



**Operations Research Society
of New Zealand**

**Proceedings of the
44th Annual Conference**

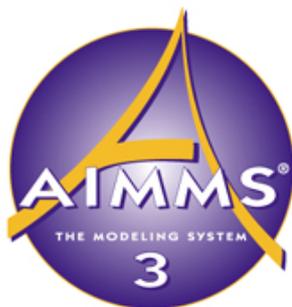
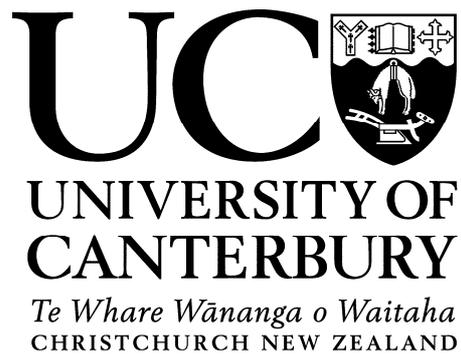
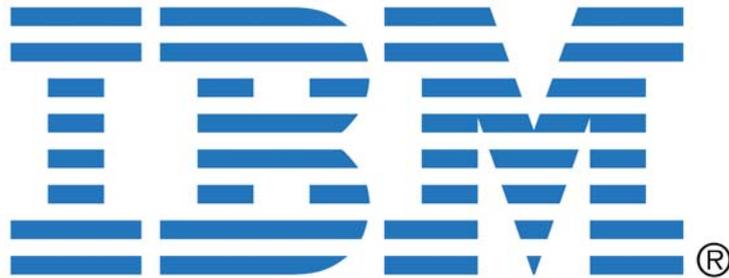
ORSNZ09

December 3 – 4th 2009

**University of Canterbury
Christchurch
New Zealand**

SPONSORS

We are most grateful for the financial support we have received from IBM ILOG, University of Canterbury, Paragon Decision Technology and Hoare Research Software.



Preface

The papers in this volume form the Proceedings of *ORSNZ09*, the 44th Annual Conference of the Operational Research Society of New Zealand (ORSNZ), held December 3 to 4 at the University of Canterbury, Christchurch New Zealand.

As chair of the Conference committee it is a pleasure to see such a strong collection of papers and, especially, such a large group of students attending and presenting papers.

The Conference Committee would like to thank the sponsors IBM ILOG, Univeristy of Canterbury, Paragon Decision Technology, and Hoare Research Software.

The conference could not have been possible without the invaluable assistance of the Conference Committee:

Shane Dye (Chair)
Pavel Catska
John George (YPP Coordinator)
John Giffin
Terri Green
Ross James
Don McNickle
Nicola Petty
John (Fritz) Raffensperger
E. Grant Read
Pulakanam Venkateswarlu

Shane Dye

Table of Contents

Thursday

Thursday Keynote Presentation (ThK): (C2 9:35–10:20) Graeme Everett, Rune Gjessing, Kjetil Vatn, Andy Philpott Norske Skog Improves Global Profitability Using O.R.	1
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---

Session Th1: Young Practitioners Prize Session #1 (C2 10:50-12:30)

Faster Ambulance Simulation Runs using Artificial Call Sampling <i>Eduard Bulog</i>	2
Catch-up Scheduling for Childhood Vaccination <i>Faramroze G. Engineer</i>	12
Making Smarter Transportation Investments <i>Kane Harton</i>	22
A Rostering Integer Programming Model for Ambulance Staffing <i>Karl Ho</i>	32
Optimisation of Mould Filling Parameters during Compression Resin Transfer Moulding Process <i>Wing Ki Kam</i>	42

Session Th2: Young Practitioners Prize Session #2 (C2 13:30-15:10)

Capacity Planning For Process Industries <i>James O. Kirch</i>	52
Trim Loss and Inventory Optimisation in Paper Mills <i>Qi-Shan Lim</i>	62
Optimisation Models and Methods for the Container Positioning Problem in Port Terminals <i>A. Phillips</i>	72
An Integrated Collaboration Platform for Sustainable Development: Project Proposal and Initial Exploration <i>Max Erik Rohde</i>	81
A Simulation of the Student Health Centre at the University of Auckland <i>Kathryn J. Trevor</i>	89

Session Th3A: MIP Theory and Applications (C2 15:40-17:00)

Improve Efficiency of OR Applications with ILOG CPLEX <i>Liu Huan Dong</i>	99
Operating Theatre Optimisation <i>Oliver Weide</i>	100
MIP-based Heuristics for the Capacitated Lot-Sizing Problem with Startup Times <i>Shaorui Li, Cheng Peng</i>	101
Area restricted forest harvesting with adjacency branches <i>Alastair McNaughton, David Ryan</i>	110

Session Th3B: Planning Over Time

(C3 15:40-17:00)

Initial use of discrete event simulation for New Zealand military workforce analysis <i>Nebojsa Djorovic, Michelle Gosse, Jason Markham, Jay Ta'ala</i>	120
MIP Models for Scheduling the Operations for a Coal Loading Facility <i>Riley Clement</i>	129
Optimal Pricing Decision for a Dynamic Inventory Problem with Constant Price Elasticity of Demand <i>Chia-Shin Chung, James Flynn</i>	139
Modelling Values of Lake Ellesmere <i>John F. Raffensperger, Ken Hughey</i>	148

Friday

Friday Keynote Presentation (FrK): (C2 8:45–9:30)

Dorit S. Hochbaum

New models and methodologies for group decision making, rank aggregation, clustering and data mining.

159

Session Fr1A: Solving Hard Problems

(C2 9:30-10:50)

Approximation Algorithm for Firefighter Problem on Trees <i>Yutaka Iwaikawa, Naoyuki Kamiyama, Tomomi Matsui</i>	160
Approximation Algorithm for Multi-Dimensional Assignment Problem Arising from Multitarget Tracking <i>Yoshitaka Sugiura, Naoyuki Kamiyama, Tomomi Matsui</i>	164
Solving the Airline Crew Pairing Problem using Subsequence Generation <i>Matias Sevel Rasmussen, David M. Ryan, Richard M. Lusby, Jesper Larsen</i>	169
Customised Column Generation for Rostering Problems: Using Compile-time Customisation to create a Flexible C++ Engine for Staff Rostering <i>Andrew J Mason, David Ryan</i>	172

Session Fr1B: Markets and Water Management

(C3 9:30-10:50)

On Asset Reallocation in the New Zealand Electricity Market <i>Anthony Downward, David Young, Golbon Zakeri</i>	180
Can Markets in Agricultural Discharge Permits be Competitive? <i>R. A. Ranga Prabodanie, John F. Raffensperger</i>	182
A Proposed Smart Market for Impervious Cover Runoff under Rainfall Uncertainty <i>Antonio Pinto, John F. Raffensperger, Thomas Cochrane, Shane Dye</i>	190
Shaping More Sustainable Communities: a Case Study in Urban Water Management <i>Robyn M. Moore</i>	200

Session Fr2: Semi-Plenary

(C2 11:20-12:20)

- Supply chain based agent simulation: Towards a normative approach 209
Luciano Ferreira, Denis Borenstein
- Some thoughts on model use in OR/MS 219
Michael Pidd

Session Fr3A: Vehicle Applications

(C2 13:20-14:20)

- Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem 222
Andrea Raith, C. Van Houtte, J.Y.T. Wang, M. Ehrgott
- Optimization of a Single Ambulance Move up 225
Lei Zhang, Andrew Mason, Andy Philpott
- Trip Assignment under Energy and Environmental Constraints 227
Kenneth D. Kuhn

Session Fr3B: Performance Evaluation and Improvement

(C3 13:20-14:20)

- The Performance Evaluation of Turkey's Export to Ireland 236
A. Ulgen Ozgul, S. Sebnem Ahiska
- Improving research students' performance by two contrasting methodologies: Theory of Constraints (TOC) and Appreciative Inquiry (AI) 246
Garoon Pongsart
- Using phone logs to analyse call centre performance 257
John Paynter

Session Fr4A: Theory and Applications

(C2 14:30-15:30)

- A Multi-plan Method for Radiotherapy Treatment Design via Finite Representation of the Non-dominated Set of Multi-objective Linear Programmes 258
Matthias Ehrgott, Lizhen Shao
- A simulated annealing approach to the inventory routing problem 269
Mohammad J. Tarokh, Nooraddin Dabiri, Mehdi Alinaghian
- Discovering Relationships between Scheduling Problem Structure and Heuristic Performance 270
Kate A. Smith-Miles, Ross J. W. James, John W. Giffin, Yiqing Tu

Session Fr4B: OR in Education

(C3 14:30-15:30)

- Educating the world about OR with viewer-paced videos on Youtube 281
Nicola Ward Petty
- Eating the Elephant! 282
Applying Theory of Constraints in Assurance of Learning
Victoria J. Mabin
- Standardising spreadsheet LP: Do textbooks make learning LP easier? 292
Shane Dye, Nicola Ward Petty

Author Index

Author	Title	Page	Session
S. Sebnem Ahiska	The Performance Evaluation of Turkey's Export to Ireland	236	Fr3B (C3)
Mehdi Alinaghian	A simulated annealing approach to the inventory routing problem	269	Fr4A (C2)
Denis Borenstein	Supply chain based agent simulation: Towards a normative approach	209	Fr2 (C2)
Eduard Bulog	Faster Ambulance Simulation Runs using Artificial Call Sampling	2	Th1 (C2)
Chia-Shin Chung	Optimal Pricing Decision for a Dynamic Inventory Problem with Constant Price Elasticity of Demand	139	Th3B (C3)
Riley Clement	MIP Models for Scheduling the Operations for a Coal Loading Facility	129	Th3B (C3)
Thomas Cochrane	A Proposed Smart Market for Impervious Cover Runoff under Rainfall Uncertainty	190	Fr1B (C3)
Nooraddin Dabiri	A simulated annealing approach to the inventory routing problem	269	Fr4A (C2)
Nebojsa Djorovic	Initial use of discrete event simulation for New Zealand military workforce analysis	120	Th3B (C3)
Liu Huan Dong	Improve Efficiency of OR Applications with ILOG CPLEX	99	Th3A (C2)
Anthony Downward	On Asset Reallocation in the New Zealand Electricity Market	180	Fr1B (C3)
Shane Dye	A Proposed Smart Market for Impervious Cover Runoff under Rainfall Uncertainty	190	Fr1B (C3)
Shane Dye	Standardising spreadsheet LP: Do textbooks make learning LP easier?	292	Fr4B (C3)
Matthias Ehrgott	A Multi-plan Method for Radiotherapy Treatment Design via Finite Representation of the Non-dominated Set of Multi-objective Linear Programmes	258	Fr4A (C2)
Matthias Ehrgott	Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem	222	Fr3A (C2)
Faramroze G. Engineer	Catch-up Scheduling for Childhood Vaccination	12	Th1 (C2)
Graeme Everett	Norske Skog Improves Global Profitability Using O.R.	1	ThK
Luciano Ferreira	Supply chain based agent simulation: Towards a normative approach	209	Fr2 (C2)
James Flynn	Optimal Pricing Decision for a Dynamic Inventory Problem with Constant Price Elasticity of Demand	139	Th3B (C3)
John W. Giffin	Discovering Relationships between Scheduling Problem Structure and Heuristic Performance	270	Fr4A (C2)
Rune Gjessing	Norske Skog Improves Global Profitability Using O.R.	1	ThK
Michelle Gosse	Initial use of discrete event simulation for New Zealand military workforce analysis	120	Th3B (C3)
Kane Harton	Making Smarter Transportation Investments	22	Th1 (C2)
Karl Ho	A Rostering Integer Programming Model for Ambulance Staffing	32	Th1 (C2)
Dorit S. Hochbaum	New models and methodologies for group decision making, rank aggregation, clustering and data mining.	159	FrK
Ken Hughey	Modelling Values of Lake Ellesmere	148	Th3B (C3)
Yutaka Iwaikawa	Approximation Algorithm for Firefighter Problem on Trees	160	Fr1A (C2)

Author	Title	Page	Session
Ross J. W. James	Discovering Relationships between Scheduling Problem Structure and Heuristic Performance	270	Fr4A (C2)
Wing Ki Kam	Optimisation of Mould Filling Parameters during Compression Resin Transfer Moulding Process	42	Th1 (C2)
Naoyuki Kamiyama	Approximation Algorithm for Firefighter Problem on Trees	160	Fr1A (C2)
Naoyuki Kamiyama	Approximation Algorithm for Multi-Dimensional Assignment Problem Arising from Multitarget Tracking	164	Fr1A (C2)
James O. Kirch	Capacity Planning For Process Industries	52	Th2 (C2)
Kenneth D. Kuhn	Trip Assignment under Energy and Environmental Constraints	227	Fr3A (C2)
Jesper Larsen	Solving the Airline Crew Pairing Problem using Subsequence Generation	169	Fr1A (C2)
Shaorui Li	MIP-based Heuristics for the Capacitated Lot-Sizing Problem with Startup Times	101	Th3A (C2)
Qi-Shan Lim	Trim Loss and Inventory Optimisation in Paper Mills	62	Th2 (C2)
Richard M. Lusby	Solving the Airline Crew Pairing Problem using Subsequence Generation	169	Fr1A (C2)
Victoria J. Mabin	Eating the Elephant! Applying Theory of Constraints in Assurance of Learning	282	Fr4B (C3)
Jason Markham	Initial use of discrete event simulation for New Zealand military workforce analysis	120	Th3B (C3)
Andrew J Mason	Customised Column Generation for Rostering Problems: Using Compile- time Customisation to create a Flexible C++ Engine for Staff Rostering	172	Fr1A (C2)
Andrew J Mason	Optimization of a Single Ambulance Move up	225	Fr3A (C2)
Tomomi Matsui	Approximation Algorithm for Firefighter Problem on Trees	160	Fr1A (C2)
Tomomi Matsui	Approximation Algorithm for Multi-Dimensional Assignment Problem Arising from Multitarget Tracking	164	Fr1A (C2)
Alastair McNaughton	Area restricted forest harvesting with adjacency branches	110	Th3A (C2)
Robyn M. Moore	Shaping More Sustainable Communities: a Case Study in Urban Water Management	200	Fr1B (C3)
A. Ulgen Ozgul	The Performance Evaluation of Turkey's Export to Ireland	236	Fr3B (C3)
John Paynter	Using phone logs to analyse call centre performance	257	Fr3B (C3)
Cheng Peng	MIP-based Heuristics for the Capacitated Lot-Sizing Problem with Startup Times	101	Th3A (C2)
Nicola Ward Petty	Educating the world about OR with viewer-paced videos on Youtube	281	Fr4B (C3)
Nicola Ward Petty	Standardising spreadsheet LP: Do textbooks make learning LP easier?	292	Fr4B (C3)
A. Phillips	Optimisation Models and Methods for the Container Positioning Problem in Port Terminals	72	Th2 (C2)
Andy Philpott	Norske Skog Improves Global Profitability Using O.R.	1	ThK
Andy Philpott	Optimization of a Single Ambulance Move up	225	Fr3A (C2)
Michael Pidd	Some thoughts on model use in OR/MS	219	Fr2 (C2)
Antonio Pinto	A Proposed Smart Market for Impervious Cover Runoff under Rainfall Uncertainty	190	Fr1B (C3)

Author	Title	Page	Session
Garoon Pongsart	Improving research students' performance by two contrasting methodologies: Theory of Constraints (TOC) and Appreciative Inquiry (AI)	246	Fr3B (C3)
R. A. Ranga Prabodanie	Can Markets in Agricultural Discharge Permits be Competitive?	182	Fr1B (C3)
John F. Raffensperger	A Proposed Smart Market for Impervious Cover Runoff under Rainfall Uncertainty	190	Fr1B (C3)
John F. Raffensperger	Can Markets in Agricultural Discharge Permits be Competitive?	182	Fr1B (C3)
John F. Raffensperger	Modelling Values of Lake Ellesmere	148	Th3B (C3)
Andrea Raith	Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem	222	Fr3A (C2)
Matias Sevel Rasmussen	Solving the Airline Crew Pairing Problem using Subsequence Generation	169	Fr1A (C2)
Max Erik Rohde	An Integrated Collaboration Platform for Sustainable Development: Project Proposal and Initial Exploration	81	Th2 (C2)
David Ryan	Area restricted forest harvesting with adjacency branches	110	Th3A (C2)
David Ryan	Customised Column Generation for Rostering Problems: Using Compile- time Customisation to create a Flexible C++ Engine for Staff Rostering	172	Fr1A (C2)
David M. Ryan	Solving the Airline Crew Pairing Problem using Subsequence Generation	169	Fr1A (C2)
Lizhen Shao	A Multi-plan Method for Radiotherapy Treatment Design via Finite Representation of the Non-dominated Set of Multi-objective Linear Programmes	258	Fr4A (C2)
Kate A. Smith-Miles	Discovering Relationships between Scheduling Problem Structure and Heuristic Performance	270	Fr4A (C2)
Yoshitaka Sugiura	Approximation Algorithm for Multi-Dimensional Assignment Problem Arising from Multitarget Tracking	164	Fr1A (C2)
Jay Ta'ala	Initial use of discrete event simulation for New Zealand military workforce analysis	120	Th3B (C3)
Mohammad J. Tarokh	A simulated annealing approach to the inventory routing problem	269	Fr4A (C2)
Kathryn J. Trevor	A Simulation of the Student Health Centre at the University of Auckland	89	Th2 (C2)
Yiqing Tu	Discovering Relationships between Scheduling Problem Structure and Heuristic Performance	270	Fr4A (C2)
C. Van Houtte	Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem	222	Fr3A (C2)
Kjetil Vatn	Norske Skog Improves Global Profitability Using O.R.	1	ThK
J.Y.T. Wang	Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem	222	Fr3A (C2)
Oliver Weide	Operating Theatre Optimisation	100	Th3A (C2)
David Young	On Asset Reallocation in the New Zealand Electricity Market	180	Fr1B (C3)
Golbon Zakeri	On Asset Reallocation in the New Zealand Electricity Market	180	Fr1B (C3)
Lei Zhang	Optimization of a Single Ambulance Move up	225	Fr3A (C2)

Norske Skog Improves Global Profitability Using O.R.

Graeme Everett*

Norske Skog Tasman

graeme.everett@norskeskog.com

Rune Gjessing, Kjetil Vatn

Norske Skog

Andy Philpott*

University of Auckland

New Zealand

a.philpott@auckland.ac.nz

Abstract

Many businesses face uncertainty about demand in the face of economic recession. Global paper maker Norske Skog is familiar with this, as the company has experienced declining demand for its products due to electronic media replacing printed publications. O.R. models have become a vital part of the decision making process, helping the company to reduce costs and enabling senior managers to make difficult decisions. The suite of MIP-based decision tools at Norske Skog was developed to optimize manufacturing, distribution, and sourcing of raw materials in Australasia. After becoming a part of Norske Skog in 2000, the methodology was further developed for use in global operations. The tactical use of the models resulted in savings of \$8 million US and \$10 million US annually in Australasia and Europe respectively. In 2008 the model was used to contribute to a strategic decision to close two paper mills and a paper machine, saving \$100 million US annually.

Faster Ambulance Simulation Runs using Artificial Call Sampling

Eduard Bulog
Department of Engineering Science
University of Auckland
New Zealand
ebul014@aucklanduni.ac.nz

Abstract

SIREN Predict is an advanced simulation tool for emergency services, developed by the Optima Corporation. Reducing the runtimes of SIREN simulations has significant benefits, as optimisation routines often require many such simulations to be performed. One method of achieving faster simulations is to introduce *artificial calls*. These are a special type of emergency (111) call which require no ambulance dispatch, but are handled in such a way that the key performance measures are still obtained.

This project involved editing the SIREN Predict source code to establish the logic for handling artificial calls. This was complicated by a number of factors arising from the manner in which ambulances respond to calls. The modified code allows SIREN to approximate the likely vehicle behaviour which would be caused by an artificial call, without requiring direct simulation.

A portion of the regular calls can instead be temporally redistributed as artificial calls, resulting in a shorter simulation period. Preliminary results indicate that this approach can achieve a 40% reduction in simulation runtimes, whilst still producing performance measures which are statistically indistinguishable from those obtained by normal simulation.

Artificial calls were also used to establish a framework which produces detailed performance measures at key locations.

1 Introduction

The Optima Corporation has developed a software package called SIREN Predict, which is an advanced simulation tool for emergency services. The main objective of these simulations is to obtain a distribution of response times, a *response time* being the time which elapses before each emergency call is reached by an ambulance. These response times can then be used to assess the performance of the given configuration of ambulance bases and staffing rosters. All simulations conducted in this project are based on test scenarios which use the Auckland road network and randomised call data.

SIREN Predict simulates the likely behaviour of an ambulance system based on a set of historic emergency calls. It does this by following the same response procedures which would be used in practice. When an emergency call occurs at a particular location, SIREN Predict runs a dispatch procedure to determine which vehicle(s) should respond to the call. Each call will have unique staff and equipment requirements, and may require more than one vehicle to be dispatched.

When a vehicle is alerted, a *mobilisation delay* is incurred which corresponds to the time it takes for staff to load the vehicle and begin the journey. SIREN Predict then sends the closest available vehicle which has the appropriate attributes to the scene of the call. Each vehicle may travel with or without lights and sirens depending on the urgency of the call. Once an ambulance has reached the scene, the patient(s) may require transport to a particular hospital. Having completed all tasks, the ambulance will then return to its base, or serve another call. This process is outlined in Figure 1.

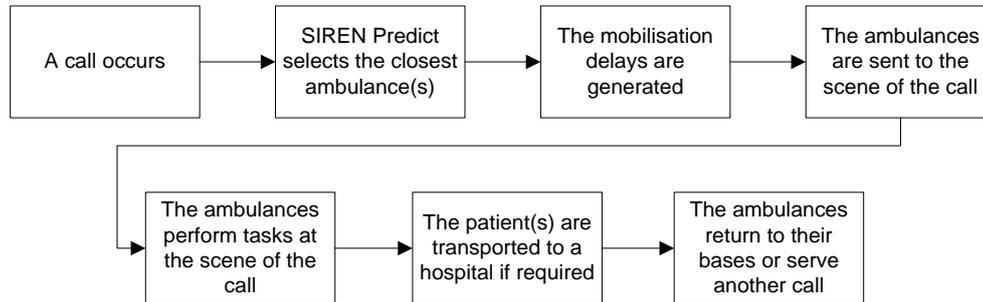


Figure 1. Outline of the Simplified Response Process in SIREN Predict

The simulation process is complicated by a number of factors, such as *diversion* which is not included in Figure 1. A diversion can occur if a high priority call takes place in the vicinity of an ambulance which is already en-route to a low priority call. In this case, the ambulance may be re-dispatched to serve the higher priority call, meaning that the low priority call must then be served by the next closest available ambulance. Consequentially, the ambulance which is initially dispatched to a particular call may not actually end up serving it.

It is common for many simulation runs to be used during an optimisation process. To find suitable locations for ambulance bases in a city, for example, a local search heuristic may be used (Kirkpatrick, 2004). This process essentially moves one base at a time, and runs a simulation at each step to see if the new configuration is an improvement. The performance of a given configuration can be assessed by examining the response times, which are commonly stipulated in ambulance performance contracts. Because a process such as this will require much iteration, and therefore many simulation runs, the Optima Corporation are particularly interested in speeding up the simulations, while still generating accurate response times. It is proposed that this can be achieved by exploiting certain properties of emergency calls.

Generating a response time can be thought of as sampling the readiness of the system to respond to a call at a particular time and location. These response times will depend on the positions of the ambulances relative to the scene of the call, and on the tasks the ambulances may be undertaking.

The historical calls contain important spatial and temporal information on arrival rates which will affect response times. While this information must be preserved, the particular week in which a call happens to occur is not important, as it is reasonable to assume that the arrival times of calls are independent provided there is no seasonality in the data. The data used in this project has been generated with arrival rates which vary across the week but have no other seasonal effects, so is ideal for this application. Consequentially, shifting the week in which a particular call occurs will still provide an equally valid response time for the call, as the call's location, day of week and time of day are preserved.

Shifting the week in which a call occurs will result in the new week having more calls than would normally be “expected.” This could potentially have an adverse effect on the response times of other calls occurring after the shifted call, because the ambulance responding to the shifted call would become busy, meaning it may no longer be available to serve future calls.

The concept of an *artificial call* is introduced to allow the week in which a call occurs to be moved, without affecting the “expected” ambulance behaviour. The vehicle movements and locations resulting from the regular calls are used as a typical environment in which a call may occur, termed the *main trace* of the simulation. When an artificial call occurs, the modified code created in this project ensures that no vehicles are dispatched and the simulation continues as if the call had not happened, but the likely response time is calculated. This *artificial response time* estimates the response time which would be observed if the artificial call were to be served normally.

Because artificial calls require no ambulance dispatch or resulting activities, the process outlined in Figure 1 is essentially reduced to the flowchart in Figure 2. This means that the response times are still obtained, but with greater efficiency. The first objective of this project is to quantify the extent to which artificial calls can be used to reduce simulation runtimes while still producing statistically valid output.

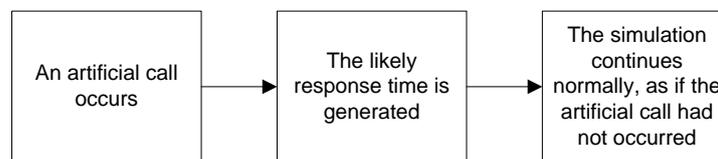


Figure 2. Outline of the Simplified Artificial Call Process

The second objective of this project is to examine how the *coverage* of key locations changes throughout a simulation run. In the context of this project, the coverage of a particular location refers to the likely response time which would be observed if an urgent call were to occur at that location. This could be evaluated by introducing artificial calls at a particular location regularly throughout the simulation. The response times of these artificial calls sample how long it would take to reach the scene if an emergency was to occur at the location at each of these times.

This report details the approach used to integrate artificial calls into SIREN Predict, the results obtained and the conclusions of the project.

2 The Artificial Call Concept

Artificial calls are introduced into SIREN Predict in order to evaluate a likely response time if a call were to occur at a given time and location. As outlined in Section 1, artificial calls require no ambulance movement, but instead estimate the likely performance which would be observed if the artificial call were to be served in the normal manner. Consequentially, artificial calls require much less computation than standard calls.

These artificial calls occur during a normal simulation run alongside the regular calls. Since an emergency call can occur at any time and location, the regular calls and resulting ambulance movements provide a typical scenario in which a call might occur. When an artificial call is introduced, the state of the simulation is essentially sampled to determine its ability to respond to such a call, if it were to occur at the given time and location.

2.1 The Main Simulation Trace

The main trace of the simulation is caused by standard (non-artificial) calls. Calls occur and ambulance movements are triggered as a result.

In order to produce meaningful response time estimates, artificial calls need to accurately capture the processes involved when a standard call is served, without disrupting the main simulation trace. If the main trace were affected by the artificial calls, then the normal number of calls occurring in a period would be exceeded, and thus the call load would no longer be representative. A simulation with artificial calls should appear to run exactly as it would without the artificial calls, meaning that the actual ambulance behaviour and attributes of the standard calls remain un-affected.

To achieve this, when an artificial call is introduced during a run, the simulation is essentially paused in order to handle this new event. The main trace of the simulation provides a sample of the system at a time when the artificial call could be introduced. A secondary trace is then created in which the artificial call is handled through some process using the current state of the main trace as its initial state. Once this secondary trace has terminated, the main trace of the simulation resumes as if the artificial call had never occurred.

2.2 Estimating Response Times for Artificial Calls

When an artificial call occurs, the main trace of the simulation is paused and a secondary trace is created in which a response time for the artificial call is estimated. In order to accurately generate these artificial response times, the system needs to mimic the dispatch behaviour of the main trace without actually simulating it in all its detail.

When a standard call occurs in the main trace, SIREN Predict begins by dispatching the closest ambulance which satisfies the particular requirements of that call. In order to achieve this, SIREN Predict implements an approach which is best represented by a Voronoi Diagram (Okabe, Boots and Sugihara, 1992), as given in Figure 3.

