\tau). \end{aligned} \quad (11)$$

Comparing the cost of this centralised rotation policy and that of decentralised rotation policy, we present Proposition 3.2.

Proposition 3.2. *Under the centralised rotation policy, the overall cost is lower or equal to the cost of decentralised rotation; when some conditions hold as follows, expiration is lower than that of decentralised rotation.*

- (i) *When $m \leq 2$, expiration under the centralised rotation policy is always lower than that under decentralised rotation.*
- (ii) *When $m > 2$, expiration under the centralised rotation policy is lower than that under decentralised rotation, when condition (12) holds:*

$$\frac{c_r}{m-2} > \frac{c_h\tau}{4}. \quad (12)$$

- (iii) *When $\frac{c_r}{m-2} = \frac{c_h\tau}{4}$, the overall cost of the centralised rotation policy equals to that of decentralised rotation; otherwise, the overall cost of centralised rotation policy B is lower than that of decentralised rotation.*

Proof. (a) For expiration, from Equation (1), (4) and (9),

$$\begin{aligned}
Q_3^2 - Q_2^2 &= \frac{2\mu(K_1 + K_2)}{3c_h + 4c_r/\tau} - \frac{2\mu K_1}{c_h} \\
&= \frac{2\mu(m+1)K_1}{3c_h + 4c_r/\tau} - \frac{2\mu K_1}{c_h} \\
&= \frac{2\mu K_1}{c_h(3c_h + 4c_r/\tau)} [(m-2)c_h - 4c_r/\tau]. \tag{13}
\end{aligned}$$

When $m \leq 2$, $(m-2)c_h \leq 0$, also $4c_r/\tau > 0$, so $Q_3^2 - Q_2^2 < 0$. That means $Q_3 - Q_2 < 0$, and so $P + 2Q_3 - \mu\tau - < P + 2Q_2 - \mu\tau$. So, when $m \leq 2$, centralised policy reduces expiration.

(b) If $m > 2$, and we want $Q_3 - Q_2 < 0$, from Equation (13),

$$\begin{aligned}
Q_3 - Q_2 < 0 &\Leftrightarrow (m-2)c_h - 4c_r/\tau < 0 \\
&\Leftrightarrow \frac{c_r}{m-2} > \frac{c_h\tau}{4}.
\end{aligned}$$

(c) For the overall cost, note that the total cost function of decentralised rotation policy and centralised rotation policy, Equation (6) and (11), have the same structure which can be generated as $TC = [\frac{\mu}{Q}(K_1 + K_2) + c_h\frac{3Q}{2} + p\mu]\tau + c_r(P + 2Q - \mu\tau)$. Since Q_3 is the only optimal solution for this cost function, it always holds $TC_3 \leq TC_2$, and the equal sign satisfies when $c_r/(m-2) = c_h\tau/4$.

This completes the proof of Proposition 3.2. \square

3.2.2 Rotation When $P < \mu\tau$

When $P < \mu\tau$, there will be no expiration under rotation, making it more beneficial to undertake rotation if the conditions discussed in Section 3.2.1 hold. We can rewrite the total cost function of decentralised and centralised rotation as follows.

The total cost of decentralised rotation is:

$$TC'_2 = \left(\frac{\mu}{Q_2}K_1 + c_h\frac{Q_2}{2} + p\mu + c_hQ_2 + \frac{\mu}{Q_2}K_2 \right) \tau. \tag{14}$$

The total cost of centralised rotation is:

$$TC'_3 = \left[\frac{\mu}{Q'_3}(K_1 + K_2) + c_h\frac{3Q'_3}{2} + p\mu \right] \tau. \tag{15}$$

with

$$Q'_3 = \sqrt{\frac{2\mu(K_1 + K_2)}{3c_h}}. \tag{16}$$

Because expiration is eliminated when $P < \mu\tau$, the conditions favourable for rotation can be relaxed. So, we present Proposition 3.3.

Proposition 3.3. *If $P < \mu\tau$, both centralised and decentralised rotation can eliminate expiration in the reserve; centralised rotation is always more cost effective than decentralised rotation; centralised rotation can reduce overall system cost, compared to non-rotation, when it satisfies condition (17):*

$$\frac{c_r}{\sqrt{3(m+1)} - 1} \geq \frac{c_h\tau}{P/Q_1} \tag{17}$$

Proof. From Equation (14) and (15), it is clear that there is no expiration when using either decentralised or centralised rotation. These two total cost functions have the same structure $TC = [\frac{\mu}{Q}(K_1 + K_2) + c_h \frac{3Q}{2} + p\mu]\tau$. Since Q'_3 gives the only optimal solution for this cost function, it always holds $TC'_3 \leq TC'_2$, and the equal sign satisfies when $m = 2$.