In this context, the dots represent the locations of the ambulances at the time when the call occurs. The Dirichlet Cell surrounding each dot contains all locations which are closer to that ambulance than to any other ambulance. Thus, the Dirichlet Cell in which a call occurs indicates the closest ambulance. SIREN Predict then dispatches this ambulance to the scene of the call. When the ambulance arrives at the scene, the response time is recorded. This general idea is outlined in Figure 3; the scene of the call is indicated by a square and the corresponding ambulance path is shown.

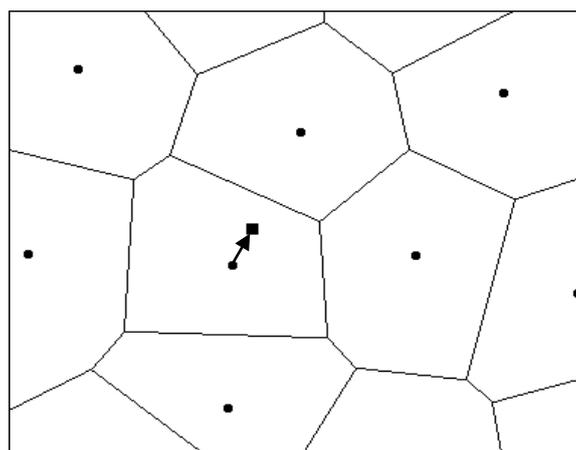


Figure 3. Voronoi Diagram Illustrating an Artificial Response Time (Adapted from Bane, 2006)

It should be noted that this approach of finding the closest ambulance is based on the travel times of an underlying road network, as opposed to simple linear distance. Furthermore, only ambulances with the attributes necessary to respond to the particular call are considered here.

In order to replicate the behaviour of the main trace, the same basic process should be used when generating artificial response times. The only difference lies in the dispatch process; instead of dispatching an ambulance and recording the response time once it reaches the scene, it is faster to simply record the likely response time based on the road network and the current vehicle locations.

Given the current location and destination of an ambulance, SIREN Predict can calculate the likely trip duration based on the travel mode of the ambulance (lights and sirens or otherwise) and the time of day, which provides the artificial response time. This duration is essentially calculated in the same way that the actual path of the ambulance would be generated in the main trace, meaning the same response times should theoretically be attained.

2.3 Handling Artificial Diversions

The main problem with this initial approach is that vehicle diversions can occur during the main trace. If an ambulance is en-route to a particular call, it may be redirected to serve a higher priority call instead. This means that the call must then be served by another ambulance, resulting in a typically longer response time.

This situation can also be represented by a Voronoi Diagram. Figure 4 shows how the Voronoi Diagram in Figure 3 can be modified to exclude the diverted ambulance, depicted by the dashed grey lines, meaning that the call will now fall in the Dirichlet Cell belonging to the closest remaining ambulance. The trip duration for this second closest ambulance can then be used in the calculation of the artificial response time for the call.

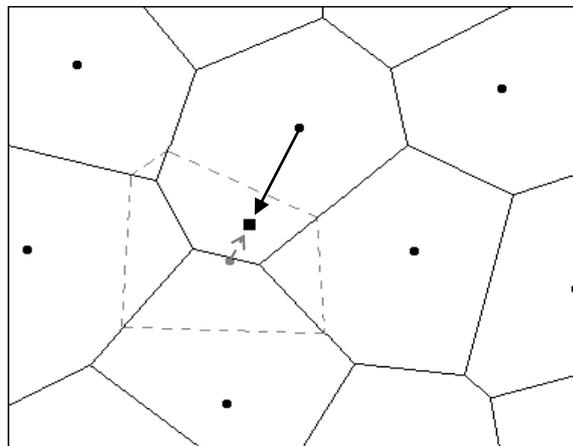


Figure 4. Amended Voronoi Diagram Illustrating an Artificial Diversion (Adapted from Bane, 2006)

The process of diversion is such that the original vehicle may be diverted at any time along its path to the original call, since diversion is triggered by the occurrence of a higher priority call. It is reasonable to assume that the call arrival process is Poisson, meaning that the durations between successive calls are exponentially distributed. However, in order for a higher priority call to cause a diversion, it must occur within the

Dirichlet Cell belonging to the appropriate ambulance. This is further complicated by the way in which the shape of this Dirichlet Cell is constantly changing as the ambulances travel. As a simplification, it is assumed that if a vehicle is diverted, the diversion is equally likely to occur at any time along its path to the scene. This assumption should be reasonable as the trip durations are relatively short compared to the period of any seasonal effects which may exist in the data.

Under this assumption, we can generate the response time for a diverted artificial call simply by adding a random, uniformly distributed portion of the first trip duration to the entire second trip duration. This approach encompasses the random portion of the first ambulance's trip duration which elapses before it is diverted. This is illustrated in Figure 4; the dashed grey vector represents the random portion of the first vehicle's trip which would be completed before diversion, and the solid vector shows the path the second ambulance would take to reach the call. Summing the durations associated with these two vectors will provide the estimate for the response time.

Another assumption made is that the location of the other ambulances will not change significantly during the first ambulance's trip before diversion occurs. In the main simulation trace, after the first vehicle has travelled towards the scene, it is likely that many of the other ambulances will also have moved. This corresponds to perturbing the nodes in the Voronoi Diagrams, which could re-configure the Dirichlet Cells. It is assumed that such changes would be minor and are unlikely to have a noticeable effect on the response time estimates.

It should also be noted that this approach does not consider the possibility that a call is diverted twice. This could occur if an ambulance is responding to a low priority (type three) call is diverted to a medium priority (type two) call, but is then diverted again to a high priority call (type one). The probability of observing a second diversion is very low, since this would require three calls of increasing urgency occurring consecutively in a close region and there are only three call priorities in the SIREN Predict Auckland data set. As a result, it should be reasonable to assume for simulation purposes that a diversion will only ever occur once per call. All simulations run to date support this.

2.4 Artificial Diversion Probabilities

The main problem when considering artificial diversions is that diversion will only occur directly in the main simulation trace. This is because the secondary trace only handles a single artificial call, meaning no other calls could occur to cause a diversion. Section 2.3 outlines the procedure for handling an artificial diversion; however it is not immediately clear when this logic should be implemented in place of the basic artificial call logic outlined in Section 2.2.

It was decided that an efficient approach was to assume that a certain proportion of artificial calls would be subject to diversion. A standard simulation was run for 182 days with no artificial calls, and the proportion of diverted calls was recorded. These proportions were split based on the priority of the call. Priority one calls, which are the most urgent, are never diverted in the simulation whereas priority three calls are most commonly diverted, being low priority.

It is logical that artificial calls should reflect this behaviour. Consequentially, artificial diversion should never occur for priority one calls, but should occur for approximately 3.9% of priority two calls and 9.1% of priority three calls. The basic artificial call logic can be applied to the remaining artificial calls for which diversion does not "occur."

It may be possible to dynamically update the artificial diversion probabilities during the simulation. The diversion probabilities outlined above were generated from one full simulation run, and as a result they may reflect trends unique to that particular data set. Perhaps a better approach would be to use these probabilities as initial values and observe the actual diversion rate in the main simulation trace. If the diversion rate deviates significantly from these initial estimates, the artificial diversion probabilities could be altered to resemble the diversion rates observed in the main trace. This approach has not implemented, but could be considered in the future.

The fact that the artificial diversion probabilities are obtained from a full simulation run may be significant. It is unclear whether this simulation would only need to be conducted once during an entire optimisation routine, or would require another run after each iteration, which would defeat the purpose of using artificial calls. It is quite possible that these probabilities would remain relatively unchanged, since each iteration in a base location heuristic, for example, involves moving one base by a relatively small amount (Kirkpatrick, 2004). As such, it may be sufficient to recalculate these probabilities once every few iterations. Alternatively, these probabilities may be dynamically re-allocated based on the observed behaviour of the main trace from each iteration. This will require further investigation before conclusions can be made.

In addition, it would be ideal if the diversion probability for an artificial call could reflect not only the priority, but also the location of a call. For example, calls occurring in the CBD may have a higher probability of being diverted than a call in a rural area due to the population density. It would also be expected that diversion would have a much less pronounced effect in a central area because it is served by many ambulances than in a remote area, meaning the distance of the second closest ambulance in the central area is likely to be significantly lower. Some of this effect may be captured by the current approach, since the second closest ambulance is computed using the same process as in the main trace of the simulation, but the spatial effects on artificial diversion remain a possible source of future work.

3 Compacting a Simulation

For reasons of computational efficiency, it may be beneficial to *compact* a simulation by converting a portion of the standard calls to be artificial calls. This process involves taking a number of weeks of calls data, and redistributing them as artificial calls occurring during the remaining weeks. This could be used to remove a period at the end of the simulation, which will shorten the simulated period. It is also expected that the runtime for the entire simulation, including both the main and secondary traces, will be reduced, since artificial calls require significantly less computation than their equivalent standard calls.

When compacting a simulation however, it is important to consider the underlying trends in the call set. Since the state of the main trace is used as a typical scenario within which the artificial call may occur, it is important that the artificial calls are re-introduced at sensible times. If this approach were not taken, the state of the main trace at the time of the artificial call may not be truly representative of the expected conditions for that call, meaning the artificial response times would likely be biased.

When compacting a simulation, the appropriate subset of calls are re-allocated by randomly changing the week in which they occur, allowing them to occur at the same time during some earlier week. This approach was taken as it will avoid adverse effects resulting from weekly patterns present in the call data. For example, a call which would

normally occur during a Friday afternoon which may typically be a busy period, should not be re-introduced as an artificial call occurring at midnight on a Sunday, which would be significantly less busy, because it would generate a shorter response time than would be expected in normal simulation.

Re-allocating the week of artificial calls will avoid adverse effects from both daily trends, such as a midday peak, and also any weekly trends, such as a weekend peak. This is because the time of day and day of week in which the call would normally occur are preserved. As a result, the state of the simulation which is sampled as the artificial call is introduced will still be valid. Annual effects, such as the different distributions of calls observed in summer and winter, could be accounted for by compacting every second week for example, rather than simply compacting a period at the end of the simulation period.

4 Improvements in Simulation Efficiency

The primary objective of simulating with artificial calls is to improve the efficiency of the simulation. This objective can be evaluated by comparing the runtimes of simulations which have been compacted to varying degrees.

To this end, 25 compacted simulations were run using the same 26 weeks of calls data. Each simulation contained a different number of *artificial weeks*; the calls occurring at the end of the simulated period which have been redistributed among the earlier weeks. The runtime of each simulation was recorded and compared with that of the standard simulation containing zero artificial weeks, as shown in Figure 5.

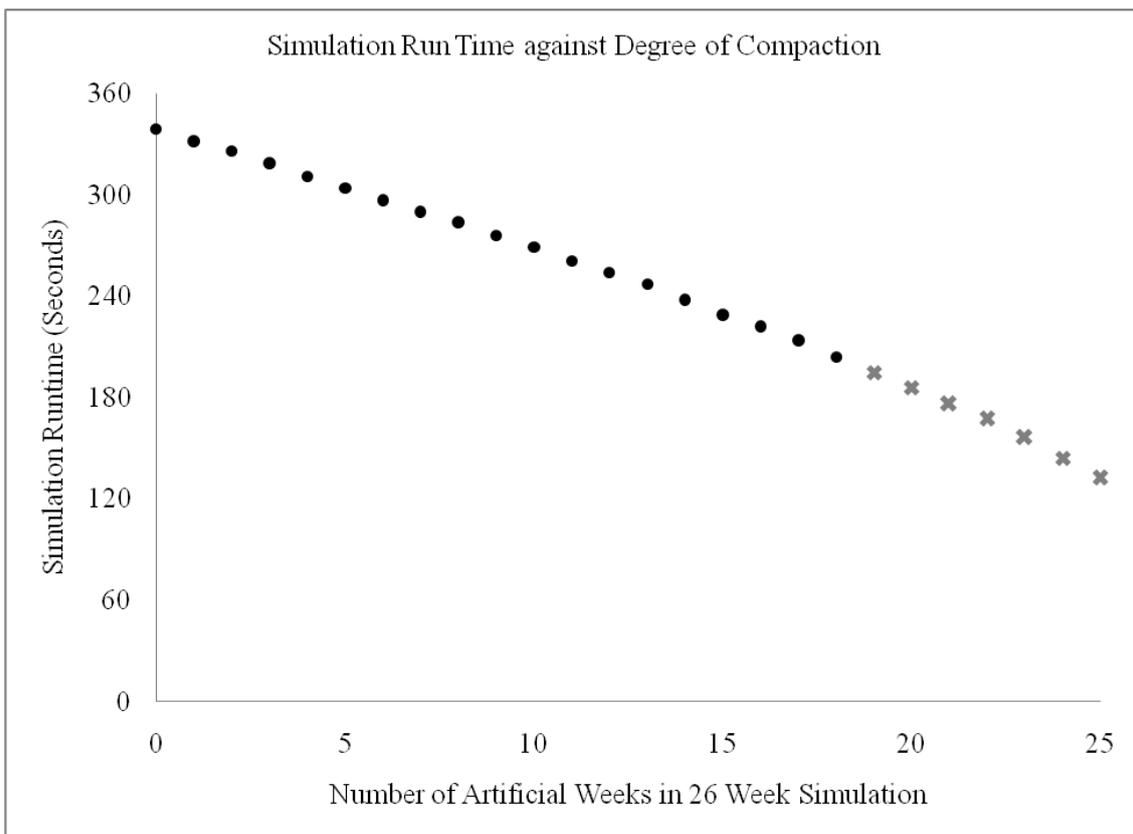


Figure 5. Plot of Runtimes for Compacted Simulations

Note that in Figure 5 the data for simulations which have been compacted beyond the upper limit of 18:8 artificial weeks to standard weeks is shown in grey. This is

because a statistical analysis, comparing the response times for each of these runs to the response times obtained from the non-compacted simulation run, yields very strong evidence that compaction beyond this limit could significantly affect the distribution of the response times. This means these runtimes may not be particularly meaningful, as the corresponding simulations are unlikely to produce useful response times. As a result, the corresponding data has been excluded from the discussion in this section.

The results indicate that compacting the simulation using artificial calls reduces the runtime significantly. There appears to be a strong linear correlation between the degree of compaction used and the resulting improvement in simulation efficiency, as illustrated in Figure 5. From these results it is estimated that a 40% reduction in runtime could be achieved using compaction, whilst still producing output which is statistically indistinguishable from the output which would be generated without compaction.

There are limits to the results generated in this section. The runtimes shown in Figure 5 necessarily include an initialisation period which cannot be reduced by compacting the simulation. As the length of this period is independent of the simulation length, compaction would have a more marked effect on longer simulations. Future work in this area could involve investigating the nature and duration of this period.

5 Estimating Coverage of Key Locations

Artificial calls can also be used to examine how the coverage of a potential high priority incident at a key location changes throughout the course of a simulation. If an artificial call is introduced at this key location at some point in the simulation, the resulting artificial response time would indicate how long it would take an ambulance to reach the scene if an emergency were to occur there. By introducing such calls regularly throughout the simulation, a profile of the likely coverage of key locations can be obtained, without disrupting the main simulation trace.

This approach would be particularly applicable to fire services, which are also of interest to the Optima Corporation. For example, the coverage of a strategic centre like the Auckland Sky Tower may be examined throughout a simulation run. This would provide a profile of possible response times if a fire were to occur, which could be compared to some acceptable maximum response time. If one or more of the artificial response times were to exceed this maximum, then it may be necessary to reconfigure base locations or take other measures to improve the responses.

Work was also completed to fit a distributions to the profile of response times. This distribution can be used to generate confidence intervals and generate in-depth performance measures for the locations of interest.

6 Conclusions

The two objectives of this project were to improve the simulation efficiency and to provide a method for estimating the coverage of key strategic locations. The conclusions relating to each of these, and general observations about artificial calls, are detailed below.

6.1 General Conclusions Relating to Artificial Calls

In general, we can conclude that the response times generated from the artificial calls are statistically indistinguishable from the response times resulting from standard calls. There is no evidence of a difference in means or variance for the two samples,

within certain limits. This verifies the implementation of artificial calls and the additional measures taken to account for diversionary effects.

6.2 Conclusions Relating to Simulation Efficiency

Simulations can be compacted using artificial calls in order to improve their efficiency. In order for this compaction to be useful, the runtime must be significantly reduced while still generating statistically indistinguishable data.

Initial tests indicate that a 40% reduction in runtime can be achieved using compaction without affecting the distribution of the response times. At this 40% level, there is no evidence of a difference in means or variance of the two samples.

This indicates that compaction effectively reduces the runtime of a simulation, and should therefore be a useful process to implement within some optimisation procedure.

6.3 Conclusions Relating to Coverage of Key Locations

By introducing artificial calls at a particular location regularly throughout a simulation, it is possible to generate a series of potential response times. This process produces useful output which has several potential applications.

Acknowledgments

First and foremost, I would like to extend my thanks to my supervisor Dr. Andrew Mason. Andrew, I thank you not only for your enduring patience and support throughout this project, but for your constant enthusiasm and belief in my abilities.

Special thanks also to the Optima Corporation for sponsoring and taking an active interest in this project. Your questions, comments and advice were greatly appreciated, and I have thoroughly enjoyed working with the SIREN software you have provided.

7 References

- Bane, A. *Voronoi Diagrams*. Adapted August 12, 2009 from <http://hirak99.googlepages.com/voronoi>
- Banks, J. And Carson, J. S. II. 1984. *Discrete-Event System Simulation*. Prentice-Hall, Inc., New Jersey.
- Cellier, F. E. And Kofman, E. 2006. *Continuous System Simulation*. Springer Science + Business Media, Inc., New York.
- Kirkpatrick, S. 2004. *Better Base Locations for the Melbourne Ambulance Service*. Department of Engineering Science, University of Auckland.
- Okabe, A, Boots, B. and Sugihara, K. 1992. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons Ltd, West Sussex.
- The Optima Corporation. *Optima Predict*. Retrieved September 23, 2009 from <http://www.theoptimacorporation.com/OptimaPredict.pdf>

Catch-up Scheduling for Childhood Vaccination

Faramroze G. Engineer
School of Mathematical and Physical Sciences
University of Newcastle
Australia
faramroze.engineer@newcastle.edu.au

Abstract

In this report, we outline the development of the core optimization technology used within a decision support tool to help providers and caretakers in constructing catch-up schedules for childhood immunization. These schedules ensure that a child continues to receive timely coverage against vaccine preventable diseases in the likely event that one or more doses have been delayed. We develop a Dynamic Programming algorithm that exploits the typical size and structure of the problem to construct optimized schedules at almost the click of a button. In using an optimization based algorithm, our approach is unique not only in methodology but also in the information, strategy and advice we can offer to the user.

The tool is being advocated by both the CDC and the American Academy of Pediatrics (AAP) as a means of encouraging caretakers and providers to take a more proactive role in ensuring timely vaccination coverage for children, as well as ensuring the accuracy and quality of a catch-up regime.

1 Introduction

With the goal of ensuring timely and accurate administration of vaccines, the Advisory Committee on Immunization Practices (ACIP) of the CDC together with the AAP and the American Academy of Family Physicians (AAFP) annually publish a *recommended* immunization schedule for children aged 0 to 6 years (see (CDC 2008)). For a child who misses the recommended time for a dose, a healthcare professional faces the challenging task of constructing a *catch-up* schedule for that child under certain rules and guidelines for the administration of the remaining doses. These rules and guidelines specify the feasible number, timing and spacing of doses of each vaccine based on the child's age, the number of doses already received, and the child's age when each dose was previously administered (see (CDC 2008) for a summary of guidelines for catch-up immunization).

Immunization programs have a significant impact on public health and have been shown to be one of the most beneficial and cost effective disease prevention measures ((Zhou et al. 2005) and (Maciosek et al. 2006)). Although the majority of school

going children in the United States that are six years and over are deemed covered against vaccine preventable diseases, most do not receive the optimal protection due to incomplete, untimely or erroneous vaccination. A comprehensive study carried out by (Luman et al. 2002) found that only 9% of children surveyed received all of their vaccinations at the recommended times and that only half received all their recommended doses by their second birthday. data gathered as part of The introduction of new vaccines to the recommended schedule adds complexity and the potential for deterioration in the overall timeliness of vaccination. Once a child falls behind the recommended schedule, statistics indicate that they often do not catch-up until close to reaching a school going age when an accelerated regime is most likely administered to meet the minimum coverage mandated by most schools.

Several factors contribute to poor and untimely vaccination rates. Some, such as parental misunderstanding and logistical difficulties affected by various environmental and socioeconomic factors are generally difficult to address and remedy. However, the problem is often exacerbated by incomplete and inaccurate catch-up schedules constructed by healthcare professionals. Constructing an accurate catch-up schedule is both a challenging and time consuming task. It therefore comes as no surprise that healthcare professionals struggle to manually construct catch-up schedules that reflect the best possible coverage for a child ((Cohen et al. 2003) and (Irigoyen et al. 2003)) and providers often fail to identify opportunities to vaccinate a child who may be at a clinic for purposes other than vaccination ((Holt et al. 1996) and (Szilagyi et al. 1993)). The complexity of the task is highlighted by the survey carried out by (Cohen et al. 2003) in which healthcare professionals were asked to construct catch-up schedules for 6 different hypothetical scenarios describing children who have fallen behind. On average, only 1.83 out of the schedules constructed for the 6 scenarios were deemed correct.

In this report, we investigate the catch-up scheduling problem and outline a Dynamic Programming (DP) algorithm that has been successfully adopted within a tool (downloadable from www.cdc.gov/vaccines/recs/scheduler/catchup.htm) developed jointly by CDC and Georgia Institute of Technology to help caretakers and providers make timely and accurate decisions with regards to childhood vaccination.

In what follows, we give a precise description of the catch-up scheduling problem in §2, outline the dynamic programming algorithm in §3, present solutions obtained for two real-life scenarios in §4, and relay some preliminary statistics about the use of the tool in practice and initial feedback from both physicians and parents in §5.

2 Problem Description and Notation

Given the current age of a child and their vaccination history (i.e., the number and timing of doses of each vaccine already administered), the catch-up scheduling problem is one of constructing a schedule for the remaining doses so that each dose is scheduled within the minimum and maximum age for that vaccine and dose, and the time separation between (not necessarily successive) doses of the same vaccine does not violate a certain minimum gap. This minimum gap may vary by vaccine, dose, current age and/or age at which some previous dose is administered. For example, the minimum gap between the second and third dose of *Hib* is 4 weeks if the current age is < 12 months, and the minimum gap is 8 weeks if the current age is ≥ 12 months and the second dose is administered at age < 15 months. In addition to

Table 1: The spacing between the first dose and remaining doses of *PCV*

dose i	dose $j > i$	age i is administered	age j is considered for administration	min gap between i and j
1	2	< 12 months	any age	4 weeks
1	3	< 12 months	any age	8 weeks
1	4	< 12 months	any age	16 weeks
1	2	< 24 but ≥ 12 months	[24, 60 months)	8 weeks
1	3,4	≥ 12 months	[24, 60 months)	∞
1	2, 3, 4	≥ 24 months	any age	∞

regulating the gap between doses of the same vaccine, doses of *live*¹ vaccines can only be administered during the same visit or at least a certain number of fixed days apart (28 days under the current guidelines). Finally, the number of *simultaneous administrations* (i.e., number of vaccinations administered during a single visit) may be discretionarily limited to avoid significant discomfort to a child.

Note that even without imposing a limit on the number of simultaneous administrations, it may not be possible to construct a schedule in which all the remaining doses can be feasibly scheduled. If a dose for some vaccine cannot be scheduled, the vaccination *series* for that vaccine is considered incomplete.

In certain cases, depending on the child’s age and/or the age at which some previous dose is given, it may be beneficial or necessary to prematurely terminate a series. For example, a child normally receives 4 doses of *PCV*, however, the second dose is deemed final if the first dose is administered at age ≥ 12 months or the current age is 24-59 months (see Table 1). In either case, the third and fourth doses are unnecessary. This form of *contraindication* can be captured by setting the required gap between the appropriate pair of doses to infinity for the appropriate range for the current age and the age when the earlier dose in the pair is administered. For example, Table 1 demonstrates how one can capture the required spacing and contraindication between the first dose and each subsequent dose of *PCV*.

Given the structure of the rules governing the spacing between doses, it may be possible to cause a contraindication by unnecessarily delaying the administration of some dose. Thus, when constructing a schedule, we are required to maximize the number of completable vaccination series, and among such candidate schedules, maximize the total number of scheduled doses and minimize the total delay from the recommended age of administering these doses.

We next introduce notation for the catch-up scheduling problem:

V	the set of vaccines.
n_v	the total number of doses that constitute the completion of a vaccination series of $v \in V$.
$t_{v,i}^{min}$	the minimum age for administering dose $i \in \{1, \dots, n_v\}$ of $v \in V$.
$t_{v,i}^{max}$	the maximum age for administering dose $i \in \{1, \dots, n_v\}$ of $v \in V$.
$t_{v,i}^{rec}$	the recommended age for administering dose $i \in \{1, \dots, n_v\}$ of $v \in V$.

¹A “live virus” vaccine is a vaccine that contains a “living” virus that is able to give and produce immunity, usually without causing illness.

$t_{v,i,j}^{gap}(t, t')$ the minimum gap required between dose i and dose $j > i$ of vaccine v when i is administered at age t , and j is being considered for administration at age t' .

$V^{live} \subseteq V$ the set of live vaccines.

t^{live} the minimum gap required between doses of any live vaccines when they are not administered during the same visit.

M the maximum number of simultaneous administrations.

Given a (possibly partial) schedule denoted by s , we use the following notation to define the number and timing of doses scheduled in s :

\mathbf{n}_v^s the number of doses of vaccine v that have been scheduled in s .

$\mathbf{t}_{v,i}^s$ is the age at which dose $i \in \{1, \dots, \mathbf{n}_v^s\}$ of v is scheduled in s .

\mathbf{t}^s an age by which time all doses scheduled in s are administered, i.e., such that $\mathbf{t}^s > \mathbf{t}_{v,i}^s$ for all $i \in \{1, \dots, \mathbf{n}_v^s\}$ and $v \in V$.

Finally, we introduce some additional notation to define some important characteristics of s :

$m(s, t)$ the number of vaccinations scheduled at age t , i.e., $m(s, t) = |\{v \in V : \mathbf{t}_{v,i}^s = t \text{ for some } i\}|$.

$c(s)$ the number of completable vaccination series, i.e., $c(s) = |\{v \in V : n_v = \mathbf{n}_v^s\}|$.

$n(s)$ the total number of doses scheduled, i.e., $n(s) = \sum_{v \in V} \mathbf{n}_v^s$.

$d_{v,i}(s)$ the delay from the recommended age of administering dose i of vaccine v , i.e., $d_{v,i}(s) = \max\{0, \mathbf{t}_{v,i}^s - t_{v,i}^{rec}\}$.

$d(s)$ the total delay from the recommended age of administering the scheduled doses, i.e., $d(s) = \sum_{v \in V} \sum_{i=1}^{\mathbf{n}_v^s} d_{v,i}(s)$.

For a given schedule s , we then define its feasibility and extension as follows.

Definition 1 (Feasibility) *A schedule s is feasible if:*

F1. s satisfies the time windows for individual doses, i.e. $t_{v,i}^{min} \leq \mathbf{t}_{v,i}^s \leq t_{v,i}^{max}$ for all $i \in \{1, \dots, \mathbf{n}_v^s\}$ and $v \in V$,

F2. s satisfies the gap requirements between doses of the same vaccine, i.e., $\mathbf{t}_{v,j}^s - \mathbf{t}_{v,i}^s \geq t_{v,i,j}^{gap}(\mathbf{t}_{v,i}^s, \mathbf{t}_{v,j}^s)$ for all $i \in \{1, \dots, \mathbf{n}_v^s\}$, $j \in \{i+1, \dots, \mathbf{n}_v^s\}$ and $v \in V$,

F3. s satisfies the gap requirement between doses of live vaccines, i.e.,

$$|\mathbf{t}_{w,j}^s - \mathbf{t}_{v,i}^s| \begin{cases} = 0, & \text{or} \\ \geq t^{live} \end{cases}$$

for all $i \in \{1, \dots, \mathbf{n}_v^s\}$, $j \in \{1, \dots, \mathbf{n}_w^s\}$ and $v, w \in V^{live}$, and

F4. No more than M doses are scheduled in s at any given age, i.e. $m(s, t) \leq M$ for all t .

Definition 2 (Extension) *A schedule s' is considered an extension of schedule s , if s' can be obtained from s by scheduling any remaining doses of s on or after \mathbf{t}^s , i.e.,*

E1. $\mathbf{n}_v^{s'} \geq \mathbf{n}_v^s$ for all $v \in V$,

E2. $\mathbf{t}_{v,i}^{s'} = \mathbf{t}_{v,i}^s$ for all $i \in \{1, \dots, \mathbf{n}_v^s\}$ and $v \in V$, and

E3. $t_{v,i}^{s'} \geq t^s$ for all $i \in \{\mathbf{n}_v^s + 1, \dots, \mathbf{n}_v^{s'}\}$ and $v \in V$.

The catch-up scheduling problem can then be stated as follows: Given a feasible schedule s , the catch-up scheduling problem is one of extending s to a feasible schedule s^* so that

$$\begin{bmatrix} c(s^*) \\ n(s^*) \\ -d(s^*) \end{bmatrix} \geq_L \begin{bmatrix} c(s') \\ n(s') \\ -d(s') \end{bmatrix}$$

for any other feasible extension s' of s . Here we use \geq_L to represent a lexicographical ordering of vectors. Thus, s^* is the best extension of s with respect to (1) the number of completable vaccination series, (2) the number of scheduled doses, and (3) the total delay from the recommended age of administering the scheduled doses, in the stated order of priority.

The catch-up scheduling problem encapsulates many of the complexities of traditional machine scheduling problems. It therefore comes as no surprise that the problem is NP-complete. In fact, it can be shown that the problem remains NP-complete under various simplifications. Despite the similarities to traditional machine scheduling, the catch-up scheduling problem differs since it may not be possible to construct a feasible schedule with all remaining doses and the required separation between doses varies with not only the current age of the child but also, the age at which some previous dose is administered. As a result, one has to employ a multi-level objective that is not typically found in the literature, but at the same time, is ideally suited for DP.

3 Solution Approach

Although the catch-up scheduling problem is NP-complete, the typical size of the problem in practice does not pose a huge challenge. Instead, the challenge we face is in being able to solve these problems consistently within a short amount of time (generally a few seconds).

In this section, we outline a dynamic programming algorithm for solving the catch-up scheduling problem. We start by identifying “dominance” of one schedule over another.

Definition 3 (Dominance) *Given two feasible schedules s_1 and s_2 such that $\mathbf{t}^{s_1} = \mathbf{t}^{s_2}$, we say s_1 dominates s_2 or $s_1 \succeq s_2$, if we can extend s_1 to a schedule that is at least as good as any schedule obtained from extending s_2 .*

Only non-dominated schedules are warranted in the construction of the optimal schedule. Unfortunately, given the complexity of the problem, it is unlikely that there exist efficient necessary conditions for proving dominance of one schedule over another. However, we can find reasonable sufficient conditions by observing that the required spacing between a pair of doses of the same vaccine is generally non-decreasing in the age the first dose in the pair is administered. We first identify within a given schedule s certain “critical” vaccine-dose pairs whose timing in s may prevent some future dose being administered at some age on or after \mathbf{t}^s :

$$\Psi(s) = \left\{ (v, i) : \begin{array}{l} v \in V, i \in \{1, \dots, \mathbf{n}_v^s\} \text{ such that either:} \\ \text{i. } \mathbf{t}_{v,i}^s + t_{v,i,j}^{gap}(\mathbf{t}_{v,i}^s, t') > t' \text{ for some } j \in \{\mathbf{n}_v^s + 1, \dots, n_v\} \\ \text{and } t' \geq \mathbf{t}^s, \text{ or} \\ \text{ii. } \mathbf{t}_{v,i}^s + t^{live} > \mathbf{t}^s \text{ and } v \in V^{live} \end{array} \right\}.$$

Dominance can then be recognized by using the criteria in the following proposition.

Proposition 1 *If $t_{v,i,j}^{gap}(t, t')$ is non-decreasing in t for all $v \in V$, $i \in \{1, \dots, n_v\}$, $j \in \{i + 1, \dots, n_v\}$, and t' , then for any two feasible schedules s_1 and s_2 such that $\mathbf{t}^{s_1} = \mathbf{t}^{s_2}$, $s_1 \succeq s_2$ if the following conditions hold:*

- D1.** $\mathbf{n}_v^{s_1} \geq \mathbf{n}_v^{s_2}$ for all $v \in V$,
- D2.** $\mathbf{n}_v^{s_1} = \mathbf{n}_v^{s_2}$ for all $v \in V^{live}$ s.t. $(v, \mathbf{n}_v^{s_1}) \in \Psi(s_1)$,
- D3.** $t_{v,i}^{s_1} \leq t_{v,i}^{s_2}$ for all $(v, i) \in \Psi(s_1)$ such that $i \in \{1, \dots, \mathbf{n}_v^{s_2}\}$, and
- D4.** $\sum_{v \in V} \sum_{i=1}^{\mathbf{n}_v^{s_2}} d_{v,i}(s_1) \leq \sum_{v \in V} \sum_{i=1}^{\mathbf{n}_v^{s_2}} d_{v,i}(s_2)$.

The dominance criteria simply state that s_1 dominates s_2 if (1) s_1 has scheduled at least as many doses as s_2 for each vaccine, (2) s_1 has scheduled the same number of doses as s_2 for any live vaccine whose last dose in s_1 prohibits the scheduling of any other live vaccine at age \mathbf{t}^{s_1} , (3) the timing of critical doses scheduled in s_1 is no later in s_1 than in s_2 , and (4) the total delay in administering doses in common is no worse in s_1 than in s_2 .

If the required gap between a pair of doses is non-decreasing in the age the first dose is scheduled, then a schedule with the critical doses scheduled earlier has more flexibility in choosing dates for scheduling future doses. The remaining non-critical doses play no part in determining the timing of any remaining doses but may contribute to the overall quality of the schedule. Given sufficient criteria such as **D1-D4** for efficiently recognizing dominance, we can then use the DP algorithm outlined in Algorithm 1 to construct an optimal extension of a schedule.

For a given schedule s , we define $\tau(s) = \{t_0, t_1, t_2, \dots\}$ to be the ordered set of time points corresponding to possible ages any remaining dose can be administered starting with $t_0 = \mathbf{t}^s$. Starting with s and age t_0 , the DP algorithm at iteration k constructs all possible schedules that can be obtained by extending all non-dominated schedules constructed for age t_{k-1} by a single time period. We denote with $\langle s', V', t \rangle$, the new schedule resulting from extending schedule s' by scheduling each vaccine in the set $V' \subseteq V$ at age t and setting $\mathbf{t}^{\langle s', V', t \rangle} = t$. Any newly constructed schedule that is infeasible is immediately discarded. Otherwise, $\langle s', V', t \rangle$ is checked for dominance against the candidate set of schedules in $\mathcal{S}(t_k)$ during the **insert** procedure in which any dominated schedules are immediately discarded.

Algorithm 1 The DP Algorithm

Input: schedule s .

Initialize:

$\tau(s) = \{t_0, t_1, t_2, \dots\}$, $\mathcal{S}_{t_0} \leftarrow \{s\}$, and $\mathcal{S}_{t_k} \leftarrow \{\emptyset\}$ for all $k = 1, \dots, |\tau(s)|$.

Main Loop:

for $k = 1, \dots, |\tau(s)|$ **do**

 /* Iteration k */

for all $s' \in \mathcal{S}(t_{k-1})$ **do**

for all $V' \subseteq V$ s.t. $|V'| \leq M$ **do**

if $\langle s', V', t_k \rangle$ is feasible **then**

insert $\langle s', V', t_k \rangle$ into \mathcal{S}_{t_k}

Output: **return** schedule $s^* \in \mathcal{S}_{t_{|\tau(s)|}}$ such that $[c(s^*), n(s^*), -d(s^*)] \geq_L [c(s'), n(s'), -d(s')]$ for all $s' \in \mathcal{S}_{t_{|\tau(s)|}}$.

It can be shown by induction that at the start of iteration k of Algorithm 1, $\mathcal{S}_{t_{k-1}}$ contains at least one schedule that can be extended to obtain some best extension of s . Thus, starting with a partial schedule s containing only the past vaccination history of the child and t^s corresponding to the current age of the child, the DP constructs the optimal schedule for administration of the remaining doses. Using the given dominance criteria, we are able to solve most instances within a second and have never encountered any practical instance that took longer than a handful of seconds to solve.

4 A Case Study of Two Scenarios

In this section, we present two solutions obtained for two different real-life scenarios for children requiring catch-up schedules. These cases present varying levels of urgency in terms of how far behind a child has fallen as well as demonstrate the impact of different rules that govern the timing, spacing, and premature termination of a series.

Case 1: A 4 month old child who has received *HepB* at birth and one each of *HepB*, *DTaP*, *Hib*, and *PCV* at 2 months of age.

Case 2: A one year old child without any vaccination.

Figures 1-2 display the different solutions obtained for each of the two scenarios. The first two rows of each chart displays the age and dates for scheduled visits. The first column corresponds to the vaccine line-up. Each box in the chart represents four possible outcomes for a scheduled dose:

AD – an already Administered Dose,

CD – a Catch-up Dose scheduled after the recommended age,

OD – an On-time Dose scheduled during the recommended age, and

PD – a Preemptive Dose scheduled before the recommended age.

At the end of each row we give a tally of doses administered/scheduled out of the total recommended for a vaccination series to be considered completed.

Consider the solution obtained for Case 1 shown in Figure 1. Note also that although this child is 4 months behind for 5 of the 9 vaccines, the schedule has the child catch-up for all but *Rota* by 6 months of age. This is indicated by the trailing **OD** boxes at 6 months of age.

The final solution (Figure 2) displays the solution for Case 2 which is often the standard scenario for internationally adopted or immigrant children presumed not to have received any vaccinations (see (Cohen and Veenstra 2006)). Since the one year old child is assumed not to have received any vaccinations, the standard recommendation would be to vaccinate the child with all 8 vaccines that can be feasibly administered on the current day. However, unless a clinic has many of these in combination, it is unlikely that they would actually administer 8 shots during a single visit. Figure 2 displays the solution when the user chooses to restrict the maximum number of simultaneous administrations to 4.

Schedule generated for: ***** on Apr 21, 2008 (04/21/2008)
 Birth Date: Dec 21, 2007 (12/21/2007). Current Age: 0 year/s, 4 month/s and 0 week/s

Age	0-4 weeks	1-3 months	3-6 months			6-12 months	12-15 months	15-18 months		18-24 months	3-4 years	4-6 years	
Rec. Date (mm/dd/yy)	12/21/07	02/21/08	Today 04/21/08	05/19/08	06/16/08	12/15/08	03/09/09	04/10/09	06/19/09	11/20/09	12/12/11	12/13/13	Tally
HepB	AD	AD			OD								3/3
Rota													0/3
DTaP		AD	CD		OD		OD				OD		5/5
Hib		AD	CD		OD	OD							4/4
PCV		AD	CD		OD	OD							4/4
IPV			CD	CD	OD						OD		4/4
MMR						OD					OD		2/2
Var						OD					OD		2/2
HepA						OD		OD					2/2

AD - Administered Dose CD - Catch-up Dose OD - On-time Dose PD - Preemptive Dose

Figure 1: A catch-up schedule constructed for Case 1

Schedule generated for: ***** on Apr 21, 2008 (04/21/2008)
 Birth Date: Apr 21, 2007 (04/21/2007). Current Age: 1 year/s, 0 month/s and 0 week/s

Age	6-12 months	12-15 months					15-18 months		18-24 months			3-4 years	4-6 years	
Rec. Date (mm/dd/yy)	Today 04/21/08	04/28/08	05/19/08	06/02/08	06/16/08	07/12/08	08/25/08	10/18/08	11/03/08	12/15/08	03/21/09	04/16/11	04/13/13	Tally
HepB		CD		CD			OD							3/3
Rota														0/3
DTaP	CD		CD		CD					CD		OD		5/5
Hib	CD				CD									2/4
PCV	CD				CD									2/4
IPV	CD		CD		OD							OD		4/4
MMR		OD										OD		2/2
Var		OD										OD		2/2
HepA		OD							OD					2/2

AD - Administered Dose CD - Catch-up Dose OD - On-time Dose PD - Preemptive Dose

Figure 2: A catch-up schedule constructed for Case 2 when $M = 4$

5 The Scheduler in Practice

The tool was downloaded over 37,110 times from CDC’s website during the first of its deployment. The tool has also been referenced in several sections of the latest edition of the AAP Red Book ((Pickering et al. 2009)) considered one of the most definitive guides to immunology and childhood immunization. It has also been featured in over 50 different (online) magazines and news articles including the Washington Post ((Kritz 2008)), U.S. News ((Shute 2008)), Discoveries and Breakthroughs Inside Science ((Ivanhoe Broadcast Network 2008)), AAP News ((Cash 2008)), and was recently featured in CDC’s 2009 back-to-school immunization campaign (www.cdc.gov/Features/CatchUpImmunizations/).

Several physicians including Dr. Bocchini (chair of the AAP Committee on Infectious Diseases) and Dr. Robert Harrison (Children’s Healthcare of Atlanta) who have used the tool have commented that in a busy office they appreciate the rapidity with which decisions can be made by using the tool when a child falls behind in his or her immunizations. They noticed that parents have brought the schedule with them to physician visits and are able to ask appropriate questions and feel they are part of the process. The amount of time saved in determining what

vaccines need to be administered when a child is behind and the confidence that the recommended vaccines are administered are major benefits to them. Moreover, physicians feel that the scheduler helps them ensure that children receive vaccines within the recommended guidelines.

6 Conclusions

Healthcare providers are faced on a daily basis with the challenging task of constructing catch-up schedules for childhood immunization. The manual process of constructing such schedules is both difficult and time consuming often resulting in inaccurate or incomplete schedules that can have a detrimental impact on coverage rates and children's health.

In this report, we examine the complicating characteristics of the catch-up scheduling problem and design a DP algorithm that constructs an optimal (with respect to the potential coverage) schedule for a child based on their vaccination history and current age. By observing and exploiting the fact that the required separation between doses of the same vaccine is non-decreasing in the age some previous dose is administered, we derive dominance criteria that are sufficiently tight in practice to solve practically sized problems very quickly.

The tool is currently available for download from CDC's website (www.cdc.gov/vaccines/scheduler/catchup.htm) and is being advocated by both the CDC and AAP as a means of encouraging caretakers and providers to take a more proactive role in ensuring timely vaccination coverage of children under their care, and ensuring the accuracy and quality of a catch-up regime. The tool has already provided considerable direction for the rule makers in establishing a more rigorous framework for maintaining consistency in the way current and future rules are stated and in dealing with an infeasible vaccination history. Although it is hard to further quantify the impact of the tool on public policy and/or practice, based on initial feedback and download statistics, we have observed that it has aroused keen interest in the health care community as well as the general public. The fact that the tool is being advocated by both the CDC and AAP additionally testifies to the general acceptance within the healthcare community of the need for such a tool in clinical practice.

Acknowledgments

The author is grateful to Dr. Larry Pickering of from the Centers for Disease Control and Prevention and Prof. Pınar Keskinocak of Georgia Institute of Technology for their vision and support in the development of this tool.

References

- Cash, S. 2008. "Online scheduler keeps track of missed immunizations." *AAP News* 29, no. 7 (July).
- CDC. 2008. *Recommended schedules and guidelines for catch-up scheduling*. United States: Department of Health and Human Services, National Center for Immunization and Respiratory Diseases. URL: www.cdc.gov/vaccines/recs/schedules/.

- Cohen, A.L., and D. Veenstra. 2006. "Economic Analysis of Prevacination Serotesting Compared With Presumptive Immunization for Polio, Diphtheria, and Tetanus in Internationally Adopted and Immigrant Infants." *Pediatrics* 117 (5): 1650–1655.
- Cohen, N.J., D.S. Lauderdale, P.B. Shete, J.B. Seal, and R.S. Daum. 2003. "Physician Knowledge of Catch-up Regimens and Contraindications for Childhood Immunizations." *Pediatrics* 111 (May): 925–932.
- Holt, E., B. Guyer, N. Hughart, V. Keane, P. Vivier, A. Ross, and D. Strobino. 1996. "The Contribution of Missed Opportunities to Childhood Underimmunization in Baltimore." *Pediatrics*, vol. 97 (April).
- Irigoyen, M., S. Findley, P. LaRussa, S. Chen, P. Sternfels, L. Cooper, A. Caesar, and M. Ewing. 2003. "Impact of Immunization Schedule Changes on Missed Opportunities for Vaccination." The 37th National Immunization Conference Chicago, IL.
- Ivanhoe Broadcast Network, Inc. 2008. "Keeping Vaccinations On Track." *Discoveries and Breakthroughs Inside Science*. URL: www.ivanhoe.com/science/story/2008/06/434a.html.
- Kritz, F.L. 2008. "For Parents, an Easier Way to Track Vaccines." *The Washington Post, Health*, no. July 15.
- Luman, E.T., M.M. McCauley, S. Stokley, S.Y. Chu, and L.K. Pickering. 2002. "Timeliness of Childhood Immunizations." *American Journal of Preventive Medicine*, vol. 110.
- Maciosek, M.V., A.B. Coffield, N.M. Edwards, T.J. Flottemesch, M.J. Goodman, and L.I. Solberg. 2006. "Priorities Among Effective Clinical Preventive Services: Results of a Systematic Review and Analysis." *American Journal of Preventive Medicine* 31:52–61.
- Pickering, L.K., C.J. Baker, D.W. Kimberlin, and S.S. Long. 2009. *Red Book: Report of the Committee on Infectious Diseases*. 28th.
- Shute, N. 2008. "A New Tool to Manage Your Child's Vaccine Schedule." *U.S. News and World Report, Online*, no. May 27.
- Szilagyi, P.G., L.E. Rodewald, S.G. Humiston, R.F. Raubertas, L.A. Cove, C.B. Doane, P.H. Lind, M.S. Tobin, K.L. Roghmann, and C.B. Hall. 1993. "Missed Opportunities for Childhood Vaccination in Office Practices and the Effect on Vaccination Status." *Pediatrics*, vol. 91 (January).
- Zhou, F., J. Santoli, M.L. Messonnier, H.R. Yusuf, A. Shefer, S.Y. Chu, L. Rodewald, and R. Harpaz. 2005. "Economic Evaluation of the 7-Vaccine Routine Childhood Immunization Schedule in the United States, 2001." *Archives of Pediatrics and Adolescent Medicine* 159:1136–1144.

Making Smarter Transportation Investments

Kane Harton
Department of Engineering Science
University of Auckland
New Zealand
kharton@gmail.com

Abstract

Transportation is vital to any city as a facilitator of a productive, efficient economy. Auckland's transportation infrastructure suffers from congestion which costs the economy 900 million dollars per year. It is therefore important to invest in a combination of transportation projects that will resolve this problem. This is framed by uncertainty of the future value of these transportation investments given uncertain oil prices.

In Auckland (and throughout the world) transportation investments are not explicitly considered as a portfolio and neither is the effect that oil may have on their future value. There is no published academic work that considers the optimal selection of transportation investments, thus this paper explores a new application for optimisation.

This paper investigates the transportation investment problem by considering a subset of investments in the Kingsland area of Auckland. This work treats these investments as a portfolio and considers oil price stochasticity when selecting the investments. By doing this we are able to significantly enhance the combined value of the total investment.

A conditional value-at-risk formulation of the problem is also explored to understand how risk can effectively be managed within transportation investments.

1 Introduction

Transportation infrastructure requires large upfront investments, yet there is a significant degree of uncertainty around the future value of these investments. Many factors will affect their performance such as oil price and population growth.

A city such as Auckland has limited funds to invest in infrastructure and therefore must choose the best investments to make at any given time; therefore it is important to make 'smart' investments.

1.1 Funding Environment

There is a complex funding system through which projects are selected and funded. Both the Auckland Regional Transportation Authority (ARTA) and the New Zealand Transportation Authority (NZTA) assess investments, while the NZTA allocates funding to these investments. Funding is primarily allocated according to the economic efficiency of investments which is determined by performing economic evaluations according to the NZTA's Economic Evaluation Manual (EEM).

1.2 Oil Price Uncertainty

There has always been and will continue to be uncertainty as to what future oil prices will be. Intuitively we know oil price will impact the demand for transportation as was recently demonstrated in Auckland with 2008's oil price hike reducing road traffic volumes by 3% over the year (Borley, 2008).

This is a well known phenomenon. Elasticities have been established in the literature (Kennedy & Wallis, 2007) which measure the change in travel behaviour for different modes of travel. It follows then that as oil price is uncertain demand is uncertain, and this should be taken into account while performing economic evaluations.

1.3 Weaknesses

There are three key weaknesses inherent in the current funding system that will be addressed in this paper:

- Oil price uncertainty – this is only implicitly considered when evaluating investments under the EEM where it is assumed oil price will not affect demand, i.e. oil price remains constant.
- Investments are considered independently – each investment is considered on its own merit – but realistically making a change to one part of the transportation network will affect the value of the entire network.
- Financial risk – each investment possesses a degree of financial risk as a result of uncertainty. Currently this financial risk is not considered. There may be a small risk associated with a single investment, however many investments with similar risk profiles will have a large collective risk exposure.

The author was unable to identify previous research on these topics in relation to transportation investments, thus this paper highlights new avenues for optimisation.

1.4 Aim

The aim of this research is to explore how selecting transportation investments as a portfolio with uncertain value, arising from oil price uncertainty, can enhance Auckland's transportation system.

2 Subset Approach

It is very difficult to find detailed information about all potential transportation investments across Auckland. Thus we decided to focus on a small subset of transportation investments that when examined would provide an insight to the larger problem.

2.1 Kingsland

ARTA were able to identify a set of transportation investments in the Auckland suburb of Kingsland which had been committed to by the NZTA.

We consider eight different investments of different types (locations shown in Figure 1), including; public transportation, travel demand management, roading, rail, cycling and pedestrian improvements.

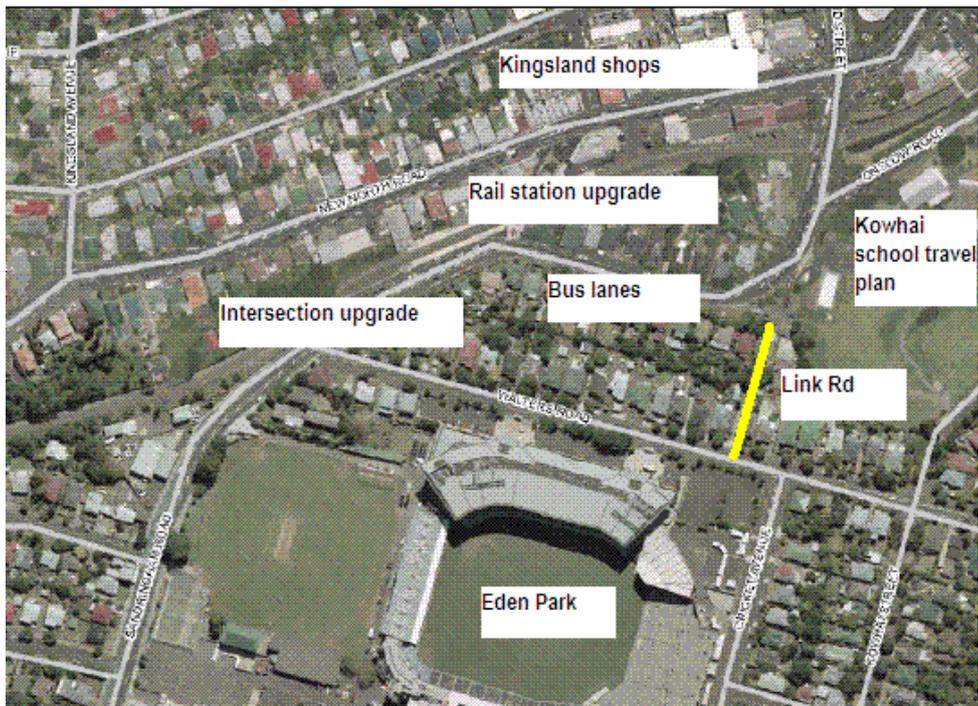


Figure 1: Kingsland subset of investments (Percy & Harton, 2009)

3 Economic Evaluations

An economic analysis assumes that all benefits and costs of a transportation investment can be quantified as a monetary value. A set of methodologies are required to determine these values and are defined by the EEM. The EEM quantifies costs and benefits such as travel time, congestion, vehicle operating costs, accident costs and health benefits. The most significant of these costs tends to be travel time.

Travel time savings do not accrue linearly on road networks thus it is necessary to simulate demand scenarios using a traffic modelling tool for road based investments.

3.1 No Induced Demand

For the purposes of this research we assume that there is no induced demand. Induced demand is the phenomenon of demand increasing as supply increases. This is a valid simplifying assumption as the investments are implemented in a small area over short distances relative to the length of an average trip making it unlikely that the changes will induce any significant demand.

3.2 Benefits as a Function of Demand

The costs and benefits evaluated using the EEM vary primarily with the level of demand for a given service. Thus we are able to express our economic evaluations as a function of demand which allows us to determine the investment benefits under different demand scenarios which result from oil price uncertainty.

4 Kingsland Model

In this section we present the portfolio optimisation model. Starting with a simple (deterministic) formulation ignoring oil price uncertainty, we develop stochastic models that allow us to assess the value of the stochastic solution.

Indices

a = investment from the set of investments A

t = year with $0 \leq t \leq n = 30$

Parameters

$\beta_{a,t}$ = the benefit of investment a in year t

$\gamma_{a,t}$ = the cost of investment a in year t

$\delta_{a,t}$ = the demand for investment a in year t

Decision variables

$x_a = 1$ if investment a selected, 0 otherwise

Model Simple Formulation

Maximise $\sum_{a \in A} (\beta_a - \gamma_a) x_a$

Where

$$(1) \quad \beta_a = \sum_{t=0}^n \left(\frac{\beta_{a,t}(\delta_{a,t})}{1.08^t} \right)$$

$$(2) \quad \gamma_a = \sum_{t=0}^n \left(\frac{\gamma_{a,t}}{1.08^t} \right)$$

Subject to

$$(3) \quad x_a \in \{0,1\} \quad \forall a$$

Explanation

The above formulation ignores oil price uncertainty and provides the optimal portfolio of investments for a single demand scenario.

The objective function we have selected maximises the profit from the set of investments. It could be argued that we should be using an objective function which maximises the economic efficiency of the investment portfolio. However, if we view the problem from the perspective of a funding agency we can see the objective is irrelevant as they will seek to allocate all funding i.e. costs are fixed, which means both objectives will maximise benefits.

Term (1) describes the benefits in each year for a given investment as a function of demand for that investment and discounts the annual benefit by 8% to reflect the NZTA's discount rate. Term (2) describes the cost in each year of a given investment and also discounts that at 8%. Term (3) recognises that this is an integer investment problem and that we can't make a fractional investment.

4.1 Stochastic Formulations

In order to take uncertainty into account in our model it is necessary to create stochastic formulations of the model. The formulations below show how they are different to the simple formulation. If a term is not mentioned it remains the same as in the simple formulation.

Expected Value (EV)

Model EV Formulation

Maximise $\sum_{a \in A} (\beta_{a,E\delta} - \gamma_a) x_a$

Where

$$(1) \beta_{a,E\delta} = \sum_{t=0}^n \left(\frac{\beta_{a,t}(E\delta_{a,t})}{1.08^t} \right)$$

Explanation

This chooses an optimal policy based solely on the expected demand scenario. The model specifies that the single demand scenario used is the expected demand scenario.

Expected Expected Value (EEV)

The expected value solution is simulated under different oil price scenarios to find the EEV. This provides a lower bound estimate of the optimal policy as the expected value policy does not take into account the effect of uncertainty.