From Equation (3) and (15), we can get the difference between centralised rotation and non-rotation:

$$\begin{aligned} TC'_3 - TC_1 &= \left[\frac{\mu}{Q'_3}(K_1 + K_2) + c_h \frac{3Q'_3}{2} - \frac{\mu}{Q_1}K_1 - c_h \frac{Q_1}{2} \right] \tau - c_r P \\ &= \left[\sqrt{6\mu c_h (K_1 + K_2)} - \sqrt{2\mu c_h K_1} \right] \tau - c_r P \\ &= \sqrt{2\mu c_h K_1} \left(\sqrt{3(m+1)} - 1 \right) \tau - c_r P. \end{aligned} \quad (18)$$

If the centralised rotation cost is less than non-rotation, it means $TC_2 - TC_1 \leq 0$, so we have:

$$\begin{aligned} TC'_3 - TC_1 \leq 0 &\Leftrightarrow c_r P \geq \sqrt{2\mu c_h K_1} \tau \left(\sqrt{3(m+1)} - 1 \right) \\ &\Leftrightarrow \frac{c_r}{\sqrt{3(m+1)} - 1} \geq \frac{c_h \tau}{P/Q_1}. \end{aligned}$$

This completes the proof of Proposition 3.3. □

3.3 Discussion

When developing rotation schemes for the medical reserve supply, we need to balance the benefit of reduced expiration with the costs of increased logistics. Proposition 3.1, 3.2 and 3.3 give the conditions favourable for different approaches.

Given the demand and cost structure in the hospital, the effectiveness of the rotation policy is dependent on the unit replacement cost c_r and the ratio of the fixed ordering cost m . Rotation will be more cost beneficial when replacement cost is big or the cost of rotation is small. Besides, high replacement cost can make centralised rotation more cost effective than decentralised rotation. This makes sense for practical situations. When it costs a lot to replace expired items, huge waste generates the need to reduce expiration through rotation, and to collaborate, that is, to use centralised rotation; but when the replacement cost is small, it may not be worthwhile to make the effort. When the cost of implementing rotation is low, it is appealing to do the rotation; but when the cost of rotation is high, the rotation cost may be greater than the savings from reduced expiration.

Though non-rotation may generate less cost than rotation in theory, rotation is superior to non-rotation in most practical situations. In practice, the minimum stock level of the reserve is usually big, the shelf life of the reserve supply is long, and the holding cost of medical supply is relatively small. These factors mean that the values of the right side of the conditions (7), (12) and (17) are usually very small numbers, so it means that under most cases in practice, these conditions will hold. Therefore, in most practical cases, rotation is more cost effective than non-rotation, and that centralised rotation is more cost effective than decentralised rotation.

4 Conclusion and Future research

In this paper, we propose a rotation policy to alleviate the serious expiration problem in the medical reserve supply, and build models to compare the effectiveness of non-rotation, decentralised rotation, and centralised rotation policy. When the required quantity in the reserve is huge, rotation may not be able to completely eliminate expiration, but it can still effectively reduce expiration and improve the overall performance of the system. Through the model analysis, we present the conditions when it is beneficial to implement rotation. We found that rotation is more attractive when replacement cost is big and rotation cost is small. Further, under most practical situations, rotation can lead to reduced expiration and overall cost compared to non-rotation, and centralised rotation can lead to more reduction in expiration and overall cost than decentralised rotation.

Future work will focus on rotation policy structure and collaboration contracts. We give a predefined rotation policy in this paper, assuming hospitals get all replenishment from the reserve under rotation. This is not realistic because hospital can only get aged items. It is desirable to develop a more realistic rotation structure. Besides, though centralised rotation can be cost effective for the whole system, it increases the hospital's cost, so it is possible that the hospital is not willing to do the rotation. We need to find appropriate incentives for the hospital to participate in this centralised decision.

Acknowledgments

I would like to thank staff from Health Benefit Limited, Wellington hospital, and Capital & Coast DHB. The interviews and talks I made with staff from these organisations enable me to better understand the problem and build the model. I would also like to thank The Chartered Institute of Logistics and Transport (CILT) in New Zealand for providing me the Transport Research and Educational Trust Board Scholarship to undertake this research.

References

- Dhankhar, P., J.D. Grabenstein, M.A. O'Brien, and E.J. Dasbach. 2010. "Cost-effectiveness of stockpiling 23-valent pneumococcal polysaccharide vaccine to prevent secondary pneumococcal infections among a high-risk population in the united states during an influenza pandemic." *Clinical therapeutics* 32 (8): 1501–1516.
- Ministry of Health. 2009. *National Health Emergency Plan: National Reserve Supplies Management and Usage Policies*. 2nd Edition. Wellington: New Zealand Ministry of Health.
- Nahmias, S. 2011. *Perishable inventory systems*. Springer.
- Shen, Z., M. Dessouky, and F. Ordonez. 2010. "Perishable inventory management system with a minimum volume constraint." *Journal of the Operational Research Society* 62 (12): 2063–2082.
- Whybark, D. 2007. "Issues in managing disaster relief inventories." *International Journal of Production Economics* 108 (1): 228–235.

Accumulating Priority Queues: A New Priority Scheme for Hospital Queues?

Ilze Ziedins
Department of Statistics
The University of Auckland
New Zealand
i.ziedins@auckland.ac.nz

Abstract

In traditional priority queues, arrivals are assigned a fixed priority and served in strict order of their initial assigned priority. This can lead to unacceptably long waiting times for low priority customers. This talk will discuss a priority scheme where customers are assigned to a priority class, but their priority increases linearly with their waiting time in the queue. The higher the priority class of the customer, the faster their priority increases. Under this scheme a customer from a low priority class may be served before a customer from a higher priority class if they have been waiting sufficiently long. Such a priority scheme has particular relevance in health-care settings, where a patient's condition may become more acute while awaiting treatment, and this motivated the current study.

This is joint work with David Stanford (Western Ontario) and Peter Taylor (Melbourne).

Key words: Priority queues, Time-dependent priorities, Hospital waiting lists.