Here and Now (HN)

Indices

i = demand scenario from the set of demand scenarios I

Parameters

p_i = probability of scenario i

Model HN Transportation Investment Selection

Maximise $\sum_{i \in I} \sum_{a \in A} p_i \times (\beta_{i,a} - \gamma_a) x_a$

Where

$$(1) \quad \beta_{i,a} = \sum_{t=0}^n \left(\frac{\beta_{i,a,t}(\delta_{i,a,t})}{1.08^t} \right)$$

Explanation

This is the stochastic formulation which represents the real world. It recognises that we can only choose one set of transportation investments now but allows for demand uncertainty by making demand a stochastic input.

Value of Stochastic Solution (VSS)

Formula VSS

$$VSS = HN - EEV$$

Explanation

The VSS examines the difference between the EEV problem and the HN problem. This tells us how a naive approach (which ignores stochasticity) compares to an approach which does allow it, and consequently informs us how important considering stochasticity is when planning transportation investments.

4.2 Implementation

In order to implement this model a number of points need to be addressed. We need to generate oil price scenarios and then turn these scenarios into demand scenarios.

4.2.1 Oil Price Scenarios

To generate oil price scenarios we use Donovan's meta-model scenario generator which has been used in reports to NZTA and ARC (Auckland Regional Council) discussing the future of oil prices (Donovan, 2008a, 2008b). The methodology behind the oil price scenario generator is discussed in detail in Donovan's report to the ARC and will be summarized here (Donovan, 2008a).

A number of reputable institutions created their own oil price forecasts which are publicly available. Donovan's meta-model combines and weights these oil price forecasts. Scenarios are then randomly generated around these forecasts using historical information about the variability between past forecasts and actual oil price.

4.2.2 Oil Price to Demand

The generated oil price scenarios need to be linked to demand and cross-price elasticities are used to do this. These cross-price elasticities describe how the change in price of one good, oil price, changes the demand for another good, demand for transportation.

The cross-price elasticities that we use have been determined empirically through research by a number of different organisations and have both a long and short term component representing how people react to price over time. Population growth in the area must also be considered in order to take into account how an increasing population also affects traffic demand. The formula below shows how we are able to transform historical demands and oil price predictions into a demand forecast.

Parameters

M = mode of transportation

n = year

δ_n^M = demand for transportation of mode m in year n

O_n = oil price in year n

ε_{LR} = long range elasticity

ε_{SR} = short range elasticity

g = annual population growth

Formula Demand as a Function of Elasticity

$$\begin{aligned} \delta_{t+4}^M = & \delta_{t+1}^M + \left(\delta_{t+1}^M \times \frac{O_{t+1} - O_t}{O_t} \times \varepsilon_{LR} \right) + \left(\delta_{t+2}^M \times \frac{O_{t+2} - O_{t+1}}{O_{t+1}} \times \varepsilon_{SR} \right) \\ & + \left(\delta_{t+3}^M \times \frac{O_{t+3} - O_{t+2}}{O_{t+2}} \times \varepsilon_{SR} \right) + ((\delta_{t+1}^M + \delta_{t+2}^M + \delta_{t+3}^M) \times g) \end{aligned}$$

4.3 Results

The graph on the next page shows how the value of the Giveway investment varies across our set of oil price scenarios. The Giveway investment is a fairly typical road investment and involves upgrading the intersection of Walters Rd and Sandringham Rd to an enhanced giveway layout. Just from looking at this graph we can tell that oil price stochasticity has a major effect on the value of an investment and should not be ignored.

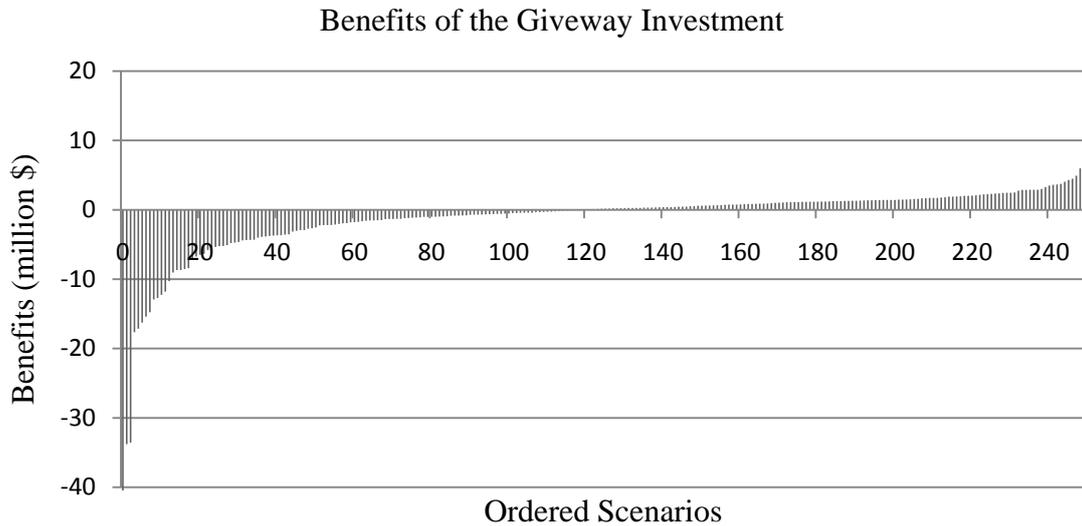


Figure 2: graph which shows how the value of one of the investments changes with the oil price scenario

The NZTA have chosen to invest in every project in the Kingsland area and the expected value of this portfolio of projects when simulated under a stochastic oil price was found to be \$-4,582,435.

The best EV portfolio that was selected matches the NZTA's portfolio (we assume the same expected oil price but do not allow for stochasticity). The value of the EV portfolio using only the expected demand scenario was \$11,153,700.

Simulating the portfolio under stochastic demand scenarios (EEV) we see that the expected value of this portfolio is \$-4,582,435.

The HN portfolio, where the best single portfolio was chosen, consisted of only a single investment from the options available. The value of this portfolio was \$318,682 with a capital cost of \$90,000.

The VSS solution has a value of \$4,901,117. This indicates that there is a significant benefit of performing a stochastic analysis.

4.4 Limitations

To fully understand our transportation investment selection model's potential it is important to address its limitations.

4.4.1 Induced Demand

While no induced demand assumption simplifies the problem it limits the model's ability to be applied to the real world, the example of a new public transport service could not be modelled under this assumption as all users of a new service would be induced.

4.4.2 Managing Risk

The current objective function is indifferent between situations with the same expected value. For example if you have a choice of two investments both with the same expected value of \$10 million, one returns \$10 million guaranteed, the other has a 50% chance of delivering \$0 and a 50% chance of delivering \$20 million the objective function is indifferent. From the perspective of a government who supplies transportation to its population it is unacceptable to have a situation where people won't

be using the transportation system, thus the objective function should not be indifferent between these scenarios.

4.4.3 Considering Existing Assets

Ideally the transportation network as a whole should be considered while selecting transportation investments. We only consider this existing network implicitly by allowing the model to choose not to make any investments. Instead it would be better to explicitly consider the value of these assets in the model.

Section 4.2.2 makes it clear that we need to manage risk, in order to do this we need to consider the risk associated with the existing transportation network, our existing assets. This is because the value of this existing network varies under demand so we should be choosing assets whose value changes counter-cyclically to our current investments.

5 Risk

The current formulation does not adequately consider risk within the objective function, this section explores how we can manage risk.

5.1 Conditional Value-at-Risk (CVaR)

VaR is a popular financial concept used to manage risk; it specifies the maximum amount you would expect to lose a given percentage of the time (φ). Instead of assuming a normal distribution VaR assumes that probability can be represented by a continuous distribution. A major problem with VaR is that it is not a convex function and thus difficult to optimise.

CVaR is the expected loss exceeding VaR, so if $\varphi = 0.99$ and $CVaR_\varphi = \$2,000,000$ then the mean loss of the 1% worst losses is \$2,000,000.

CVaR is generally considered a better measure of risk, as it considers the size of the tail of these worst losses and unlike VaR it also has the advantage of being convex and can therefore be easily optimised (Rockafellar & Uryasev, 2000). Therefore we explore how the CVaR technique can be applied to the transportation investment problem to manage the risk of these investments. The formula for CVaR is shown below where \mathbf{y} represents uncertainty, $f(\mathbf{a}, \mathbf{y})$ is the loss and N the number of scenarios being used.

Formula CVaR

$$CVaR_\varphi(\mathbf{A}) = \min_{VaR \in \mathbb{R}} VaR + N(1 - \varphi)^{-1} \sum_{i \in I} [f(\mathbf{A}, \mathbf{y}) - VaR]^+$$

This can be formulated as a linear programme as shown in the model below with a different objective function and additional constraints compared to the simple model.

Model CVaR Transportation Investment Selection

$$\text{Minimise} \quad VaR + \frac{1}{N(0.95)} \sum_{i \in I} u_i$$

Subject to

$$\begin{aligned} (1) \quad & u_i \geq 0 \quad \forall i \\ (2) \quad & \sum_{a \in A} (\beta_{i,a} - \gamma_a) x_a + VaR + u_i \geq 0 \quad \forall i \\ (3) \quad & \sum_{i \in I} \sum_{a \in A} \beta_{i,a} x_a - (1 + r) \sum_{i \in I} \sum_{a \in A} \gamma_a x_a \geq 0 \end{aligned}$$

Explanation

This formulation takes the CVaR formula and expresses it in a linear form, additional constraints (1) and (2) are necessary to enforce this objective function. Constraint (3) specifies that the solution must minimise risk subject to a minimum return.

5.2 Results

The optimal portfolio using the CVaR formulation was found to be the same portfolio using the HN technique (none of the investments apart from one). The portfolio has a CVaR of \$-318,682, which means the expected loss of these worst 5% of scenarios is actually a benefit of \$318,682.

The reason that we found this portfolio to be optimal is because the single investment which was chosen had a high return and no risk. Through selecting this investment we have minimised CVaR and satisfied our return constraint. We can see that the inclusion of an investment with a constant, high return will always result in the selection of this lone investment.

When we consider a more realistic interpretation of the transportation investment problem where we already hold assets whose value varies under demand (our current road network) we would expect a different result where a CVaR formulation could be useful.

6 Future Work

This project has primarily been an opportunity to begin to explore the possibility of using optimisation techniques to enhance the selection of transportation investments and hence leaves a number of avenues for future work.

6.1 Multistage Formulation

A multi-stage formulation creates a model in which there is not only a choice of whether to make an investment now, it also provides an option to delay the investment. A tool capable of doing this could be very useful; a real-world example is the planned alternate harbour crossing in Auckland. A multi-stage stochastic tool could indicate the optimal time to build this crossing.

6.2 Induced Demand

As discussed in section 3.1 the no induced demand assumption limits the ability to apply our model to projects that may in fact induce demand, such as a new service, or any large changes to the network.

Removing this assumption will make the model more viable for real-world applications; however it is likely that through removing this assumption non-linearity will be introduced into the model.

If we allow induced demand, the demand for each transportation investment will vary according to the project choices which have been made, this means that demand for each investment will vary within the model. This will add complexities to the model as the benefit of each investment under a different scenario would need to be found within the model. This, however, can be overcome by using piece-wise linearisation within the model to represent the profit according to demand for each investment. Piece-wise linearisation can be used to deal with non-convex profit functions (provided they are separable) so non-linearity in profit should not be a problem (Gabriel, Garcia-Bertrand, Sahakij, & Conejo, 2005).

The major problem with removing the no induced demand assumption comes when we consider how to assign demand to each investment within the model. In order to allow induced demand it is necessary to either formulate an additive approach to assigning demand, or allow multiplicative terms and somehow control their order and then introduce techniques to linearise these.

Thus the effect of removing this constraint should be explored and if non-linearity is introduced ways to overcome this problem should be examined.

6.3 Existing Assets

Section 5 suggests that we should consider assets that we already own. Therefore a method to quantify the value of existing transportation assets should be formulated. A possible starting point for valuing these assets is the equilibrium model of Auckland's transport network which should give an indication of how long people spend on roads.

If we were to combine this with the CVaR formulation given the dominance of roading in the current Auckland transportation portfolio we would expect the optimal investment portfolio to consist primarily of public transport investments.

7 Conclusion

This research has studied the transportation investment problem and has:

- successfully formulated a model to evaluate a portfolio of investments,
- applied this model to a subset of investments to see if we could select transportation investments better,
- established through the use of the VSS measure that evaluating the portfolio as a whole is better than evaluating a single transportation investment,
- shown that a CVaR formulation can be used to manage risk within transportation investments,
- identified the limitations of the model,
- and identified how the model could be enhanced in the future.

8 References

- Borley, C. (2008). Auckland leaves the car at home, traffic cut 3pc *NZ Herald*, from http://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=10517584
- Donovan, S. (2008a). *Managing transport challenges when oil prices rise* (No. 357): McCormick Rankin Cagney.
- Donovan, S. (2008b). *Price Forecasts for Transport Fuels and Other Delivered Energy Forms*.
- Gabriel, S. A., Garcia-Bertrand, R., Sahakij, P., & Conejo, A. J. (2005). A practical approach to approximate bilinear functions in mathematical programming problems by using Schur's decomposition and SOS type 2 variables. *J Oper Res Soc*, 57(8), 995-1004.
- Kennedy, D., & Wallis, I. (2007). *Impacts of fuel price changes on New Zealand transport*.
- Percy, A., & Harton, K. (2009). *The economic performance of transport projects: Using and interpreting the benefit/cost ratio to compare different transport initiatives*. . Paper presented at the Australasian Transport Research Forum 2009.
- Rockafellar, R. T., & Uryasev, S. (2000). Optimization of Conditional Value-at-Risk. *The Journal of Risk*, 2(3), 21-41.

A Rostering Integer Programming Model for Ambulance Staffing

Karl Ho

Home: (09) 443-3165

Mobile: 021547148

Email: cho046@aucklanduni.ac.nz

Abstract

The rostering problem is an important factor affecting the response times of units arriving at emergency calls within a region. A roster is defined to be a combination of time slots (periods in a day) that a staff member can be allocated to work. This project produced a rostering integer programming model for ambulance staffing. The main objective was to find a cost effective and efficient solution while maintaining a high level of performance around the region under investigation.

The project focused largely on the set partitioning problem which has proven to be effective when determining flight schedules for airline crew. By using SIREN (An emergency service simulation program developed by The Optima Corporation) to run simulations on historic data and an implementation of an approximate hypercube model (developed by Richard C. Larson of MIT), different performance measures were calculated and used as costs in the set partitioning problem.

With the use of the AMPL (A Mathematical Programming Language), a column-wise formulation of the set partitioning problem can be created to solve for a realistic model of the rostering problem for ambulance staffing.

Chapter One: Introduction

1.1 Overview

Emergency services around the world all have different staff members allocated to a base which needs to be given a roster which they work to. With the emergency services being the first line of contact when accidents occur, the response times and ability to reach all calls as quickly as possible becomes a large factor in how the rostering of the bases is to be organized, as delays in response times can be fatal. With this, the project aims to generate a cost efficient schedule for the emergency services whilst maintaining a high level of quality of service in terms of a specified performance measure for the district which it is providing medical service for. When we use the term quality in this report we are considering two types of quality. We have performance qualities and rostering qualities.

We mainly discuss performance measures as the performance quality, but here we will try and give an insight to rostering quality. In the emergency service occupation, staff members are exposed to dangers throughout their time at work. This forces health and safety organizations to define rules and regulations to govern the health and safety of all the staff members in hazardous work places. With this in mind we try to have different combinations of roster qualities which stay within the health and safety requirements whilst also maintaining a high level of performance quality.

The project will make use of the set partitioning problem and implementing quality into the set partitioning problem to organize the schedules of the bases involved in the problem. The use of simulation data collection and an approximate hypercube queuing model will be used to approximate real life conditions in the model to find costs corresponding to realistic events to provide a more accurate model.

Chapter Two: Literature Review

2.1 Introduction to Approximate Hypercube Queuing Model

An approximate hypercube model is used for approximating the performance of urban emergency service systems researched by Richard C. Larson of Massachusetts Institute of Technology.

There have been various studies of models simulating spatially distributed emergency service systems, such as ambulance, police and fire departments. While a lot of models have been developed for these systems there is still a large demand for an approximate method. The demand for an approximate method is in high demand due to the methods ability to generate performance measures which take considerably less time and computational space than an exact method. The approximate hypercube model can be used to analyze a number of different resource allocation problems including the districting problem, location problem and the work load balancing problem.

The districting problem focuses on how the region should be “partitioned into areas of responsibility (districts) so as to best achieve some level or combination of levels of service”, while the location problem focuses on how the “N response units be located or positioned while not responding to calls for service” and finally the workload problem looks into how the “units should be positioned and selected for dispatch in order to balance the workloads among units”. All these problems have an effect on our overall problem of maintaining a high level of quality in our schedules for ambulance staffing.

NOTE: With the districting problem a district for an ambulance is a region in which the ambulance will handle the calls if the ambulance is available; if this ambulance is not available then an out-of-district ambulance would be assigned to this call. If all N ambulances are busy then the call will enter a queue for dispatch later.

There are many situations where the use of an approximate solution would be enough. For example when the data has inaccuracies, the data would not be good for highly accurate models. Other factors that affect the collection of important data such as legal, political or administrative constraints may make gathering accurate data difficult or not possible; this makes precise estimates unnecessary. Also with the use of precise models a lot of memory is required to store the data which could also increase execution time which makes computational factors too costly.

The benefit of approximating the performance characteristics over an exact analytical model is that with N servers, it only requires N equations rather than 2^N equations which are required by the analytical model. The measures of performance that can be calculated from the model that we could consider in the project include the following: region wide mean travel time, response unit mean travel times and point specific mean travel times.

The main features of using the approximate hypercube model are that we can assume that the dispatcher has rank order of the preferred units to be dispatched from each of the geographic atoms. It is always assumed that the dispatcher will dispatch the more preferred available free unit. In addition we assume that the probability of dispatching the jth preferred unit to a call from a particular atom can be approximated to be proportional to the product of the utilization factors of the first preferred units and whether the unit is available. By considering the simple M/M/N queuing model we can determine the constant of proportionality which depends on j. Assuming a situation in which j servers are selected randomly without replacement from the M/M/N system. We can then generate N simultaneous nonlinear equations relating the N unknowns to the dispatch policy and the call rates from the various atoms. The N simultaneous equations are then solved iteratively which then produces estimates of the workloads of the units.

Chapter Three: Model Description

3.1 Scheduling Section

Before we begin the introduction to the rostering model for ambulance staffing we have to consider other models which have been created to deal with this similar problem. A similar model which involves staff rostering is the scheduling of airline crew, Ryan & Falkner (1988). The way that Professor David Ryan of the University of Auckland has approached this problem is with the set partitioning problem or set covering models in the scheduling applications he has looked into.

We start with the introduction of the set partitioning problem where we let I and P defines a partition of I if and only if $S_j \cap S_k = \emptyset$ and $\bigcup S_j = I$. We let c_j be the cost associated with S_j and C be the cost of partition P . Hence given S , find the minimal cost partition P of I . Below is a visual representation of the A matrix:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	...	Cn
1	1	1	1	1	1	1	1	1	1	1		1
2	0	0	0	0	0	0	0	0	0	0		1
3	0	0	0	0	0	0	0	0	0	0		1
4	1	1	1	1	1	1	1	1	1	1		1
5	0	0	0	0	0	0	0	0	0	0		1
6	0	0	0	0	0	0	0	0	0	0		1
7	1	1	1	1	1	1	1	1	1	1		1
8	0	0	0	0	0	0	0	0	0	0		1
9	0	0	0	0	0	0	0	0	0	0		1
10	1	1	1	1	1	1	1	1	1	1		1
...
m	0	0	0	0	0	0	0	0	0	0		1

Figure 1 - Set partitioning matrix representation

We can then let the columns form a matrix A . We then associate a 0-1 variable x_j with column j . Hence we can see that the general set partitioning problem has the form:

where A is an $m \times n$ matrix of zeros and ones and C is a vector of costs. The rows of the A matrix correspond to the different time slots of the day which need to have an ambulance allocated to. We had a constraint that forced all time slots of the day to have an ambulance base allocated to that time period of the day. This was to prevent having periods of the day where no bases are activated at all, as this would result in a lot of delays in responses to calls which can be fatal. The columns of the A matrix correspond to all the possible shifts that an ambulance staff member can take. The elements of A can be defined as $a_{ij} = 1$ if the task i is performed by shift j , and 0 otherwise.

From the above description of the columns of the A matrix we can see that to solve the full set partitioning problem can be quite difficult due to the number of variables that are involved in the problem. As the each column in the A matrix represents a different variable in the problem. But it may be possible to identify many combinations of shifts which are infeasible. (E.g. double shifts which cause long hours which may be against regulations provided by health and safety). Therefore it may be possible to find and remove many variables which are unlikely to be in the 'optimal' solution thus reducing the size of the A matrix dramatically. By doing so this the problem will have improved natural integer properties of the set partitioning problem. A solution to the problem will consist of a partition of the tasks by the shifts.

Airline crew scheduling in the aviation industry involves the "construction of an 'optimal' schedule of duties, each of which performs a sequence of runs. A run can be considered as the performance of an activity." Ryan & Falkner (1988). The difference in crew scheduling and ambulance staff rostering differs slightly here. In crew scheduling you have to consider variables like start time and start location and also the finish time and finish location. In the

ambulance staff rostering model we did not consider where the base is to be located or where the ambulance is located at time t , but we do consider the start times and finish times of bases being active. We keep in mind of the level of quality of the surrounding area that the base can provide its service area as a performance measure based on these start and finish times. When we describe a base to be active we mean that a base has a person or ambulance allocated to the base to receive calls. We initially assumed that when a base is active there is only has one ambulance or staff member active at the base.

The project begins to differ from airline crew scheduling work done by Professor Ryan when we begin to implement the effects that quality has on the ambulance staff rostering set partitioning problem. We begin this by introducing more variables into the A matrix which correspond to the quality of the schedule, provided that a particular arrangement of bases are activated. This then means that the A matrix can be broken up into two sections, a scheduling section and a quality section. As described before the scheduling section is all possible combinations of shifts that a staff member can take for a set number of time periods for set days of the week.

3.2 Quality Section

The quality section of the matrix consists of columns indicating which bases are active during a particular time slot for that day. The columns of the quality section provide all the possible combinations of active bases, hence the combinations of bases are found with a binary count of active bases during the different time periods. The rows of the quality matrix correspond to the same constraints as the scheduling section of the A matrix, the only alteration that is done with the entire problem is the right hand side vector needs to be changed to a vector of all zeros. Hence the new formulation of the problem:

where A is a matrix with the scheduling matrix and quality matrix attached end to end. Below is a preview of the extended A matrix:

	Col 1	Col n	Col n+1	Col n+2	Col n+3	Col n+4	Col n+5	Col n+6	Col n+7	Col n+8	Col n+9	Col n+10	Col n+p	
1	1	0	-1	0	-1	0	-1	0	-1	0	-1	0	0	= 0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	= 0
3	0	1	0	0	0	0	0	0	0	0	0	0	-1	= 0
4	1	0	0	-1	-1	0	0	-1	-1	0	0	-1	0	= 0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	= 0
6	0	1	0	0	0	0	0	0	0	0	0	0	-1	= 0
7	1	0	0	0	0	-1	-1	-1	-1	0	0	0	0	= 0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	= 0
9	0	1	0	0	0	0	0	0	0	0	0	0	-1	= 0
10	1	0	0	0	0	0	0	0	0	-1	-1	-1	0	= 0
...
m	0	1	0	0	0	0	0	0	0	0	0	0	-1	= 0

Figure 2 - Extended A matrix (with quality columns)

From the combination of the two sections which forms the A matrix, we can see the interaction between the scheduling section and the quality section. We can see that for a quality column to be selected, there has to be a coefficient of 1 in the same row of a corresponding scheduling column. This ensures that when the solver produces a solution, it selects a base that is active when a shift for the same base is chosen. This also works in the opposite direction in the way that a shift must be allocated to the base when a quality column with a coefficient of -1 in a respective row is selected for its particular performance measure. This ensures that if a shift was to be selected, the base for that shift must also be activated.

Having considered both the scheduling section and the quality section of the A matrix there are a set of constraints on the set partitioning problem which governs that only one shift is chosen per base and that one quality column is selected per time period, these are referred to as the gub constraints. these constraints forces the property that only one scheduling column

is selected for a particular base and only one combination of bases can be selected for a particular time period. Below is a preview of the A matrix with gub constraints indicated by the box outline:

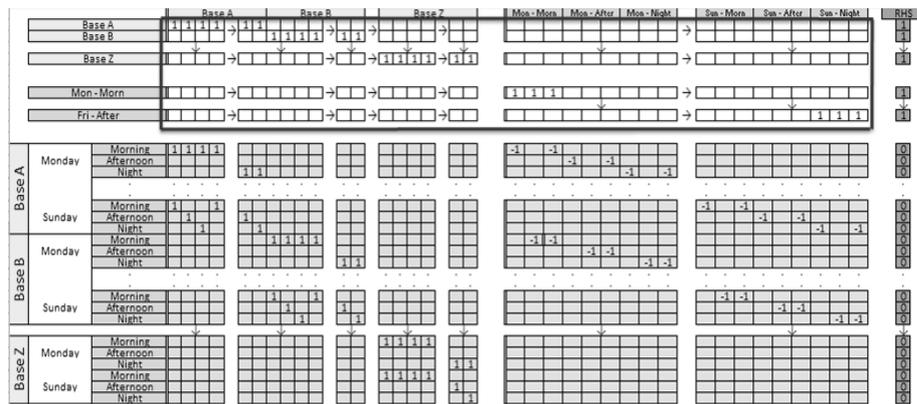


Figure 3 - Preview of entire A matrix with gub constraints

From the above figure we can see 3 distinct sections of the A matrix. We have the scheduling section which governs the scheduling constraints (the bottom left block), the quality section which governs the quality constraints (the bottom right block) and finally the gub constraints indicated by the white section and the box outline. We can see that the right hand side (RHS) in the matrix changes as it moves down the vector. The constraints that correspond to the gub have a value of 1 in the RHS vector as we only wish to have one column per base in the scheduling section and one column per time period in the quality section.

We can see from the problem description that the problem consists of minimizing the cost of the shifts chosen and maximizing the chosen performance measure. We can then conclude that we have a bi-objective problem which can both be minimized or maximized provided we have the correct weightings/multipliers applied to the problem. This requires different cost weightings to be applied to the costs of the columns of the A Matrix. We can see later on that if we consider the coverage of calls as a percentage in the A matrix we apply a negative value to the performance percentages to allow for a minimization of the objective.

Chapter Four: AMPL Model

4.1 Column-wise Formulation

The AMPL (A mathematical programming language) model for the project was based on scheduling model written in the AMPL book written by Fourer, Gay & Kernighan (2002). The method that was discussed in the book to solve a scheduling problem was a column-wise formulation. For certain types of linear programs it may be preferable to define the objective and the constraints before the variables. This is largely due to the fact that there may be a simple pattern in the definition of the coefficients down the constraints of a variable. This means that it will be easier to define the matrix of constraint coefficients (the A matrix) column-wise rather than row-wise.

4.2 Implementation of Column-wise Formulation

4.2.1 - Model Implementation

We start off the model by introducing a collection of subsets of shifts:

```

set SCHEDULES = 1..NSched;           #Number of schedules
set BASE = 1..Bases+1;              #Number of Bases + 1
set SHIFT_LIST{SCHEDULES} within PERIODS;

param shiftcost {SCHEDULES} default 0;
param qualitycost {1..QualRows, PERIODS} default 0;
param requiredSHIFT {BASES,PERIODS} default 0;
param qualitylist {1..QualRows,BASES};

```

We see from this that for each i in `SCHEDS`, the shifts that a person works on for schedule i is contained in the set defined by `SHIFT_LIST[i]`. We initially set the parameters `shiftcost` and the `qualitycost` to a default value of 0, this allows us to initialise the values later on in the model. Next we have `requiredSHIFT` which is the right hand side vector of 0's described in

3.2 Quality Section, such that for all `BASES` and `PERIODS` there is a 0. NOTE: This only covers the scheduling and quality sections of the right hand side vector. The gub constraints part of the RHS vector will be covered later in the model. Last we have defined the `qualitylist` which is a matrix indicating the different combinations of bases that can be active.

Before we start defining the A matrix we have to declare the objective and the constraints first, with the variables left out:

```

minimize Total_Cost;
subject to Shift_Needs{b in BASES, i in PERIODS}: to_come = requiredSHIFT[b,i];
subject to OneColPerBase{b in BASES}: to_come = 1;
subject to OneColPerPeriod{p in PERIODS}: to_come = 1;

```

Here we have defined the objective to be a minimization of the `Total_Cost`. We have three constraints for the problem; we have a `Shift_Needs` which is the constraint that for each b in `BASES` and i in `PERIODS` it must meet a required number of staff members provided by `requiredSHIFT[b,i]`. The next constraint is a gub constraint for the number of shifts per base which required that only one column per base was chosen. This was done due to the assumption that we only had one ambulance per base at a time. A similar constraint was applied to the quality matrix, where only one column per period was to be chosen. This prevents the model from choosing to have the column where the only base active is base A, the column where only base B is active and then choosing the third column of both base A and B being active at the same time as a possible solution to the problem. By selecting the column that covers both bases A and B it covers the columns where bases A and B are active by themselves.

We have a variable called `shift[j,b]` which represents a person being assigned to work the schedule j at base b . We have another variable called `quality[p,q]` where p is the period and q corresponds to the index allocated to a specific combination of bases being active. `quality[p,q]` represents a column being chosen for the period p with the base configuration q .

The coefficients of `shift[j,b]` appear in the var declaration as described by the method of column-wise formulation. In the objective it has a coefficient of `shiftcost[j]`. In the constraints, `SHIFT_LIST[j]` tells us that `shift[j,b]` has a coefficient of 1 in constraint `Shift_Needs[b,i]` for each i in `SHIFT_LIST[j]` and a coefficient of 0 in the other constraints. The coefficients of `quality[p,q]` appear in another var declaration, where in the objective `quality` has a coefficient of `qualitycost[q,p]` where p represents the periods and q corresponds to the number allocated to the combination of the bases. The variable `quality[p,q]` we define to be a binary value with its allocated costs as `qualitycost[q,p]`. We allocate a coefficient of -1 in locations where in `qualitylist[q,b] = 1`. In both variables `shift` and `quality` we have the gub constraints `OneColPerBase[b]` and `OneColPerPeriod[p]` where they are allocated a coefficient of 1.

```

var shift {j in SCHEDS, b in BASES} binary,
    obj Total_Cost shiftcost[j],
    coeff{i in SHIFT_LIST[j]} Shift_Needs[b,i] 1,
    coeff OneColPerBase[b] 1;

var quality {p in PERIODS, q in 1..QualRows} binary,
    obj Total_Cost qualitycost[q,p],
    coeff{b in BASES: qualitylist[q,b] = 1} Shift_Needs[b,p] -1,
    coeff OneColPerPeriod[p] 1;

```

The `obj` command gives the variable `shift` and `quality` their costs from their respective sets in the data file. The `coeff` command indicates where a coefficient of value 1 is placed in the variables `shift` for every i in `SHIFT_LIST` and b in `BASES`. A coefficient of -1 is placed in the variable `quality` for every b in `BASES` given that `qualitylist[q,b]` is equal to 1. The `coeff` for

oneColPerBase and oneColPerPeriod are values at 1 as these coefficients correspond to the gub constraints of the problem. The rest of the right hand side vector is constructed in the data file explained later on in the report.

4.2.2 – Data File Implementation

After having completed the model files we then have to construct the data files. The construction of the data files was done manually, but the use of VBA code made the construction of some of the data required easier.

We start off the data file by defining the periods of each day this is done manually for the size of our project:

```
set PERIODS :=      Mon1 Tue1 wed1 Thu1 Fri1 Sat1 Sun1
                   Mon2 Tue2 wed2 Thu2 Fri2 Sat2 Sun2
                   Mon3 Tue3 wed3 Thu3 Fri3 Sat3 Sun3;
```

Here we have used the value 1 to indicate morning shifts, 2 for afternoon shifts and 3 for the night shifts. Next the bases have to be defined, in our case since the bases are just general bases we have called them BaseA, BaseB, BaseC and BaseD.

```
set BASES := BaseA BaseB BaseC BaseD;
```

We then have to define the parameters of the problem, here we have to do calculations before hand to determine these values. As described earlier the number of combinations of schedules is determined by T^n where t is the number of time slots and n is the number of days in the roster. So in our example here we have 3 time slots and 7 days in the roster week. Therefore our initial NSched = $3^7 = 2187$ and the definition of the number of quality rows is defined by all the arrangements of bases being active – 1 so in our case the number of combinations of arrangements that 4 bases can take is 4^2 . We subtract 1 as we exclude the combination of no bases active as this would produce an unreasonable quality level in our model, and finally the number of bases is defined as an integer.

```
param NSched := 2187;
param Bases := 4;
param QualRows := 15;
```

We needed to construct the set which holds all the combinations of shifts that a person can choose from and apply it to a column-wise formulation template. The template will have the following format:

```
set SHIFT_LIST[1] := Mon1 Tue1 wed1 Thu1 Fri1 Sat1 Sun1;
set SHIFT_LIST[2187] := Mon3 Tue3 wed3 Thu3 Fri3 Sat3 Sun3;
```

As explained from above the set SHIFT_LIST provides the model with the positions within the matrix to place the coefficient of 1.

Next in the data file we have to initialise the right hand side of the model. But as we have already constructed the gub constraints with the .mod file we only have to construct the lower section of the right hand side vector. The lower section of the right hand side vector will be made up of all 0's. The lower section of the right hand side vector is made up of 0's because if we consider the constraints of the scheduling section we have placed positive 1's in its corresponding slot and when we consider the constraints of the quality section we have placed negative 1's therefore if you sum across the rows of the lower section of the A matrix it should sum up to a value of 0. We have set a parameter that the default value of the requiredSHIFT to be 0.

We created this as a parameter to allow for changes to the assumption that a single ambulance is at one base, when the base is considered to be active. For our current model we assumed this to be true, therefore we have requiredSHIFT equal to 0. If we were to relax this constraint we are able to formulate the requiredSHIFT in a matrix format.

After this the construction of the qualitylist was required to show the different combinations of the arrangements of bases in a city. A binary type count was used in the construction of the matrix.

After the construction of the qualitylist we construct the quality costs. With our initial formulation we have the periods across the top indicating the columns and we have the qualityrow number for the rows of the qualitycost matrix. This indicates that for a qualityrow of 1 there is a probability of 25% of the calls being met. NOTE: These are simple artificial values and do not represent actual data; the actual data collection is described in chapter 5. We can see from this that these values have a negative value this is due to the fact that the problem is a minimization problem and to make the values of higher percentages more appealing to the model we have multiplied the values by -1.

Below is a small indication of the format of the qualitycost matrix:

```
param qualitycost :
    Mon1  Mon2  Mon3  Tue1  Tue2  Tue3  wed1  wed2  wed3  Thu1  Thu2
    Thu3  Fri1  Fri2  Fri3  Sat1  Sat2  Sat3  Sun1  Sun2  Sun3 :=
1      -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25
      -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25 -0.25
      .
15     -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00
      -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00 -1.00
```

Next we initialise the shiftcost parameter, we are not able to attain realistic values for these costs. Therefore the costs that are provided to the model are artificial costs. Such that morning shifts are given a value of 1 afternoon costs are given a value of 2 and night shifts are given a value of 3. A summation of the shifts work periods are the values that are used in the model.

Below is a small indication to the parameter shiftcost:

```
param shiftcost :=
1      7
      .
2187  21
```

Chapter Five: Data Collection

The collection of realistic data required the use of SIREN predict. SIREN predict is a simulation tool used in the emergency medical services. It allows for different situations to be simulated, reviewed and optimized using a model which simulates real-life performances. The data collection process is broken up into two different methods, the simulation process and the use of a hypercube approximation model.

5.1 Simulation Data Collection

The first method we look at is the simulation data collection method; this involves reading in calls from historic data and running it through SIREN for a period of time. Since we are looking at a week the simulation period is broken down into a week and 8 different hours a day during the week. With the use of SIREN we are able to activate and deactivate bases and save each combination of active and deactivate bases as a scenario. After all the scenarios have been created we are able to run a simulation of all the different scenarios gathering average response times in output files which SIREN writes out to. We will use the average response times as a cost for the quality matrix which provides the model with more realistic data.

5.2 Hypercube Model Data Collection

The second method of data collection that was discussed was the hypercube approximation model. One of the main benefits of the approximate hypercube model was that we are able to generate a performance measure without running simulations. We are then able to use these

performance measures as the costs for the quality columns in our AMPL model. This is done in a similar way as the simulation data collection phase, in the way that we need to manually construct the matrix for the costs. The results will be shown in 6.2 Hypercube Data Results.

Chapter Six: Results

6.1 Simulation Data Results

```

AMPL Version 20080102 (x86_win32)
9063 variables, all binary
109 constraints, all linear; 70971 nonzeros
1 linear objective; 9063 nonzeros.

CPLEX 11.0.0: optimal integer solution; objective 370.5333333
258 MIP simplex iterations
0 branch-and-bound nodes
Total_Cost = 370.533

varname[j] [*] :=
  3 "Shift[1,'Base32',1]"      8885 "Quality['Med2',.21]"
  4 "Shift[1,'Base35',1]"      8900 "Quality['Thu2',.21]"
4374 "Shift[1094,'Base40',1]"  8915 "Quality['Fri2',.21]"
8745 "Shift[2187,'Base10',1]"  8930 "Quality['Sat2',.21]"
8760 "Quality['Mon1',.12]"     8945 "Quality['Sun2',.21]"
8775 "Quality['Tue1',.12]"     8959 "Quality['Mon3',.11]"
8790 "Quality['Wed1',.12]"     8974 "Quality['Tue3',.11]"
8805 "Quality['Thu1',.12]"     8989 "Quality['Wed3',.11]"
8820 "Quality['Fri1',.12]"     9004 "Quality['Thu3',.11]"
8835 "Quality['Sat1',.12]"     9019 "Quality['Fri3',.11]"
8850 "Quality['Sun1',.12]"     9034 "Quality['Sat3',.11]"
8855 "Quality['Mon2',.21]"     9049 "Quality['Sun3',.11]"
8870 "Quality['Tue2',.21]"
;

```

Figure 4 - Simulation Data Results

After the collection of the simulated data, we updated the data file in the AMPL model to create a more realistic model. The model was run through with a CPLEX solver and produced the solution shown on the left.

The optimal solution was 370.5333 with an average response time was 15.3 minutes

The combinations are as follows:

```

SHIFT_LIST[1]      := Mon1  Tue1  Wed1  Thu1  Fri1  Sat1  Sun1;
SHIFT_LIST[1094]  := Mon2  Tue2  Wed2  Thu2  Fri2  Sat2  Sun2;
SHIFT_LIST[2187]  := Mon3  Tue3  Wed3  Thu3  Fri3  Sat3  Sun3;

```

From the quality combinations we see that the solver has chosen quality index of 1, 2, and 12. These are shown below:

	Base10	Base40	Base32	Base35
1	1	0	0	0
2	0	1	0	0
12	0	0	1	1

6.2 Hypercube Data Results

After collection of the hypercube approximation costs for the quality matrix, we created a matrix of costs with these values and produced the following solution:

```

AMPL Version 20080102 (x86_win32)
9063 variables, all binary
109 constraints, all linear; 70971 nonzeros
1 linear objective; 9063 nonzeros.

CPLEX 11.0.0: optimal integer solution; objective 309.2
179 MIP simplex iterations
0 branch-and-bound nodes
Total_Cost = 309.2

varname[j] [*] :=
  2 "Shift[1,'Base40',1]"      8884 "Quality['Wed2',.11]"
  19 "Shift[15,'Base32',1]"    8999 "Quality['Thu2',.11]"
4357 "Shift[1090,'Base10',1]"  8914 "Quality['Fri2',.11]"
8748 "Shift[2187,'Base35',1]"  8932 "Quality['Sat2',.11]"
8754 "Quality['Mon1',.61]"     8947 "Quality['Sun2',.11]"
8769 "Quality['Tue1',.61]"     8966 "Quality['Mon3',.81]"
8784 "Quality['Wed1',.61]"     8981 "Quality['Tue3',.81]"
8799 "Quality['Thu1',.61]"     8996 "Quality['Wed3',.81]"
8814 "Quality['Fri1',.61]"     9011 "Quality['Thu3',.81]"
8826 "Quality['Sat1',.31]"     9026 "Quality['Fri3',.81]"
8841 "Quality['Sun1',.31]"     9041 "Quality['Sat3',.81]"
8854 "Quality['Mon2',.11]"     9056 "Quality['Sun3',.81]"
8869 "Quality['Tue2',.11]"
;

```

Figure 5 - Hypercube Data Results

For the hypercube data set the model has optimized to an objective value of 309.2. This is a summation of the costs for the chosen shifts and the costs of the quality columns. We found that the average response time was 12.4 minutes.

The CPLEX solver in AMPL found the following combination of bases and quality columns as its optimal solution:

```

SHIFT_LIST[1]      := Mon1  Tue1  Wed1  Thu1  Fri1  Sat1  Sun1;
SHIFT_LIST[5]      := Mon1  Tue1  Wed1  Thu1  Fri1  Sat2  Sun2;
SHIFT_LIST[1090]   := Mon2  Tue2  Wed2  Thu2  Fri2  Sat1  Sun1;
SHIFT_LIST[2187]   := Mon3  Tue3  Wed3  Thu3  Fri3  Sat3  Sun3;

```

	Base10	Base40	Base32	Base35
1	1	0	0	0
3	1	1	0	0
4	0	0	1	0
6	0	1	1	0
8	0	0	0	1

Chapter Seven: Discussion

From the simulation results we can see that the results are what we previously expected in terms of the shift combinations. When we constructed the initial excel formulation we concluded that the solver will find shifts with the least number of changes, and this was proven with the simulation model. The shifts that are allocated to the bases are constant throughout the week. By constant we mean that there are no changes to the shifts pattern. Though this may cause problems for the company as staff members may get tired later on due to the lack of change in their rostering. Problems like these may occur and one of the ways of fixing this problem is to have communication with the emergency services to determine which shift combinations are more appealing to its members. This will allow us to maintain low costs and a high level of performance whilst maintaining high staff morale (which can correspond to shift quality). Though this may be a solution to a simple problem, this may cause other types of problems in the model because including options such as preferred shift combinations forces more constraints on a model which can expand dramatically in size.

The hypercube approximation model produced results which were different from the simulation models, but the difference in shift combinations between the two solutions wasn't that significant. We could see that both the simulation model and the hypercube model both found solutions where the numbers of changes in time periods that an ambulance staff member would work in a shift were quite small. This is a positive as many changes in a shift tend to be a bad factor for people working in occupations that require shifts.

Chapter Eight: Conclusion

From the work that has been provided above, we were able to construct realistic rosters with the help of the approximate hypercube method and a simulation method in the data collection process. We can conclude that we have constructed an integer programming model for ambulance staff rostering that is able to find feasible solutions. The model was able to find suitable shifts for the bases to generate an optimal solution given a selected performance measure and cost for the shifts.

Acknowledgements

I would like to thank Dr. Andrew Mason for his knowledge and guidance throughout this project. The help Andrew has provided was a large factor in the success of this project. Without his knowledge and help, the completion of this project would have been a lot more difficult.

I would also like to thank The Optima Corporation for their feedback and the opportunity to work on a project that we both are interested in. Without their help and guidance this project would not have been possible.

References

- D.M Ryan and J.C Falkner. 1988. "Integer properties of scheduling set partitioning models". European Journal of Operations Research.
- R.Fourer, D.M. Gay and B.W. Kerningham. 2002. "A Modeling Language for Mathematical Programming". Brooks/Cole Publishing Company.
- R.C. Larson. 1975. "Approximating the Performance of Urban Emergency Service Systems". INFORMS, Volume.23 (No.5)
- R.C. Larson. 1973. "A Hypercube queuing model for facility location and redistricting in urban emergency services". The New York City Rand Institute.

Optimisation of Mould Filling Parameters during Compression Resin Transfer Moulding Process

By Wing Ki Kam, Department of Engineering Science, The University of Auckland
Email: wingkikam@gmail.com

Abstract.

Compression Resin Transfer Moulding Process is a manufacturing method for composite materials. Composite materials have high potentials but require reduction of manufacturing costs to gain greater dominance in industries. One way to reduce the cost of manufacture is to improve the manufacturing performance. This research aims to find out the optimal process parameters that produce the best performance.

To this end, the three dominant process parameters, resin injection pressure, mould height during resin injection and velocity of mould during compaction, are optimised. The objective of the optimisation is to minimise both the maximum force requirement on the pressing machine and the total process time.

The maximum force requirement and total process time are evaluated from different combinations of the three process parameters through the use of SimLCM, a program that utilises finite element method and simulates the process. Genetic algorithm, a heuristic solution searching method is used to optimise this non-convex and non-linear problem.

At the end of the research, a decision making assistant tool is developed. It is able to find the best combinations of process parameters, for most production models. It is also capable of aiding manufacturers with machinery upgrade and job scheduling problem solving.

1. Introduction

In modern day manufacturing industries, businesses are constantly looking for ways to gain an advantage to remain competitive. Optimisation of the current manufacturing processes thus becomes crucial to the survival of their products and businesses.

Advantages gained from optimising manufacturing processes include increased rate of production and reduced cost of manufacture. These in turn translate into more profits. As a result, most manufacturers want to find ways to improve their production methods.

This study focuses on the Compression Resin Transfer Moulding process for the manufacture of composite materials. To realise the full potential of the Compression Resin Transfer Moulding Process, manufacturing parameters are required to be optimised. An optimal combination of parameters will minimise the maximum clamping force requirement and process time. It will then result in minimised manufacturing costs and maximised production efficiency.

While this study primarily focuses on the Compression Resin Transfer Moulding process, the techniques and methods developed in the research can be widely adopted in many other manufacturing processes.

The aim of this project is to develop a tool to provide relevant information for making manufacturing decisions. Such decisions include but are not limited to optimal process parameters and machinery upgrade. To that end, a software package, which combines the a simulation package of the CRTM process with a multi-objective evolutionary algorithm to optimise the parameters is developed.

The results from the simulations and optimisations are then analysed and interpreted. To verify the practicality of the methodology used, a three dimensional test case is simulated and optimised using our method.

2. Problem Definition

2.1 Compression Resin Transfer Moulding

Compression Resin Transfer Moulding (also called Injection Compression Moulding) is a popular composite material manufacturing process. This process is most noted for its ability to create products fast while maintaining a low compaction force and injection pressure under operating conditions.

The Compression Resin Transfer Moulding Process can be divided into five steps.

Step One: Preform Manufacture and Lay Up

Reinforcement materials are manufactured into a mat like preform and placed in the mould.

Step Two: Initial Dry Compaction

During this phase, the reinforcement preform is compacted inside the mould through application of a force to the mould. The degree of compaction is specified by the manufacturer.

Step Three: Resin Injection

After dry compaction, a set volume of liquid resin is injected into the mould under pressure.

Step Four: Final Wet Compaction

In this final compaction phase, the resin is driven through the whole mould to completely wet the reinforcement fibres and fill up any gaps. It is done by closing the mould downward at a constant velocity to the final required composite material thickness.

The Compression Resin Transfer Moulding can be summarised in Figure 2.1.

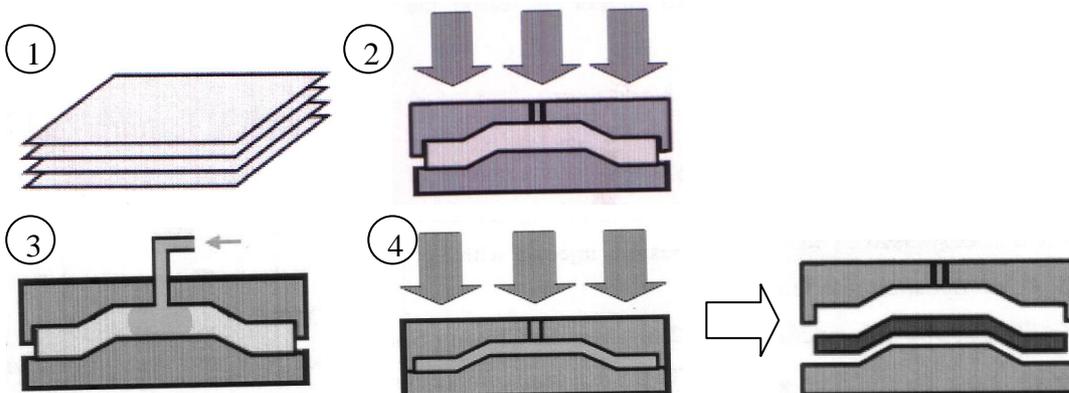


Figure 2.1 Compression Resin Transfer Moulding Process

2.2 Design Variables

Three process parameters are identified to be the dominant factors affecting the clamping force requirement and process time and chosen to be the only design variables in this research. These are the Injection Height (H_{inj}), Injection Pressure (P_{inj}) and Wet Compaction Velocity (V_{wet}). These parameters can be easily changed via control interfaces in the process machines.

Injection Height (H_{inj})

Injection Height is the height of the mould after the dry compaction phase and before the resin injection phase. Injection Height is illustrated in Figure 2.2.

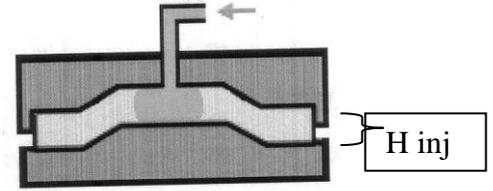


Figure 2.2 H_{inj}

Injection Pressure (P_{inj})

Injection Pressure is the operating pressure used to inject the resin into the mould and through the preform.

Wet Compaction Velocity (V_{wet})

Wet Compaction Velocity is the velocity of the mould as it is closed to the final height during the wet compaction phase.

These three process parameters are the variables that are optimised in this research. Other parameters, such as the geometry, injection node location, material properties are kept constant in each model used.

2.3 Design Objectives

In this research, there are two objectives to be minimised. They are the clamping force requirement and process time. Together they represent the manufacturing process performance. Both objectives are dependent on the three design variables.

The clamping force requirement is illustrated in figure 2.3. It is the maximum force a compaction machine uses to force the preform to its desired thickness. In real practice, the force requirement is limited by the compaction machine's capability. That is the force that the machine is capable of producing. Force output also relates to power consumption. To produce higher force output, more power is required. Generally a lower value for this objective is desirable.

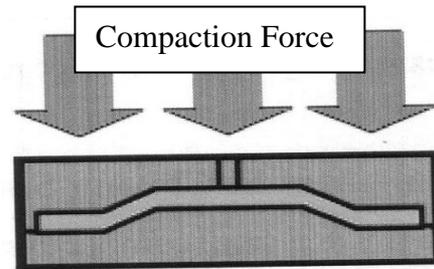


Figure 2.3 Illustrating clamping force

Process time is the time required to complete the manufacturing process. Process time has a significant effect on power consumption. The longer the process, the more power is used. Process time is also very important in job scheduling for manufacturers. Shorter process times for different jobs make scheduling them to meet due dates or minimise late jobs much easier. Therefore the process time objective needs to be minimised.

The two objectives functions are

Minimise: Objective $f = F_n(H_{inj}, P_{inj}, V_{wet})$

Minimise: Objective $t = G_n(H_{inj}, P_{inj}, V_{wet})$

Here, f stands for force and t stands for process time.

Clearly, this is a multi-objective optimisation problem. Ultimately it is hoped that both the clamping force requirement and process time are as low as possible. However, it can be very difficult to obtain a good solution because the two objectives counter-balance each other. When the process time is adjusted to low, the force requirement inevitably increases, and vice versa.

2.4 Previous Research

In Sam Na's master's research [6], a global optimal solution to Compression Resin Transfer Moulding problem was sought. The problem is formulated to optimise the Injection Height, Injection Pressure and Mould Closing Speed.

The problem is found to be non-convex, which indicates that there are multiple local optima. As a result, typical numerical optimisation methods are not suitable for the problem, because these methods terminate when they find an optimum, regardless of whether it is local or global.

To overcome the problem, an exhaustive search technique is used to locate the global optimal solution. Exhaustive search guarantees an (approximate) global optimal solution but are very time consuming and therefore not useful for normal industrial use. Na tested ten scenarios with three different geometries using the exhaustive search method.

3. Objective Evaluation and Test Models

3.1 SimLCM Software

To simulate the Compression Resin Transfer Moulding process, SimLCM Software is used. SimLCM was developed at the University of Auckland. It utilises the finite element analysis.

The main motivation to use this finite element analysis software is to reduce the cost and time of the experiments to find new solutions.

Finite Element Analysis is a numerical method used to find an approximate solution of partial differential equations or integral equations. The finite element method divides a complex problem into smaller problems in a smaller subspace. Then it solves each sub-problem one by one. All sub-problems solutions are linked together with common boundary conditions. As a result it is able to solve a problem that is otherwise too huge for analytical methods.

SimLCM was developed to simulate the Resin Transfer Moulding or Compression Resin Transfer Moulding processes. It was created to predict clamping forces and stress distributions within the mould throughout the manufacturing process.

3.2 Test Models

A simple testing model is created for the purpose of testing the simulation software and optimisation techniques.

It is a rectangular shaped, flat planar plate with liquid polymer resin fluid flowing in one direction. The flow moves from the injection nodes on the left hand side to the vents on the right hand side.

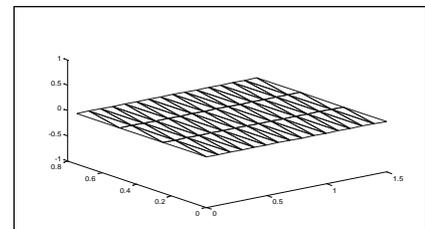


Figure 3.1 Thin Plate Mesh

The dimension of the plate is 1.5m x 0.75m with a final product thickness of 3mm.

A model of a real practical object is needed to show that the simulation and optimisation techniques are suitable for real industrial practice and not only for laboratory research.

In this research, a safety helmet is chosen to be the model for this purpose.

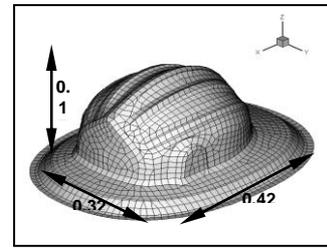


Figure 3.2 Fire Helmet

4. Optimisation Process

4.1 Multi-Objective Optimisation

In order to choose an appropriate method to optimise a problem, the problem's properties need to be examined. This research problem has been defined to be a multi-objective non-convex optimisation problem above.

The two objectives of this problem, clamping force requirement and process time are highly correlated. When process time is to be reduced, the clamping force requirement must increase to achieve it, and vice versa.

This is an optimisation problem where both objectives are minimised by adjusting the three design variables, Injection Height, Injection Pressure and Wet Compaction Velocity. This minimisation is illustrated in Figure 4.1. Plotting Force Requirement on the y-axis and Process Time on the x-axis, points on the graph represent possible solutions.

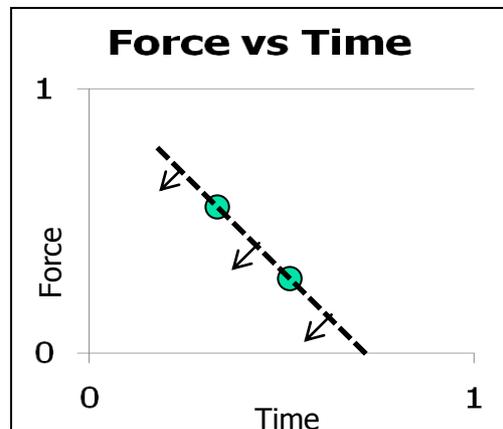


Figure 4.1 Multi Objective Optimisation

In order to find the optimal solution, it is essential to be able to identify good solutions and bad solutions. In multi-objective optimisation, a good solution is called an Efficient Solution, and a bad solution is called an Inefficient Solution. Any optimisation method identifies and keeps efficient solutions while neglects or eliminates inefficient solutions

An inefficient solution is a dot on the graph that is dominated by another dot. A solution is dominated if a solution exists that has both lower force and time. An efficient solution and an inefficient solution are illustrated in Figure 4.2.

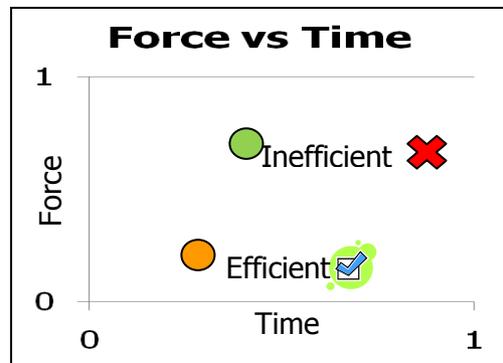


Figure 4.2 Efficient and Inefficient Solution

4.2 Optimisation Difficulties

The Compression Resin Transfer process is a complex manufacturing process. The elastic material model, fluid flow model, viscosity and permeability factors all contributed to the complicated objective functions. In fact, the clamping force and process time are calculated from the design variable values using the SimLCM software. Hence the functions to be optimised are not known analytically.

The research problem is non-convex. It means that multiple local minima exist in the objective functions. This can be seen in Figure 4.3. Point “b” is the global optimal point. Point “a” and “c” are called local optimal points.

This is critical to our selection of optimisation method. Most iterative methods are not suitable for non-convex problems, because they terminate and declare their solutions are optimal when an optimum point is reached, regardless of it is whether local or global. Moreover they require information about the function to be optimised (such as gradients) which are not available in our case.

Therefore in order to find the global optimum solution, special optimisation methods are needed.

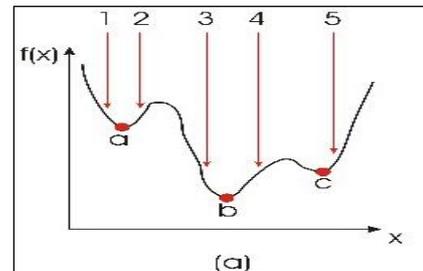


Figure 4.3 Non convex function

4.3 Genetic Algorithm

Genetic Algorithms are based on the principles of evolution. Evolution is a biological process of constant adaptation to the environment where only the fittest species survive and the less well adapted ones are eliminated.

In a genetic algorithm, the principles of evolution are incorporated and form the optimisation method that is described here. Through the observation of nature, genetic algorithms hope to use these principles to improve solutions in the optimisation process.

Genetic algorithms also have a fine balance between exploitation and exploration of good results. Reproduction of good populations represents the exploitation of good solutions, as it encourages the keeping of already good properties to try to use this experience to produce even better populations of solutions. Exploration is best implemented by including mutation. The randomness of mutation allows genetic algorithm to escape the (local) optima of an elitist gene pool and explore potentially better optima.

A genetic algorithm can be outlined in the following steps:

Step one: Initialisation of population

To start the optimisation, there needs to be a population of random solutions. A population of solutions is similar to the population of species. Each individual in the population represents a solution. In this study, an individual in the population represents a combination of the three design variables. To initialise the population, initial solutions are generated through randomly adjusting the three design variables within the allowed limits.

Step two: Selection and survival of population

Each individual of the population is evaluated and compared to each other. Inefficient solutions are identified and eliminated.

Step three: Reproduction

The good solutions that remain, are grouped into pairs. These pairs are called “parents”. A new population is then created by combining properties of the parents. This process is called “crossover”. In this study, it is by randomly combining the three design variables from the parent population. This process can be described in Figure 4.4.

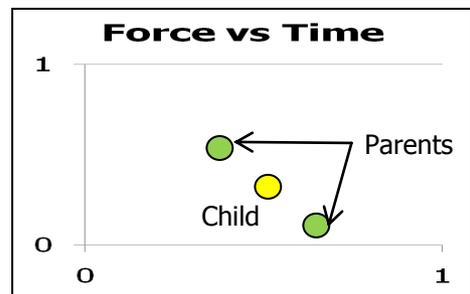


Figure 4.4 An idealistic reproduction result

Step four: Mutation

To simulate the natural phenomenon, new populations are generated from random alteration of the remaining good solutions. In this study, it is done by randomly adjusting the three design variables of a good solution. This is illustrated in Figure 4.5.

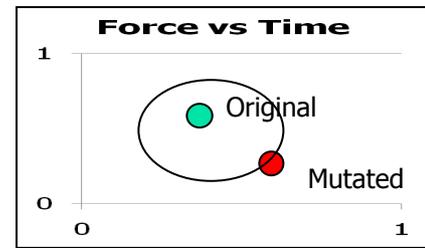


Figure 4.5 An idealistic mutation result

Step five:

Back to step two. As the algorithm iterates, it is hoped that the population of solutions are getting better and better.

Genetic algorithms are suitable to use in this research problem because they can deal with two main difficulties of the optimisation problem. They do not require derivative information but only objective function evaluations, which can be complex. Genetic algorithms have also been shown to work well with non-convex problems [4].

A genetic algorithm's ability to explore a wider search area by using mutation is also well suited to the research problem. It is able to escape local optima and keep looking for better optima.

As a result, genetic algorithm is determined to be the best optimisation method to use in this research study.

4.4 NSGA 2 Source Code and Modifications

The genetic algorithm used in this study, is called NSGAI [REF]. A source code of the algorithm is obtained from Dr. Kalyanmoy Deb, Indian Institute of Technology, Kanpur, India.

Objective Evaluation

The NSGAI algorithm needs to be adapted to call the SimLCM software for function evaluation. This is done by changing code that represents simple mathematical functions to a shell script command call to run the SimLCM.program executable.

Linking NSGA2 feedback to SimLCM input

Once the genetic algorithm has performed the steps selection, reproduction and mutation (i.e. computed a new set of solutions or combinations of values of the three design variables), the new population needs to be evaluated (i.e. the objective function values of solutions need to be computed). To do this output from NSGAI needs to be written to the SimLCM input text files.

Linking SimLCM output to NSGA2 evaluation collection

To pass on the evaluation of the two objectives from SimLCM to the NSGA2 for selection, the two objective values need to be written to the SimLCM output text file. The process time objective is the time evaluation at the end of a SimLCM simulation run. The clamping force requirement objective is the maximum z direction total force throughout the simulation run.

Store population plot every generation

In order for the outputs of NSGA2 to be more useful, the ability to produce population plots is extended. Instead of storing only the final population plot, the program is instructed to store each population plots as new file as they are generated.

Population selection and error handling

SimLCM outputs errors rarely when evaluating objectives from the changing variables. In cases of error occurring during simulation, it is detected by checking the Data_OutputErr.txt text file after each simulation run. The objective values are then forced to a negative number if any error is detected. NSGA2 code will then ignore the negative objective functions as it is not allowed to include infeasible solutions.

5. Optimisation Setup and Results

5.1 Optimisation Parameters

Population Size

The size of population affects the efficiency of the optimisation. Sixteen is determined to be the most suitable population number. With sixteen individuals, NSGAI allows sixty different combinations of parents, which is sufficient for effective crossover. Sixteen final solutions are also a reasonable number for decision making.

Number of Generations

The number of generations affects the efficiency and quality of the final result. In this research, the number of generations is set to be thirty. It allows for plenty of generations for convergence to occur, but still takes reasonable time to complete the optimisation for industrial use.

Crossover rate and mutation rate

They mainly affect the convergence rate. A higher crossover rate encourages the optimisation to converge quicker, however also risks terminating on a local optimal point. A higher mutation rate encourages the optimisation to explore a wider search area to locate more optima, however if set too high, convergence will occur too slowly.

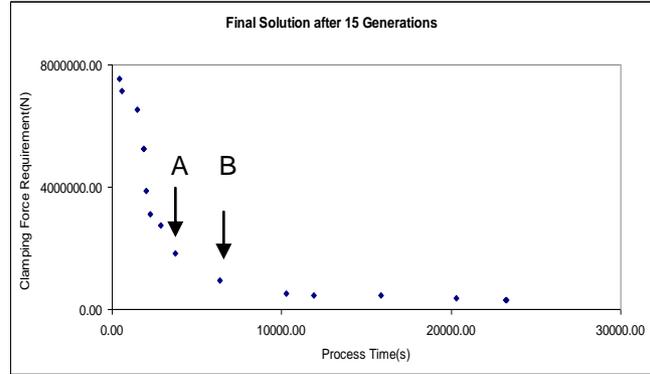
Choosing the best crossover rate and mutation rate is an optimisation problem on its own. However as this study is primarily focused on the optimisation of the mould filling process parameters, it is determined to be the most suitable to leave the rates at the default values. The default values have been used in a lot of previous researches in a wide range of topics, so it can be expected that they are likely to work well for this research, too.

5.2 Optimisation Results

The optimisation of mould filling parameters for the fire helmet model is completed using the modified GA code after 15 generations. The objective values of the final population are shown in Figure 5.1.

The two solutions A and B identified in Figure 5.1 appear to be the most desirable ones, it is because they offer the best balance between clamping force and process time.

While all other efficient solutions are no worse than solutions A and B, it is important to note that clamping force and process time are limited by the capability of the pressing machine and job due time, respectively, in the real world.



	Solution A	Solution B
Injection Pressure(Pa)	401566.80	340915.60
Injection Height(m)	0.002223	0.002341
Compaction Velocity(m/s)	0.000004	0.000002

Figure 5.1 Final Optimisation Result

5.3 Other Applications

Other than finding the best mould filling parameters for the manufacturing process, the tool developed in this research is also able to provide information for everyday management.

Machine Upgrade Problem

One of the situations a manufacturer may face is the question of whether they should upgrade their pressing machine. A better pressing machine will be able to produce composite material faster, increasing the production rate and therefore their competitiveness. A technique to solve this problem is demonstrated in Figure 5.2. Given upgrading to a new machine will increase production rate by a certain percentage. The upgrade is justified if the increase of production rate is worth the investment.

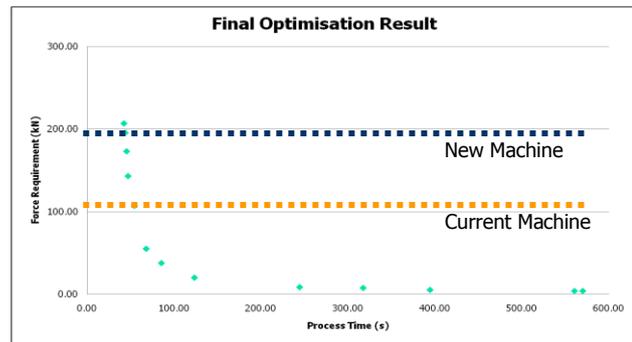


Figure 5.2 Using result in upgrade problem

Job Scheduling

Another problem manufacturers often face is job scheduling, and the utilisation of machines and minimising job lateness. Using the efficient solutions from the optimisation result, a manufacturer can choose from sets of mould filling parameters to minimise downtime or to increase production rate to meet a deadline, hence allowing the manufacturer greater flexibility in scheduling each job.

6. Conclusion

This project has successfully developed a software package that optimises the Compression Resin Transfer Moulding process and gives essential information for management decision makings. We summarise key findings and possibilities for future work.

Key Findings

- Analysis of the three variables' effects on the two objectives has shown that the optimisation problem is non-convex.
- Genetic Algorithm is a competent tool for use in a multi-objective non-convex optimisation problem where objective function values are only available through simulation.
- Final results from the optimisation have proven to be very useful information for everyday manufacturers' management and decision makings.

Future Work

- Simulation for a practical object takes too long to be useful in the manufacture industries. The simulation can be optimised and speeded up in future research.
- This project only focuses on the three process variables. Future study can include more parameters.
- There are other heuristic method besides genetic algorithm and can be tested and their performance compared with the genetic algorithm.
- Several parameters for the genetic algorithm can be optimised to give faster convergence result.

7. References

- 1) Ellyar, D. (2000). Putting it together – the science and technology of composite materials. Australian Academy of Science.
<http://www.science.org.au/nova/059/059key.htm>
- 2) Astrom, M.M. (2000). Manufacturing of Polymer Composites, London: Chapman and Hall.
- 3) Bicketon, S. and M. Abdullah (2003). Modelling and evaluation of the filling stage of injection compression moulding. Composite Science and Technology 63:1359-1375
- 4) Deb, K. (2002) Introduction to Evolutionary Multiobjective Optimisation. Publisher? Website?
- 5) Deb et al NSGAI paper
- 6) Lin M., M. Murphy, and T.H. Hahn (2000). Resin Transfer Moulding Process Optimisation. Composites Part A 30:361-71
- 7) Na, S. (2008) Global Optimisation of Mould Filling Parameters during the Constant Speed Injection Compression Moulding Process. Master Thesis, University of Auckland
- 8) Riche, R., A. Saouab., and J. Breard (2002). Coupled compression RTM and composite layup optimisation. Composites Science and Technology.63: 2277-2287.
- 9) Ropars, A. (2006). Optimisation of the Moulding Filling During Liquid Composite Moulding. Summer studentship report. University of Auckland.

Capacity Planning For Process Industries

James O. Kirch
Department of Engineering Science
University of Auckland
New Zealand
jkir053@aucklanduni.ac.nz

Abstract

A capacity planning model is an optimization model that can be used to assist managers in making optimal strategic capacity-related decisions over a time horizon. Capacity planning models are of particular interest to those within process industries where capital items are typically of a large scale, involving large costs and risks. In a deterministic setting, these are time-staged mixed integer programming models. When future uncertainties can be modeled by random variables, these models become multi-stage stochastic integer programming problems, a problem class that is extremely difficult to solve. We show how Dantzig-Wolfe decomposition methods can be used to efficiently solve multi-stage stochastic integer programming problems in this setting for realistically-sized models of capacity planning in the pulp-and-paper industry.

Key words: Dantzig-Wolfe Decomposition, Pulp and Paper, Capacity Planning, Multi-Stage Stochastic Integer Programming.

1 Introduction

A capacity planning model (CPM) is a decision-support system that is used to compute optimal or near-optimal planning decisions regarding the operational capacity of a given set of capital items. Many items (e.g. machines, network links) have an associated operational capacity that defines an upper bound on the production or operational usage of that item. A CPM simply allows for strategic capital planning, with the objective of increasing overall profitability or efficiency over a time horizon. CPMs are of particular importance to the process industry, as capital expenditure decisions can often be extremely costly and time-consuming to implement.

Although literature on capacity planning has long recognized the importance of uncertainty (see e.g. Manne (1961) and Eppen, Martin & Schrage (1989)), practical problems often involve complicated constraints that make stochastic versions very large scale mixed-integer programs. Capacity expansion decisions have therefore either been looked at using simplified real-option models (see Brealey and Meyers (1996)) or treated as deterministic mixed-integer programming problems.

A stochastic integer programming model allows for a more realistic representation of the decision process, and presents managers with an optimal policy of capacity-related decisions, whilst allowing them to also investigate the financial risk or variability associated with the given venture.

In this paper, we attack this problem using Dantzig-Wolfe decomposition and delayed column generation. This approach was first studied in capacity planning by Ahmed, King & Parija (2003) in a model where inventory linked consecutive periods. In our setting of process industries, where capacity expansion takes years to implement, each planning stage is of significant duration (e.g. a year), which means that the inventory transferred from stage to stage is small compared with the production during the stage. The linking variables instead are the capacity levels arising from expansion decisions in previous stages. This model then falls into the class of problems recently studied by Singh, Philpott & Wood (2009), which enables us to adopt their approach.

We show that these methods can be used to efficiently solve multi-stage stochastic integer programming problems for realistically-sized models of capacity planning in the process industry.

We illustrate this with some computational experiments on a capacity planning model from the pulp-and-paper industry.

In the remaining section of this chapter, we introduce the pulp-and-paper CPM known as SOCRATES. In the next section, we extend SOCRATES to a stochastic framework and investigate the applicability of decomposition methods in its solution process. In the closing chapters, we briefly discuss several key observations and state our conclusions.

1.1 Socrates

SOCRATES (Strategic Optimization of Capital and Resource Allocations Traversing Extreme Scenarios) was a model developed by Everett, Philpott & Cook (2000) for Fletcher Challenge’s Canadian paper milling operations (FCCL).

SOCRATES was designed to cover all aspects of FCCL’s pulp and paper supply chain; from sourcing raw materials from suppliers, to final delivery of the product to market. Initially, it encompassed a 10 year deterministic planning horizon, over which all parameters and costs were assumed to be known with certainty. It also included *capital constraints* which incorporated capacity contraction and capacity adaptation decisions into the model formulation. The addition of these constraints put SOCRATES firmly within the broad family of models known as CPMs. To be more specific, SOCRATES was a discrete-time multi-stage deterministic capacity planning model.

FCCL’s Canadian operations at the time consisted of two paper mills (‘Crofton’ and ‘Elk Falls’), each with a set of distinct paper machines. Yearly demand for different paper-related products was forecast for five different international markets (demand forecasts for each year were fully deterministic) and used as data for the model.

SOCRATES introduced the important concept of a *capital array* associated with a given paper machine (which will become a key concept in the stochastic model). Put simply, a capital array defines what capital items are needed to be purchased before a certain grade of paper can be produced on a given machine. An example is shown below in figure 1. Here, without any further capital expenditure this machine can produce only paper grade 1. Capital item 2 must be purchased and installed before it can produce grade 2, and so on.

Figure 1 - Capital Array Table displaying items needed for production

	Paper Grade 1	Paper Grade 2	Paper Grade 3	Paper Grade 4
Capital Item 1			X	X
Capital Item 2		X	X	
Capital Item 3			X	X
Capital Item 4				

The addition of the concept of capital items allows for *capacity adaptation* decisions. Essentially through capital expenditure the manufacturer is allowed to diversify its production capabilities and branch out into producing alternate grades of paper. For example, in the case of declining demand for newsprint, through *capacity adaptation* decisions it could switch to producing bleached paper, enabling higher overall profits. It is important to note here that capital items must be bought a year before production can begin (i.e. there is a set up time, as well as a cost).

SOCRATES also included the option of shutting down machines and mills, effectively adding *capacity contraction* decisions to the formulation. To shut down a mill, one must first shut down all machines associated with it. Each machine and mill has fixed operating costs associated with it, and hence it may be optimal to shut down one or more machines in certain situations. Obviously, a machine is not allowed to produce products if it is shut down.

The model incorporated a 10 year planning horizon, as well as an 11th artificial year to record slack binary variables in shutdown constraints. All machines and mills were assumed operational in the first year, and forced to be shut down (if they weren’t already) in the artificial year. The

SOCRATES model is described in detail in Everett *et al* (2000). Here we give an abbreviated description.

Set Definitions:

- Z Paper Mills – indexed by **z**
- M Paper Machines – indexed by **m**
- C Capital Items – indexed by **c**
- J Paper Grades – indexed by **j**
- T Planning Years – indexed by **t**
- M{z} – the set of machines **m** associated with mill **z**
- K Consumer Markets – indexed by **k**
- R Raw Materials – indexed by **r**
- Q Suppliers – indexed by **q**

Variable Definitions:

- x_{mjt} – Number of tonnes of product **j** produced on machine **m** in year **t**
- w_{mrt} – Total amount of raw material **r** used by machine **m** in year **t**
- v_{mrqt} – Total amount of raw material **r** supplied by supplier **q** to machine **m** in **t**
- y_{mjkt} – Total amount of product **j** shipped from machine **m** to market **k** in year **t**

Binary Variables:

- σ_{mcs} – Set to 1 if capital item **c** is bought for machine **m** in year **s**
- α_{ms} – Set to 1 if machine **m** is shut down in year **s**
- μ_{zs} – Set to 1 if mill **z** is shut down in year **s**
- ρ_{mjt} – Set to 1 if paper grade **j** is produced on machine **m** in year **t**

Parameters:

- e_{mjt} – Annual production capacity of machine **m** producing product **j** in year **t**
- h_{mjr} – Tonnes of raw material **r** used by producing product **j** on machine **m**
- d_{jkt} – Demand for product **j** at market **k** in year **t**
- p_{jk} – Unit price paid for product **j** at market **k**
- c_{mj} – Variable costs associated with producing product **j** on machine **m**
- f_{mjk} – Shipping costs of shipping product **j** from machine **m** to market **k**
- r_{mrq} – Material costs of purchasing material **r** for machine **m** from supplier **s**
- F_m – Yearly Fixed costs associated with operating machine **m**
- F_z – Yearly Fixed costs per machine associated with operating mill **z**
(Distributed amongst all mill machines, i.e. total is divided by number of machines)
- τ_t – Discount factor for year **t** (to find present value of cash flow)
- β_c – Cost of purchasing capital item **c**

Objective Function (maximize):

$$\begin{aligned}
 & \sum_{t=1}^T \sum_{z=1}^Z \sum_{m \in M\{z\}} \tau_t \left[\underbrace{\left(\sum_{j=1}^J \sum_{k=1}^K y_{mjkt} (p_{jk} - f_{mjk}) \right)}_{\text{Revenue from Sales}} - \underbrace{\left(\sum_{j=1}^J x_{mjt} c_{mj} \right)}_{\text{Variable Costs}} - \underbrace{\left(\sum_{r=1}^R \sum_{q=1}^Q r_{mrq} v_{mrqt} \right)}_{\text{Resource Costs}} \right] \\
 & - \underbrace{\left(1 - \sum_{s \leq t} \alpha_{ms} \right) F_m}_{\text{Fixed Machine Costs}} - \underbrace{\left(1 - \sum_{s \leq t} \mu_{zs} \right) F_z}_{\text{Fixed Mill Costs}} - \underbrace{\sum_{c=1}^C \sigma_{mct} \beta_c}_{\text{Capital Item Costs}}
 \end{aligned}$$

As seen above, the objective of SOCRATES is to maximize total discounted profits over the time horizon. In this context, profits consist of the pure revenue received from sales, minus total costs incurred over the time horizon.

Constraints:

$$\sum_{j=1}^J \frac{x_{mjt}}{e_{mjt}} \leq 1, \quad m = 1 \dots M, \quad t = 1 \dots T \quad (\text{Production Constraints}) \quad (1)$$

$$\sum_{j=1}^J h_{mjr} x_{mjt} = w_{mrt}, \quad m = 1 \dots M, \quad r = 1 \dots R, \quad t = 1 \dots T \quad (\text{Resource Usage Constraints}) \quad (2)$$

$$\sum_{q=1}^Q v_{mrqt} = w_{mrt}, \quad m = 1 \dots M, \quad r = 1 \dots R, \quad t = 1 \dots T \quad (\text{Resource Supply Constraints}) \quad (3)$$

$$\sum_{k=1}^K y_{mjkt} = x_{mjt}, \quad m = 1 \dots M, \quad j = 1 \dots J, \quad t = 1 \dots T \quad (\text{Shipping Constraints}) \quad (4)$$

$$\sum_{m=1}^M y_{mjkt} \leq d_{jkt}, \quad j = 1 \dots J, \quad k = 1 \dots K, \quad t = 1 \dots T \quad (\text{Demand Constraints}) \quad (5)$$

$$\sum_{s \in S} \sigma_{mcs} \leq 1, \quad m = 1 \dots M, \quad c = 1 \dots C \quad (\text{One Capital Item Purchase}) \quad (6)$$

$$\sum_{s \in S} \alpha_{ms} = 1, \quad m = 1 \dots M \quad (\text{Machines Must Close}) \quad (7)$$

$$\rho_{mjt} \leq \sum_{s>t} \alpha_{ms}, \quad m = 1 \dots M, \quad t = 1 \dots T \quad (\text{No production on Closed Machines}) \quad (8)$$

$$\rho_{mjt} \leq \sum_{s<t} \sigma_{mcs}, \quad \text{if machine } \mathbf{m} \text{ requires capital item } \mathbf{c} \text{ to produce } \mathbf{j}, T > 1 \quad (9a)$$

$$\rho_{mjt} = 0, \quad \text{if machine } \mathbf{m} \text{ requires capital item } \mathbf{c} \text{ to produce } \mathbf{j}, T = 1 \quad (9b)$$

$$x_{mjt} \leq \rho_{mjt} e_{mjt}, \quad m = 1 \dots M, \quad j = 1 \dots J, \quad t = 1 \dots T \quad (\text{Only produce if binary is set to 1}) \quad (10)$$

$$\sum_{s \in S} \mu_{zs} = 1, \quad z = 1 \dots Z \quad (\text{Mills Must Close}) \quad (11)$$

$$\mu_{zt} \leq \sum_{s \leq t} \alpha_{ms}, \quad n = z \dots Z, \quad m \in M\{z\}, \quad t = 1 \dots T \quad (\text{Machine/Mill Closure}) \quad (12)$$

Constraint (1) ensures the combination of products produced on a given machine is within total capacity. Constraint (2) calculates the amount of raw material used given certain production levels. Constraint (3) sources raw materials needed for production from different suppliers. Constraint (4) ensures the amount of product shipped to a market from a machine is equal to the amount produced. Constraint (5) ensures the amount of product shipped to a market is less than or equal to demand. Constraint (6) ensures that at most one of a certain capital item is bought for a given machine. Constraint (7) ensures that all machines are forced closed by the artificial year. Constraint (8) ensures no production on closed machines. Note that the summation will be zero if the machine is already closed in a previous year, or one if it closes at a later date. Constraints (9a/b) ensure that capital items are purchased before production of restricted products can begin. Note that the ‘if’ statements relate back to the capital array of a machine. Constraint (10) ensures that a machine produces within its capacity limits, whilst including the binary produce variable to link with the capital constraints. Constraint (11) ensures that all mills are forced closed by the artificial year. Constraint (12) ensures that all machines associated with a mill must be closed before the mill can close.

2 A Stochastic Model

As has been previously discussed, the optimization model known as SOCRATES was originally fully deterministic in nature. Whilst a fully deterministic model may be useful in the presence of accurate parameter forecasts, this is almost never the case. Uncertainty, especially in product demand, is almost always present, and hence should be accounted for.

To account for uncertainty in demand in the case of a the SOCRATES model, we shall develop a stochastic programming model, based on the split-variable formulation and associated methods that were proposed by Singh, Philpott & Wood (2009) for application to electricity networks. Uncertainty shall be expressed through the creation of a rooted scenario tree, with numerous scenario paths (indexed by ω) traversing multiple possible states of the world at each stage (represented as nodes). For simplicity it is assumed that demand is the only stochastic parameter. It may be noted that other parameters such as product price, electricity price or exchange rate could also be implemented within a stochastic framework if considered appropriate.

Discrete capacity-related decisions link each stage, having an effect on subsequent periods, whilst what could be considered as individual single-stage capacity planning models are solved at each node. This gives the overall problem a typical ‘block-diagonal’ structure, lending the problem well to solution by Dantzig-Wolfe decomposition methods. In essence, the capital constraints related to machine/mill closures and capital item purchases will be time-dependent (i.e. depend on decisions made in previous or subsequent periods), whilst the remaining constraints apply solely to the current state of the world.

2.1 Forming the Deterministic Equivalent

To extend the previously defined SOCRATES formulation to encompass a stochastic framework, we first must discard the notion of an ordered ‘planning horizon’ of years. We instead introduce the following concepts:

- A set of ‘nodes’ (indexed by \mathbf{n}) which each represent a ‘unique state of the world’ with individual demand realizations.
- A set of scenarios (indexed by ω), each composed of a set of nodes that define the path leading up to the leaf node. We define the set of nodes that comprise a scenario path to be $\mathbf{SP}\{\omega\}$.
- For each node, we define the set of scenarios that contain the given node on their scenario paths. We define this to be $\mathbf{NS}\{\mathbf{n}\}$.
- For each node, we define the probability of the occurrence of that particular state of the world. This parameter is named $\mathbf{NP}\{\mathbf{n}\}$.
- For each scenario path we also add an additional artificial year after the final leaf node, for use in forcing machine/mill closures.

It is important to note that in the stochastic model; production, demand, supply and resource constraints remain essentially the same, except they are now indexed by node instead of year. The following capital constraints are altered for the stochastic model:

$$\begin{aligned} \text{for: } m = 1 \dots M, \quad c \in \mathbf{CAP}\{m\}, \quad j = 1 \dots J, \quad n = 1 \dots N \\ \text{if: } a_{mcj} = 1 \\ \text{then: } \rho_{mjn} \leq \sum_{x \in \mathbf{P}\{n\}} \sigma_{mcx} \end{aligned}$$

For each node \mathbf{n} , let $\mathbf{P}\{\mathbf{n}\}$ denote the set of direct predecessor nodes of \mathbf{n} , and $\mathbf{S}\{\mathbf{n}\}$ denote the set of direct successor nodes of node \mathbf{n} . Let $\mathbf{CAP}\{\mathbf{m}\}$ denote the set of capital items available for machine \mathbf{m} , and \mathbf{a}_{mcj} the capital array (keeping in mind that if $a_{mcj} = 1$, then capital item c is needed by machine m to produce product j). ρ_{mjn} are the binary production variables, and σ_{mcx} are the binary capital item purchase variables. This is the equivalent stochastic constraint to the capital item constraints in SOCRATES. It ensures that at a given node, you may not produce certain items unless you have purchased the appropriate capital items in a previous time period.

$$\begin{aligned} \text{for: } m = 1 \dots M, \quad j = 1 \dots J, \quad n = 1 \dots N \\ \rho_{mjn} \leq \sum_{x \in \mathbf{S}\{n\}} \alpha_{mx} \end{aligned}$$

Here α_{mx} are the binary machine closure variables. This constraint ensures that no production occurs if a machine has already closed. This is achieved by scanning the nodes branching from the node in question (i.e. the successor nodes). If the summation is greater than zero then the machine is still open (i.e. it closes at a later date), and therefore production can be allowed (note that the capital item constraint must also be satisfied to enable production).

$$\begin{aligned} \text{for: } m = 1 \dots M, \quad c \in \mathbf{CAP}\{m\}, \quad \omega = 1 \dots \omega \\ \sum_{n \in \mathbf{SP}\{\omega\}} \sigma_{mcn} \leq 1 \end{aligned}$$

The constraint above is equivalent to the ‘‘One Capital Item’’ constraint in the previous SOCRATES definition. It ensures that a given capital item is purchased at most once for a given machine on any given scenario path.

$$\text{for: } m = 1 \dots M, \omega = 1 \dots \omega$$

$$\sum_{n \in SP[\omega]} \alpha_{mn} = 1$$

This stochastic equivalent constraint ensures that machines are forced to be closed by the end of the simulation (note that $SP\{\omega\}$ includes final artificial nodes).

$$\text{for: } z = 1 \dots Z, m \in M\{z\}, n = 1 \dots N$$

$$\mu_{zn} \leq \sum_{x \in P\{n\} \cup n} \alpha_{mx}$$

This equivalent constraint ensures that mills are unable to close if machines are still considered to be operating. Note $M\{z\}$ is the set of machines associated with a given mill z .

$$\text{for: } z = 1 \dots Z, \omega = 1 \dots \omega$$

$$\sum_{n \in SP[\omega]} \mu_{zn} = 1$$

This constraint forces all mills closed by the end of the simulation (essentially the same constraint as the mill closure constraint).

As well as these constraints, the objective function was also altered to reflect the now probabilistic nature of the model (to maximize *expected* profits).

2.2 Computational Results

A 3-stage, 9-scenario model was formulated as above, with 2 mills, 6 machines, 6 markets, 35 products and up to 30 capital items available for each machine (mirroring the original SOCRATES model in all aspects, except for the time-horizon). As expected, traditional branch and bound solution methods via CPLEX proved very ineffective.

After ~43 hours of computational time a MIPGap of only 5.67% had been reached. The solution procedure was subsequently terminated. Considering the relatively small size of the model (only 9 scenarios), this result further emphasized the need for an efficient solution method that exploited the specific problem structure.

To form a Deterministic Equivalent that would actually reach an optimal solution in reasonable time, further model simplifications would have to take place. Reaching an optimal solution was considered particularly important, as any optimal solutions produced would be used to check that subsequent Dantzig-Wolfe decomposition methods were coded correctly and reaching proper solutions. In subsequent model revisions it was therefore decided to remove all aspects of machine/mill closures. As well as expectantly increasing Deterministic Equivalent tractability, closure decisions would be difficult to implement within the SV1-Dantzig-Wolfe framework described by Singh, Philpott & Wood (2009). In their framework, they allow for a single type of ‘capacity expansion’ decision. Extension of the SV1 framework to encompass multiple types of different capacity-related decisions remains as possible future work in the area.

Whilst the removal of ‘capacity contraction’ decisions from the model would be detrimental to the realism of the model, we believe that the remaining aspects of the model (specifically ‘capacity adaptation’ decisions captured by the inclusion of capital items) will still be of great value and interest. We shall now describe the steps taken in adapting this ‘Deterministic Equivalent’ stochastic formulation to adhere to the SV1 framework of Singh, Philpott & Wood (2009).

2.3 Dantzig-Wolfe Decomposition

As previously discussed, models of significant size often require efficient solution methods to improve tractability. Dantzig-Wolfe decomposition (G. Dantzig & P. Wolfe, 1960) is one such method for efficiently solving large linear programs of a special ‘block-diagonal’ structure (based on delayed column generation).

As mentioned earlier, Singh, Philpott & Wood (2009) proposed that Dantzig-Wolfe decomposition provides an efficient solution method for multi-stage stochastic capacity planning problems with binary or integer variables. They specifically applied the method to solve models related to electricity networks, and we shall apply their ‘SV1’ Dantzig-Wolfe approach to our stochastic SOCRATES formulation. Note that multi-stage stochastic programs are traditionally solved using Bender’s Decomposition. This will not work in general for problems with binary or integer variables in stages apart from the first one; hence we must adopt a different approach.

2.4 A Split-Variable Formulation (SV1)

Before we could apply the Dantzig-Wolfe algorithm to the initial Deterministic Equivalent model mentioned in the previous section, several alterations had to be made.

Firstly, all constraints relating to machine or mill closures had to be removed, as these could not be incorporated within the SV1 framework. Secondly, the binary decision variables relating to capital item purchases were ‘split’ (hence the name) into capital item *requests* and capital item *grants*. Note that the ‘1’ in SV1 refers to the additional condition that only 1 expansion may occur per item. We effectively already incorporate this condition, as an item may only be bought at most once for a given machine on a given scenario path.

We then define sub-models that essentially consist of production planning models (one sub-model for each node). Sub-models are able to make ‘requests’ for capital item purchases, these ‘requests’ can then be satisfied within the master problem in the form of capital item ‘grants’. A sub-model can make a large amount of different combinations of ‘requests’ for different items; each of these different combinations is known as a Feasible Expansion Plan (FEP). It is the task of the master problem to decide what the most profitable combination of these FEPs is, whilst ensuring overall feasibility is maintained. Note that only 1 FEP can be chosen per sub-problem, and that if a specific FEP is chosen by the master problem, the requests made by it must be satisfied at some predecessor node.

The Dantzig-Wolfe procedure iteratively looks for the most profitable combination of FEPs (FEPs represented as columns in the master problem). Note that due to new simplifications and the addition of new ‘split-variables’, a new Deterministic Equivalent had to be constructed. The Deterministic Equivalent forms somewhat of a ‘scientific control’, enabling comparison and verification of the Dantzig-Wolfe methodology.

We define the following SV1 Master Problem formulation:

SV1-MP:

$$z_{SV1-MP}^* = \min \sum_{m=1}^M \sum_{c \in CAP[m]} \sum_{n=1}^N \varphi_n k_{mcn} + \sum_{n=1}^N \sum_{f \in F_n} \varphi_n v_n^f w_n^f$$

s.t:

$$\sum_{f \in F_n} \hat{x}_{mcn}^f w_n^f \leq \sum_{h \in P_n} x'_{mch} \quad , \quad n = 1..N, \quad m = 1..M, \quad c \in CAP[m] \quad [Dual\ Variables: \boldsymbol{\pi}_{mcn}]$$

$$\sum_{h \in P_n \cup n} x'_{mch} \leq 1 \quad , \quad m = 1..M, \quad c \in CAP[m], \quad n = 1..N$$

$$\sum_{f \in F_n} w_n^f = 1 \quad , \quad n = 1..N \quad [Dual\ Variables: \boldsymbol{\mu}_n]$$

$$k_{mcn} = \frac{q_{mc} x'_{mch}}{(1+r)^{Y_n}} \quad , \quad m = 1..M, \quad c \in CAP[m], \quad n = 1..N$$

Here φ_n represents the probability of node \mathbf{n} , k_{mcn} represents the capital item costs at node \mathbf{n} (calculated in a constraint below), v_n^f represents the profit/loss made from operations at node \mathbf{n} under FEP \mathbf{f} , w_n^f are the binary FEP choice variables (1 if FEP \mathbf{f} is chosen for node \mathbf{n} , 0 otherwise), and F_n represents the set of possible Feasible Expansion Plans (FEPs) at node \mathbf{n} . In the constraints \hat{x}_{mcn}^f are binary variables representing capital item requests for FEP \mathbf{f} , x'_{mcn} are binary variables representing capital item grants, P_n represents the set of predecessor nodes to node \mathbf{n} , q_{mc} represents capital item costs, r is the discount rate, and Y_n represents the cardinality of the year in which node \mathbf{n} lies. The sub-problems are:

SV1-SP[n]:

$$z_{SV1-SP[n]}^* = \min \varphi_n q_n^T y_n - \hat{\pi}_{mcn}^T x_{mcn} - \hat{\mu}_n$$

s.t.: $\rho_{mjn} \leq x_{mcn}$, $m = 1..M$, $c \in CAP[m]$, $j = 1..J$, (if \mathbf{c} needed to produce \mathbf{j} and $\mathbf{n}>1$)
 $\rho_{mjn} = 0$, $m = 1..M$, $c \in CAP[m]$, $j = 1..J$, (if \mathbf{c} needed to produce \mathbf{j} and $\mathbf{n}=1$)
 $y_n \in Y_n$

Here $\hat{\pi}_{mcn}^T$ and $\hat{\mu}_n$ are dual variables from the previous master problem solution, x_{mcn} are binary capital item requests, y_n represents a variety of operational decisions including resource allocation, production allocation and product distribution, q_n^T are the operational costs/benefits associated with these operational decisions, ρ_{mjn} represents the binary production variables (1 if product \mathbf{p} is produced on machine \mathbf{m} in node \mathbf{n} , 0 otherwise) and Y_n represents the feasible region for operating decisions at node \mathbf{n} .

Note that capital item requests are incorporated into the sub-problems by requiring them to occur when production is desired on a machine that requires a given capital item it does not have, and that if a machine does not have a required capital item in the initial year, then it is forced to not produce that item (because requests can only be satisfied in predecessor nodes to the current node, and the initial year has no predecessors).

2.5 Dantzig – Wolfe Implementation

Note that the traditional Dantzig-Wolfe procedure applies specifically to linear programs, and we currently have integer/binary variables in both our master and sub problems. In order to overcome this, we solve an LP-relaxation of the master problem (MP-LP) whilst solving sub-problems to optimality as integer programs. As shall be discussed in a later section, we tend to obtain integer solutions regardless of the relaxation at optimality, but in the case of fractional solutions it has been proposed that a branch and price procedure may be applied.

We apply an ‘interior point duals stabilization’ procedure, by solving master problem relaxations with interior point algorithms. The aim of this is to reduce the overall number of columns generated by the procedure, and to hence reach optimality in fewer iterations. ‘MP-LP-I’ denotes the LP relaxation of the master problem, incorporating an interior point duals stabilization procedure.

We are able to define a lower bound on the optimal Master Problem objective by the following formula:

$$z_{lower} = z_{MP-LP-I} + \sum_{n=1}^N s_n$$

(Where $z_{MP-LP-I}$ is the LP-relaxation of the master-problem objective and s_n is the objective of sub-problem n)

The adapted Dantzig-Wolfe solution algorithm for the SV1 formulation follows the following iterative procedure (shown below in pseudo code):

```

Generate Initial trivial FEP sets with high cost for each node;
Repeat {
  For each node {
    Solve integer sub-problem;
    If resultant FEP prices < 0 (as this is a minimization problem) {
      Enter FEP to Master Problem;
    }
  }
  Solve LP relaxation of Master Problem;
  Record dual values for use in subsequent sub-problems;
  Calculate Lower Bound;
  Write Key Statistics to File;
  If no columns were added this major iteration {
    Terminate algorithm: optimality reached;
  }
}

```

Figure 2 - Dantzig-Wolfe for SV1: Pseudo Code

2.6 Computational Results

Using the same data sets as for the previous Deterministic Equivalent (to enable comparison and validation), the SV1 model was solved using the Dantzig-Wolfe algorithm in AMPL using CPLEX. The settings of various solver parameters (including probing, cuts generated, and crossover) were varied to find the optimal combination with the lowest solution times. A moderate level of probing is most beneficial (option `cpex_options "probe 2"`) in solving sub-problems, and the algorithm converges faster with MIP-cuts turned off.

The final stochastic model was tested with multiple scenario-tree structures in order to compare computational times and test the flexibility of the model to encompass larger scenario trees. These were then compared with the computational times for the Deterministic Equivalent formulations. Displayed below is the resulting table, displaying total computational times to reach optimality for a selection of different tree structures.

Figure 3 - Computational Comparison Table

No. of Stages	No. of Scenarios	Solution Time: Deterministic Equivalent (s)	Solution Time: Dantzig-Wolfe (s)
2	3	3	44
3	9	240	177
4	27	1172	630
5	81	10812	1747
2	2	2	36
3	4	173	153
6	32	5691	872
8	128	(Out of Memory)	4361

As can be observed, the Dantzig-Wolfe procedure improves the solution time in all but the most trivially small problems. The computational benefit of the procedure increases as problem size increases, with the largest problems experiencing drastic improvements in solution times.

The largest problem we were able to solve was a 255-node, 128-scenario, 8-stage problem. This corresponded to a Deterministic Equivalent composed of over 200,000 variables and 300,000 constraints.

It is important to note that all solutions proved to be fully integral at optimality (remember that we are solving the LP relaxation of the MP); we have yet to find an instance where fractional solutions occur at the successful termination of the SV1 Dantzig-Wolfe procedure.

3 Discussion

3.1 Computational Issues

A key observation related to the Dantzig-Wolfe solution method concerns the integrality of the optimal solutions. Note that in the master problem we have binary capital item ‘grant’ variables as well as binary FEP choice variables. In our Dantzig-Wolfe iterative procedure, we solve the LP-relaxation of this master problem and hence relax the binary conditions on these variables. It logically follows that therefore it should be possible for fractional solutions to occur at the successful termination of the Dantzig-Wolfe algorithm. Whilst fractional solutions occur in intermediate (sub-optimal) master iterations, solutions were found to tend towards integer solutions at optimality. Despite an extensive search for fractional solutions at optimality, none were found. Further investigation into the reason for this apparent inherent integrality of the master problem is needed.

3.2 Model Relevance

For reasons previously discussed, the final stochastic model was simplified significantly from the original SOCRATES formulation. Whilst many of these simplifications were appropriate (in terms of creating a generic formulation for the process industry), other simplifications were less than desirable (specifically the removal of capacity contraction decisions). Nevertheless, we argue that the remaining decisions contained within the model (capacity adaptation) are still of particular relevance to the current economic climate. In the face of declining global demand in one particular product, it is often the case that through capital expenditure, machines used in the production of one type product can be adapted to produce an entirely different product. This would allow the manufacturer to effectively ‘escape’ from this declining market and focus on other potential products.

4 Conclusions

In this report, we have extended the SOCRATES model to encompass uncertainty in the form of a stochastic programming model. We have shown that through the application of Dantzig-Wolfe decomposition, we are able to greatly increase the tractability of the model.

Key areas of possible future work include the investigation into the apparent inherent integrality of the solutions produced by the SV1 framework, improving the memory efficiency of the AMPL formulations, and the application of the model to industry.

5 References

- Ahmed, S., King, A.J., & Parija, G. (2003). *A multi-stage stochastic integer programming approach for capacity expansion under uncertainty*. J. Global Opt. 26
- Brealey R.A. & Meyers, S.C. (1996). *Principles of Corporate Finance*, McGraw-Hill.
- Dantzig, G., & Wolfe, P. (1960). *Decomposition Principle for Linear Programs*. Operations Research, 8, 101–111.
- Eppen, G.D., Martin, R.K., & Schrage, L. (1989) *A scenario approach to capacity planning*. Operations Research, 37, 517-527.
- Everett, G., Philpott, A.B., & Cook, G. (2000). *Capital Planning Under Uncertainty at Fletcher Challenge Canada*. Proceedings of the 35th Annual Conference of ORSNZ, 231-240.
- Manne, A.S. (1961) *Capacity expansion and probabilistic growth*. Econometrica. 29, 632--649.
- Singh, K., Philpott, A., & Wood, K. (2009). *Dantzig-Wolfe Decomposition for Solving Multistage Stochastic Capacity-Planning Problems*. Operations Research, 57, 1271-1286.

Trim Loss and Inventory Optimisation in Paper Mills

Q. Lim

Department of Engineering Science
University of Auckland
New Zealand
qlim001@aucklanduni.ac.nz

Abstract

Generating efficient production schedules that minimise costs is of great importance in the paper industry. This planning problem combines two well known Operations Research problems: the lot sizing problem and the cutting stock problem. Separately, both have received extensive attention and many variants of each can be solved efficiently using mixed integer programming formulations. In the paper industry, however, these two problems are coupled. The resulting models have proven much harder to solve, with most approaches in the literature so far being of a heuristic nature.

In this project we look at the scheduling problem for one of the paper machines at Norske Skog's Tasman paper mill. The nature of this particular problem allows some simplifications to be made in the modelling process. To generate production schedules we develop an integer programming formulation, which minimises trim and inventory costs. The column generation procedure for the 1-dimensional cutting stock problem is adapted to solve this model. Using historical data we compare the schedules generated by the model with those used at the plant.

Key words: Lot sizing, cutting stock, integer programming, paper production, scheduling.

1 Introduction

In the complex business of paper production, mathematical techniques can be applied in all stages of the supply chain. In this paper we focus on the problem of minimising costs in the production of reels of paper from pulp, the process depicted in figure 1.

The process can be divided into three stages. First, paper is formed from wet pulp. The pulp is fed into a *paper machine*, which lays it onto a mesh to form a sheet. The paper is then dried and smoothed. The output of this stage is a *jumbo reel*. This is an extremely large reel of paper, a typical width being close to 7m.

In the second stage, the jumbo reels are unwound and slit into smaller reels of varying widths. These are then rewound onto smaller *product reels*, with widths and

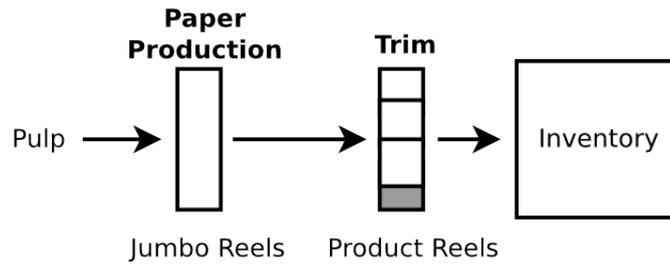


Figure 1: The paper production process

diameters as required by customers. Finally, the product reels are held in inventory before being shipped to the customer.

Various costs are incurred throughout this process which need to be minimised. The three most important costs are changeover costs, trim loss costs and inventory costs. *Changeover costs* are incurred in the first stage. Paper mills may produce many different grades of paper, however a paper machine is only able to produce one grade of paper at a time. Different grades of paper are made from different types of pulp, and while the paper machine is transitioning between two grades, the paper that is produced will meet the standards of neither grade and cannot be sold. Changeover costs are a measure of this lost production.

Trim loss costs are incurred in the second stage, when the jumbo reels are sliced up into smaller product reels using a *cutting pattern*. A cutting pattern may not occupy the entire width of the jumbo reel, and the resulting wastage is known as the *trim loss* (figure 2), which ends up being re-pulped (with an associated cost). A paper mill will seek to minimise such wastage by choosing efficient cutting patterns as this trim loss represents under-utilised production capacity. When modelling these problems we can assign a trim loss cost to each cutting pattern, which is the economic cost incurred each time the pattern is used.

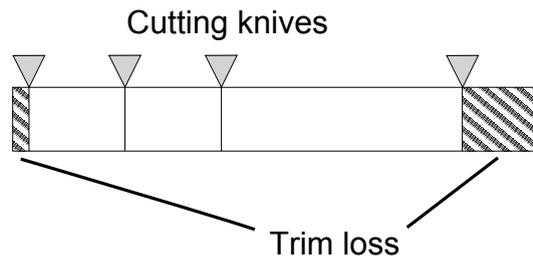


Figure 2: Schematic of jumbo reel and cutting pattern

Finally, *inventory costs* are the costs associated with carrying finished goods in inventory.

1.1 Previous work

Modelling this production process combines two well-known Operations Research problems: the Lot Sizing Problem (LSP) and the Cutting Stock Problem (CSP). Both problems have received considerable attention from OR practitioners as they arise in many industrial applications.

Lot sizing models are used to determine the size of a batch of production before switching to a different product; they explore the tension between changeover and

inventory costs. Many variations on the problem exist with different types of costs and structure (Drexler and Kimms 1997); some of these admit tight reformulations such as those based on the facility location and shortest path problems. This has led Pochet and Wolsey (2006) to claim that “there is a nontrivial fraction of practical lot-sizing problems that can now be solved by nonspecialists just by taking an appropriate a priori reformulation of the problem, and then feeding the resulting formulation into a commercial mixed-integer programming solver”.

The (1 dimensional) cutting stock problem seeks to find the optimal way of cutting stock pieces of varying width so that demand for each width is satisfied and trim loss wastage is minimised. An explicit integer programming formulation for this problem contains a variable for each cutting pattern. For practical problems there may be many millions of valid cutting patterns, making the model hard to solve. This difficulty was overcome in the seminal paper of Gilmore and Gomory (1963) who outline how delayed column generation can be used to handle these variables.

In paper production, the lot sizing and cutting stock problems are *coupled* – that is, the solution of one will affect the other and both need to be solved simultaneously to arrive at the optimal solution. The resulting problem is much harder to solve than the two problems separately; the inclusion of cutting patterns prevents standard path-based lot-sizing reformulations from being used effectively. In industry the problems are often solved separately, or software employing heuristics such as genetic algorithms are used.

Research into coupled problems has, for the most part, only occurred relatively recently. Respício (2003), in her PhD thesis (Portuguese), looked at a multi-item coupled cutting stock and lot-sizing problem as part of the development of an integrated Decision Support System for the paper industry. By ignoring setup costs, a branch and price algorithm was developed to solve this problem. Poltroniere et al. (2008) investigate iterative heuristic methods for solving the coupled problem in the paper industry. Their approaches decouple the problems into separate lot-sizing and cutting stock problems, which are solved using methods including Lagrangian relaxation. The 2-dimensional coupled problem has also received attention in papers such as Nonås and Thorstenson (2000) and Gramani and França (2006).

1.2 Norske Skog Tasman

Norske Skog is the world’s largest producer of newsprint and magazine paper. Located in Kawerau, New Zealand, the Norske Skog Tasman paper mill is one of three mills in Australasia and with two paper machines boasts annual production capacity of over 300,000 tonnes. Production is flexible across all paper machines in the region, which means that strategic decisions must be made concerning the allocation of production to each mill. However, in this report we focus on the lower level planning problem faced by each mill once the demands that they must satisfy are known.

The company currently uses what amounts to a two-step procedure for solving this problem. The lot sizing problem is solved heuristically by a production planner, which determines the orders that will be produced on the paper machine (the grade of jumbo reels to make). A commercial software package is then used to solve the cutting stock problem, which determines how the jumbo reels are to be cut up. The aim of this project is to demonstrate that by considering both problems simultaneously, more production schedules with lower costs can be implemented.

Modelling the entire production problem at the Tasman mill is complex. The

mill has two paper machines, each of which is capable of producing a subset of the paper grades manufactured at the plant. In this paper we restrict ourselves to a simpler problem, namely the scheduling of production on just one of the paper machines. This machine produces predominantly newsprint grades of paper, which have relatively small changeover costs between grades. Furthermore, the demands for the various grades of newsprint are such that the machine spends most of its time producing a single grade. As lot-sizing models which include changeover variables and constraints are much harder to solve than those without, we omit these from the optimisation. However, for comparison we still calculate their value as part of the total cost of the schedule.

1.3 Outline

In the next section we present a mathematical formulation of the Tasman problem which minimises trim and inventory costs, called ‘TIMBER’. The model is solved using a column generation approach which is explained, and some computational results are given.

In section 3 we apply this model to historical data obtained from the mill. The schedules generated are compared with those used in practice and estimates of the possible cost savings attainable are given.

2 Model TIMBER

We now present an integer programming model for finding schedules which optimise trim and inventory costs. The formulation is based on a standard big-bucket lot sizing model (that is, we place no restriction on the number of grades which may be produced in each period). The model has been extended with the addition of variables that represent the use of cutting patterns and the necessary constraints to link these to the production variables.

2.1 Definitions

Sets

- \mathbb{G} = The set of all grades of paper produced
- \mathbb{W}_g = The set of all widths of paper of grade g demanded by customers
- \mathbb{P}_g = The set of all cutting patterns for grade g

Parameters

- T = number of planning periods
- L = the width of a jumbo reel
- d_t^{gw} = demand for grade g and width w during period t
- ρ_w^{gp} = number of times width w occurs in cutting pattern p for grade g
- s_{initial}^{gw} = initial inventory of grade g width w

Costs

- h_t^{gw} = cost of storing one unit of grade g width w in inventory during period t
- r^{gp} = trim loss cost of pattern p for grade g

Decision variables

x_t^{gw} = number of reels of grade g width w produced in period t

s_t^{gw} = number of reels of grade g width w held in inventory in period t

q_t^{gp} = number of times cutting pattern p for grade g is used in period t

2.2 Master Problem

$$\min \sum_{g \in \mathbb{G}} \sum_{t=1}^T \left(\sum_{w \in \mathbb{W}_g} h^{gw} s_t^{gw} + \sum_{p \in \mathbb{P}_g} r^{gp} q_t^{gp} \right) \quad (1)$$

subject to

$$s_{t-1}^{gw} + x_t^{gw} = d_t^{gw} + s_t^{gw} \quad \text{for } g \in \mathbb{G}, w \in \mathbb{W}_g, t = 1, \dots, T \quad (2)$$

$$\sum_{p \in \mathbb{P}_g} q_t^{gp} \rho_w^{gp} = x_t^{gw} \quad \text{for } g \in \mathbb{G}, w \in \mathbb{W}_g, t = 1, \dots, T \quad (3)$$

$$\sum_{g \in \mathbb{G}} \sum_{p \in \mathbb{P}_g} q_t^{gp} \leq C \quad \text{for } t = 1, \dots, T \quad (4)$$

$$s_0^{gw} = s_{initial}^{gw} \quad \text{for } g \in \mathbb{G}, w \in \mathbb{W}_g \quad (5)$$

$$s_t^{gw}, x_t^{gw} \geq 0, \text{ integer} \quad \text{for } g \in \mathbb{G}, w \in \mathbb{W}_g, t = 1, \dots, T$$

$$q_t^{gp} \geq 0, \text{ integer} \quad \text{for } g \in \mathbb{G}, p \in \mathbb{P}_g, t = 1, \dots, T$$

Equation (2) is a flow balance constraint for each product (grade and width). At each period the inventory carried over from the previous period plus the amount of production in the current period must be equal to the sum of the current period's demand and inventory level. Equation (3) links the number of products produced in each period to the cutting patterns that are used. Equation (4) enforces a capacity on the number of jumbo reels that can be produced in a period. Equation (5) specifies the initial inventory for each product.

2.3 Restricted Master and Subproblems

The formulation above includes variables for all feasible cutting patterns in the sets \mathbb{P}_g . In practice, this may amount to many millions of variables. We can adapt the column generation procedure presented by Gilmore and Gomory (1963) to this formulation. The *restricted master problem* (RMP) is formed by replacing \mathbb{P}_g with $\mathbb{P}'_g \subset \mathbb{P}_g$. We initially populate \mathbb{P}'_g for each grade g with the following trivial cutting patterns: For a given grade g and width w^* ,

$$\rho_w = \begin{cases} \lfloor \frac{L}{w^*} \rfloor & w = w^* \\ 0 & \forall w \in \mathbb{W}_g, w \neq w^* \end{cases} \quad (6)$$

is a valid cutting pattern (cutting as many rolls of width w^* as possible from our jumbo reel of width L).

To generate new columns we wish to find new q_t^{gp} pattern variables with negative reduced cost. The reduced cost expression for one of these variables is

$$\text{rc} = r^{gp} - \sum_{w \in \mathbb{W}_g} \pi_w \rho_w^p - \gamma \quad (7)$$

where π_w are the dual variables associated with the production constraints (3) and γ is the dual variable on the capacity constraint (4). Note that this means there is a subproblem for every grade g and time period t . It can be shown that the pattern generation subproblem for grade g and time period t is equivalent to the integer knapsack problem

$$\max \sum_{w \in \mathbb{W}_g} (w\phi_g + \pi_w)\rho_w \quad (8)$$

$$\sum_{w \in \mathbb{W}_g} w\rho_w \leq L \quad (9)$$

$$\rho_w \geq 0, \text{ integer} \quad \text{for } w \in \mathbb{W}_g \quad (10)$$

If the value of the objective function is greater than $L\phi_g - \gamma$, then the reduced cost for this new pattern is negative. A new variable q_t^{gp} can be added to the RMP which represents the number of times the pattern defined by ρ is used in period t .

The integer knapsack problem can be solved using a dynamic programming recursion given by Gilmore and Gomory (1963).

2.4 Solution process

As TIMBER is an integer programme a branch and price approach is required to find an optimal solution. Several branch and price approaches for the cutting stock problem, which maintain subproblem structure, are reviewed in Vanderbeck (2000) and could be extended to this model.

Algorithm 1 TIMBER solution algorithm

Require: Initial feasible solution (use trivial patterns)

repeat

for all time periods $t \in 1, \dots, T$ **do**

for all grades $g \in \mathbb{G}$ **do**

repeat

 Solve the LP relaxation of the RMP

 Solve the pattern generation problem for grade g period t using the optimal RMP duals

if reduced cost ≤ 0 **then**

 Add a variable for the cutting pattern in period t to the RMP

end if

until No new patterns generated by subproblem

end for

end for

until No new patterns generated by subproblems

For this study we elected not to take this approach, sacrificing optimality for simplicity of implementation. Instead, we use column generation to solve the linear

Table 1: Computational results for LSCSP-BB

Problem	Size		Patterns			Gap %	Time (s)		
	T	G	Vars	Gen	Used		Patt	IP	Total
MAR2-30	30	8	1562	247	36	0.53	8	6	14
MAR2-60	60	8	8535	1341	104	0.52	195	79	274
MAR2-90	90	8	12206	1687	104	0.62	371	122	493
MAR2-120	120	8	22654	3542	134	0.88	992	637	1629
APR2-30	30	6	1960	341	23	0.30	7	4	11
APR2-60	60	6	8057	1388	50	0.27	143	27	170
APR2-90	90	6	11592	1437	84	0.29	268	46	314
APR2-120	120	6	18740	2074	119	0.55	651	81	732
APR2-180	180	6	32764	3806	133	0.65	1895	230	2125

‘Vars’ is the total number of pattern variables added to the model, ‘Gen’ is the number of distinct cutting patterns generated and ‘Use’ is the number of distinct cutting patterns used in the optimal integer solution. ‘Gap’ denotes the relative difference between the LP optimal and the integer solution found. The times given are the time spent in the pattern generation part of the algorithm, and that spent solving the final integer programme, as well as the total solution time.

relaxation of the master problem. We then enforce the integrality constraints on the master problem containing the columns generated for the relaxation and solve this IP using traditional branch and bound. This gives a feasible, but not necessarily optimal, integer solution.

An open question is the choice of subproblem to be solved on each iteration of the column generation algorithm, that is, the pricing strategy to use. One strategy is to solve all $|g| \times |T|$ subproblems (i.e. full pricing), however this takes a significant amount of time and results in many variables being added which may be of limited use. We found that a simple but effective method is to select one subproblem to solve on each iteration, which we choose to be the same subproblem as that solved on the previous iteration if it yielded an entering column (otherwise we just choose the next subproblem obtained by iterating through the grades and periods). This results in Algorithm 1. The iterations terminate once all subproblems no longer produce entering columns.

2.5 Computational results

TIMBER was implemented in C# utilising the CPLEX .NET API to solve the LP and IP problems. Computational results for the model are presented in Table 1. The problems were solved on a desktop machine with a Pentium 4 3.20GHz CPU and 2.00GB of RAM. In all instances the integer solution found is within 1% of the bound provided by the LP relaxation. We find that we are able to solve problems of a practical size in reasonable time - for instance a 6 grade, 60 period problem (2 month planning horizon using daily discretisation) can be solved in less than 3 minutes.

Table 2: Comparison of historical and TIMBER-optimised costs

Prob	T	Hist/Opt	Costs (\$000)				% Saving
			Inventory	Trim	Chg	Total	
A	39	Hist	491.4	220.8	15	727.2	
		Opt	416.2	122.7	45	583.9	19.7
B	23	Hist	208.6	97.1	6	311.7	
		Opt	165.8	55.8	12	233.6	25.1
C	31	Hist	407.0	119.2	15	541.3	
		Opt	353.9	42.1	36	396.0	26.8
D	50	Hist	605.1	236.7	0	841.8	
		Opt	491.6	174.4	0	660.0	21.6
E	35	Hist	606.0	131.0	0	736.9	
		Opt	507.9	87.6	0	595.5	19.2
F	39	Hist	454.5	177.5	0	632.0	
		Opt	439.6	48.7	0	488.3	22.7

3 Application to Norske Skog Tasman

In this section we use TIMBER to benchmark historical production at Norske Skog’s Tasman paper mill, analysing the operation of one of their paper machines.

3.1 Methodology

Estimates of the inventory, trim and changeover costs relevant to this problem were provided by Norske Skog staff. Detailed data sets of customer orders and the production on the paper machine over an entire year were also provided.

A number of planning windows between 1 and 2 months in length were selected for analysis. A program was written in C# to estimate the costs incurred at the mill over these planning windows using the data supplied by the mill. An equivalent planning problem was then solved using TIMBER to give a production schedule optimising trim and inventory costs. The changeover costs for these schedules (not considered in the optimisation process) were then added to give a total cost for these schedules, which were then compared to the historical cost found.

3.2 Results

The results for six problem instances are listed in Table 2. For each problem the estimated historical costs are compared with the costs of the optimised schedules produced by TIMBER. In problems A, B and C, the machine is required to produce multiple grades of paper and thus changeovers occur. In problems D, E and F, the demand during the planning window is for only one grade of paper and no changeovers are required.

We find that savings of approximately 20% are achieved. The resulting production schedules delay production of items until very close to their due date to minimise inventory costs. This has implications for the robustness of the schedules, which are discussed in the next section.

Table 3: Comparison of cost savings of robust schedules

Prob	T	% Total cost saving		
		0 periods	1 period	2 periods
A	39	19.7	12.1	4.5
B	23	25.1	18.7	7.7
C	31	26.8	19.5	12.2
D	50	21.6	13.1	4.6
E	35	19.2	10.7	2.2
F	39	22.7	14.5	6.2

Interestingly, despite this, trim costs are still able to be reduced. Other things being equal one would expect trim costs to increase if production is delayed to reduce inventory costs. This suggests that trim savings may be achieved simply by grouping together product widths more effectively.

3.3 Robust schedules

As noted above, the major cost-savings achieved in the optimised schedules arise from the delaying of the production of products until just before their due date. Consequently, the optimised schedules are less robust than those used in practice, as a small delay in production may cause items to be late.

In order to estimate the magnitude of this effect we modified the model to add a degree of robustness to the schedules obtained. This was achieved by adjusting the due dates to be n periods earlier than actually required. As a result, products are produced earlier and the inventory component of the total cost is increased.

Table 3 compares the percentage cost savings of robust schedules with a 1-period and 2-period ‘safety factor’ are compared with the savings of the original ‘non-robust’ schedules. With a 1-period safety factor all cost savings are still greater than 10%. When this is increased to two periods the media cost saving is close to 5%.

3.4 Discussion

A number of limitations apply to our analysis which mean that the cost savings obtained are likely to be overstated. Additionally, the cost parameters of the problem (changeovers, trim and inventory) are extremely hard to quantify precisely.

For the problems with changeovers (A, B and C) our optimised schedules switch between grades more frequently than the historical schedules. The effect of changeover times (which reduce the effective capacity of the machine) was not considered in the model so results for these problems in particular will be optimistic.

Inventory handling costs and other logistics-related costs are not considered. A customer’s order may consist of multiple products of various widths. If these are not all produced at similar times then the tracking and handling of inventory becomes more complicated.

Finally, the rate of production has been assumed to be the same for all grades. Further analysis of the data is required to determine if this is a significant issue.

4 Conclusions

An integer programming model for the coupled lot sizing and cutting stock problem was developed ('TIMBER'). The column generation procedure for the 1-dimensional cutting stock problem is adapted to solve this model, which is able to solve problems of a practical size to near-optimality using commodity hardware.

TIMBER was used to benchmark part of the operations at Norske Skog's Tasman paper mill. Using historical data we generate optimised production schedules with savings of up to 20%. However, given the limits of the model this figure should only be taken as an indicative upper bound. When additional constraints to increase the robustness of the schedules are imposed, savings in excess of 10% are still achieved. These results suggest that further investigation into the feasibility of implementing such an approach are warranted.

Acknowledgments

The author would like to thank his supervisor, Professor Andy Philpott, for his expertise, guidance and encouragement over the course of this project. Thanks also to Graeme Everett of Norske Skog for sharing his wealth of knowledge concerning the paper industry and for providing the data used in this project. In this regard the assistance of Warren Anderson and Bob Wilson is also greatly appreciated.

References

- Drexl, A., and A. Kimms. 1997. "Lot sizing and scheduling: survey and extensions." *European Journal of Operational Research* 99 (2): 221–235.
- Gilmore, P.C., and R.E. Gomory. 1963. "A linear programming approach to the cutting stock problem-part II." *Operations Research*, pp. 863–888.
- Gramani, M.C.N., and P.M. França. 2006. "The combined cutting stock and lot-sizing problem in industrial processes." *European Journal of Operational Research* 174 (1): 509–521.
- Nonås, S.L., and A. Thorstenson. 2000. "A combined cutting-stock and lot-sizing problem." *European Journal of Operational Research* 120 (2): 327–342.
- Pochet, Y., and L.A. Wolsey. 2006. *Production planning by mixed integer programming*. Springer Verlag.
- Poltroniere, S.C., K.C. Poldi, F.M.B. Toledo, and M.N. Arenales. 2008. "A coupling cutting stock-lot sizing problem in the paper industry." *Annals of Operations Research* 157 (1): 91–104.
- Respício, A.L.C.C. 2003. "Apoio à Tomada de Decisão no Planeamento e Escalonamento da Produção ." Ph.D. diss., Faculdade de Ciências da Universidade de Lisboa.
- Vanderbeck, F. 2000. "On Dantzig-Wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm." *Operations Research*, pp. 111–128.

Optimisation Models and Methods for the Container Positioning Problem in Port Terminals

A. Phillips
Department of Engineering Science
University of Auckland
New Zealand
aphi038@aucklanduni.ac.nz

Abstract

In today's globalised economy, almost all international freight is transported via containerised shipping. Workers must rapidly unload these ships, and then reload them with other containers stored in the yard. The decision of where to store containers in the yard under strict space constraints is called the Container Positioning Problem (CPP), which has traditionally been solved using heuristic methods. This project investigates recent literature which proposes optimisation models for the CPP, but concludes they are too difficult to solve due to the complex constraint enforcing the Last-In First-Out (LIFO) condition on container stacks.

The project objective is to demonstrate that a large proportion of LIFO constraints in the model are inactive and that faster solve times can be achieved by exploiting this property. A reduced instance of the CPP without LIFO constraints is formulated. This is then solved iteratively, adding violated constraints. Eventually, the reduced problem yields a solution which is optimal for the full problem. This process is enhanced with a predictive algorithm which identifies a set of critical constraints, and adds them a priori.

This method proves highly effective, and optimal solutions were found for much larger problems than those in the literature. However it is concluded that in order to solve large industry problems, a rethink of the formulation will be required.

1 Introduction

Almost all industries rely on a timely and frequent supply of goods originating in a foreign country. These goods can be either bulk commodities which are shipped on specialised carriers (e.g. oil), or containerised goods which can include consumables, electronics, machinery, medical supplies, etc (Murty, Liu, Wan, Linn, 2005).

Containerised shipping has been described by the World Trade Organisation (2008) as "chief among [the] driving forces of globalisation", with global container traffic volumes predicted to increase by more than double in the 2005 to 2015 period (United Nations, 2007).

In addition to this, the container shipping industry is a highly competitive market (Coronado, Acosta, del Mar Cerban, del Pilar Lopez, 2008) whereby there is intense pressure on the ports to rapidly and efficiently unload and reload container ships. This calls for a streamlined performance of all operations which take place in the Container Terminal Yard.

The main operations which take place in a Container Terminal are pictured in Figure 1 (Sibbesen, 2008) and can be divided into:

- Quayside Operations - the loading and unloading of container ships using quay cranes
- Yard Operations - the placement and extraction of containers in the storage block by the yard cranes
- Landside Operations - the loading and unloading of land vehicles (trains and/or trucks) with containers

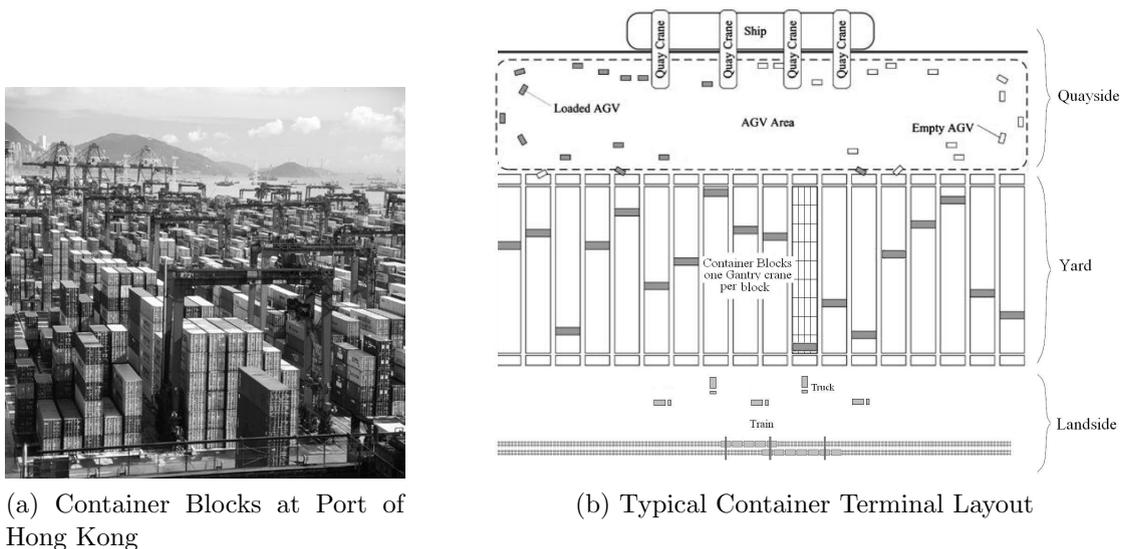


Figure 1: Container Terminal Layout

Yard operations are fundamentally important to the container terminal. This is because all containers which are waiting to be loaded onto a ship, or collected by a land vehicle, must be stored in a designated container block in the yard and then retrieved promptly when its transport arrives. In most container terminals, each container block has one or more gantry cranes which perform all yard movements for that block.

Generating an efficient Yard operations plan requires the Container Positioning Problem (CPP). The CPP aims to find a good sequence of positionings for containers within a block, or equivalently, it attempts to find a good sequence of crane movements to efficiently move containers in a storage block. Since it is commonplace for containers to be relocated whilst in the container block, the problem is more complicated than finding a single assignment of containers to positions. The CPP is further complicated by the Last-in First-out (LIFO) constraint, whereby only the top container from any given stack of containers can be accessed by the crane. Since containers are stacked to the height limited by the overhead gantry crane (which can be up to seven high), it is inevitable that the container next to depart cannot always

be on top of its stack. This will mean that in order to extract containers on lower layers, all containers stacked on top will need to be relocated to other positions.

In 2008 Sibbesen identified in her PhD thesis that the key deficiency in all the previous research was that the impact of the LIFO constraint had not been modelled. The thesis proposes optimisation models and a heuristic for solving the CPP with the LIFO constraint enforced on container stacks. However, the optimisation models took a long time to solve even for simple CPP instances which was attributed to the inclusion of the LIFO constraint. As a result Sibbesen concluded that the use of optimisation methods for solving the CPP may be an intractable research direction.

The research outlined in this report investigates and improves the use of optimisation techniques on the Container Positioning Problem. My research is based on the ‘CPPT’ model proposed in Sibbesen’s thesis, with the initial aim of demonstrating that it can be solved much more quickly by exploiting the problem structure. This is based on the observation that many of the LIFO constraints in any given model are likely to be inactive and nonbinding on the solution. Using this principle I trial more efficient optimisation methods for solving the CPP. A secondary objective of this research is to analyse the existing CPP model proposed by Sibbesen, with the intention of verifying to what extent the model represents the real process.

2 Solution Techniques for the CPPT

2.1 The CPPT Model

The CPPT model variables, objective and constraint classes are defined in Table 1. The CPPT model uses binary variables to represent states, rather than events. The location of containers in a single container block is modelled across a finite discretised time domain.

Decision Variables	
$x_{ctp} \in B$	1 if container c is at position p at time t , 0 otherwise
$xm_{cp} \in B$	1 if container c is ever placed in position p , 0 otherwise
$y_{ct} \in B$	1 if container c is being moved at time t , 0 otherwise
Objective Function	
$\text{minimise : } \sum_c \sum_p xm_{cp} + \sum_c \sum_t y_{ct}$	
the sum of all ‘xm’ variables (positionings) + the sum of all ‘y’ variables (moving time periods)	
Constraints	
Constraint Class	Description
Arrival/Departure	Containers must arrive and depart at scheduled times
Flow	Containers must have a logical continuous motion through the block
LIFO	If a container is removed from a stack, there must be no containers still in the stack which were added more recently than the removed container
Capacity	Crane capacity and Maximum Stack Height

Table 1: Full CPPT Formulation

2.2 Critical LIFO Constraint Identification (CLCI)

As mentioned previously, the key focus of this project is to deal more efficiently with the main LIFO constraint than solving the full MIP directly. In the CPPT, the LIFO constraint class makes up the vast majority of the constraints. “With a planning horizon of one week, discretised into 336 time slots ... [there are] 1.5 billion constraints, 96% concerning the LIFO principle” (Sibbesen, 2008). As a result, the following “Critical LIFO Constraint Identification” (CLCI) Algorithm (Algorithm 1) could potentially be faster than solving the original problem directly. The effectiveness of this algorithm relies on the omission of LIFO constraints dramatically reducing the solve time, and there being only a small number of active constraints.

Algorithm 1 CLCI Algorithm

```
Formulate Reduced problem by omitting all LIFO constraints
while Reduced problem solution contains violated constraints do
    Add any violated constraints from previous iteration
    Solve Reduced problem
    Check solution for violated constraints
end while
An optimal solution to the full problem is found
```

The CLCI Algorithm is guaranteed to eventually find the optimal solution to the original problem, without having to deal with all the constraints. Each iteration before the final one will provide a list of constraints which prevented that solution from being feasible. Since optimisation methods find a solution which is optimal for the reduced problem, performing a series of iterations identifies the constraints which are restricting on the full problem.

The general pattern of the CLCI tends to be short iterations at the start, which progressively (as more constraints are added) increase in duration as the objective function also increases (worsens). It is also observed however, that many iterations do not change the objective function. This is because there may be multiple ways to perform similar actions, all of which must be constrained in order for the reduced problem’s optimal objective value to change.

2.3 Improving the CLCI Algorithm

In order for the CLCI Algorithm to be useful, the time to perform all iterations needs to be significantly less than the total time required to solve the problem directly. This section contains methods which were used to make the CLCI process run faster.

To reduce the amount of time spent checking for violated constraints (which must occur at each iteration), a Just-In-Time compiler for Python called Psyco was used. This enables algorithm-style Python code to execute at near the speed of low level languages. It was also found that searching the solution for violations and identifying the parameters of the violated constraints is much more efficient than iterating through all omitted constraints performing checks on the solution.

For any potential path which a container can take as it moves through the storage block, there will exist many symmetrical paths which are essentially equivalent.

In Figure 2, the upper and lower routes shown are symmetrical to one another. Extending this, any feasible solution can have a symmetrically paired feasible solution which has the same objective function. In the CLCI, when the LIFO constraint is enforced on a specific set of binary variables it must also be applied to all symmetrical binary variable sets. This method can save a large number of iterations on a symmetry-prone problem.

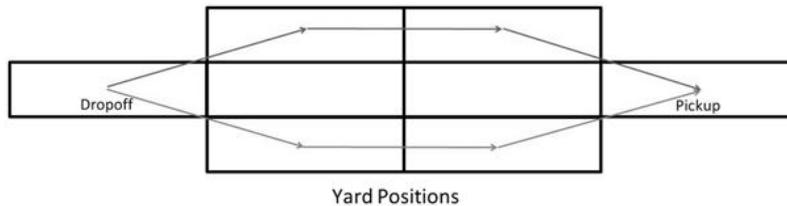


Figure 2: Equivalent Routes through a Container Block

In the CPPT model, when a solution breaks the LIFO rule it causes a range of related constraint violations to occur. These constraints are dependent upon each other, whereby the addition of one can cause others to become unlikely or impossible to be violated. Identifying the dependent groupings of constraints after a CLCI iteration enables many fewer constraints to be added to the model.

The final method involves making a prediction about which LIFO constraints are likely to be violated during the CLCI, based solely on problem data. These deductions are made by estimating when it will be most inconvenient to follow the LIFO rule. Given that the purpose of solving each iteration of the CLCI is to obtain the current set of binding constraints, information about these active constraints is extremely valuable. If all active LIFO constraints can be predicted and included a priori, the reduced problem can be solved in just one iteration. A range of prediction algorithms were written with differing criteria for including a constraint in the ‘likely active’ set. Even the most inclusive or ‘liberal’ predictive algorithm written (which is used in the computational results section), still identifies less than 5% of the total number of LIFO constraints.

3 Computational Results

All computational results were obtained using the ILOG CPLEX 11.0 Callable Library on a 3.2Ghz core of an Intel P4 Dual Core processor, with 2GB of RAM. Four CPP instances from Sibbesen’s thesis were used to compare the model size, and the effectiveness of the different solve techniques. The problems used are titled ‘4S.t’, ‘5S.t’, ‘6S.t’ & ‘7S.t’ whereby the number signifies how many containers (i.e. size) are in the problem, and the ‘S.t’ is a standard CPP instance.

The three methods for which times are shown in Table 2 are: the full model solve (as was performed in the thesis), the CLCI without predictions, and the CLCI using inclusive predictions. For the CLCI method which used inclusive predictions, the set of LIFO constraints added a priori caused only a single solve to be required in all cases.

Solve Times	Problems			
	4S.t	5S.t	6S.t	7S.t
Full Model Solve	1.34	43.6	508.4	39113.6
CLCI No Predictions	2.5	25.7	135.9	16514.2
CLCI Predictions	0.7	12.4	83.2	3184.0

Added LIFO Constraints				
Full Model Solve (all)	824	4080	7680	22680
CLCI No Predictions	56	57	195	510
CLCI Predictions	49	200	335	980

Model Specifications				
Total Variables	268	525	702	1155
Total Constraints	1573	5940	10184	27727
Non-LIFO Constraints	749	1860	2504	5047
Block Dimms (rows*cols)	2*1	3*1	3*1	2*2
Time Domain	18	22	28	34

Table 2: Computational Results

The solve-time results obtained which compare the full CPPT model to using the CLCI (Figure 3), conclusively show that solving the full CPPT model directly is an inefficient method. Furthermore, it is shown that predicting the critical LIFO constraints a priori is the most efficient approach.

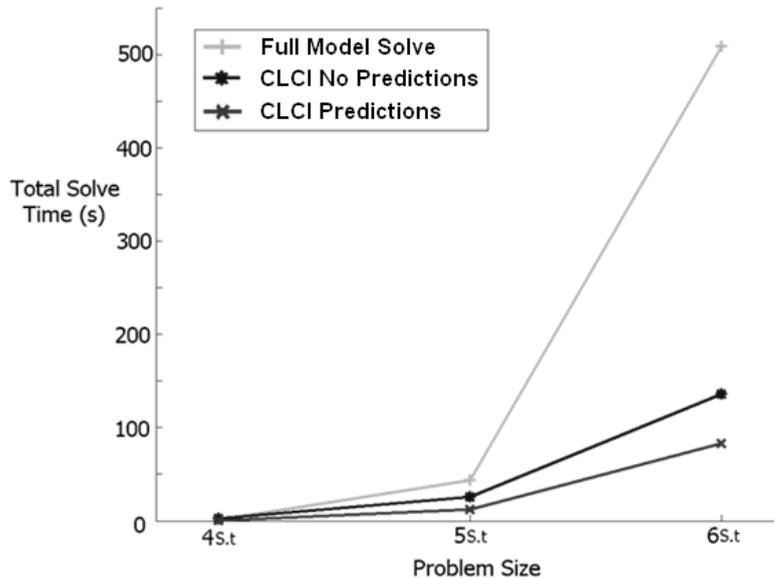


Figure 3: Comparison of methods on different problems

Whilst the CLCI without predictions adds the least LIFO constraints to find an optimal solution, it takes significantly longer than the CLCI with Predictions to reach this solution. When predictions are not used, many iterations must be performed to identify a complete binding set of LIFO constraints. By contrast, the predictive

method can find the optimal solution in a single iteration. Assuming the predictions algorithm can identify a sufficiently small set of LIFO constraints, it is much more efficient to solve this rather than performing many iterations. It is noted from the ‘Added LIFO Constraints’ table, that both the CLCI methods only need to include a very small proportions of the total number of LIFO constraints.

Finally, it is also noted that increasing the problem size dramatically increases the solve time required for all methods. Whilst the methods maintain their efficiency ranking, each method experiences an exponential rise in solver time.

4 Limitations of the CPPT

The secondary aim of this research was to analyse the CPPT model presented by Sibbesen. Having demonstrated the existence of more efficient techniques to solve the CPPT, it is important to have verification that the real-world process is being realistically and efficiently modelled.

4.1 Problem Data

The time required to move between positions in the model container block forms the representation of the physical layout of the container block positions. Reasonable accuracy in these times is required for a feasible and optimal schedule of crane movements, however these times in the CPPT have been simplified significantly to make the problem more tractable. The primary issues identified are:

- Movements between ‘rows’ and ‘columns’ take an equivalent amount of time. In reality, containers have a much greater length than width.
- The time spent attaching Gantry crane connections to the containers is not modelled. This skews the output solution to favour many small movements.

Resolutions for these issues could be incorporated in the model, however it would require a greater level of time discretisation. As acknowledged by Sibbesen, the size of the CPPT mathematical model is extremely sensitive to changes in the size of the time domain.

4.2 Objective Function

The objective function in the CPPT is defined as the number of container placements, plus the number of time periods spent moving containers. Several issues arise with this:

- The objective function consists of the summation of a unitless placement count, and a time measure. In order to make this summation, a weighting factor would need to be applied to represent the cost of a placement in minutes.
- Container repositionings are penalised twice in the objective. Repositioning inefficiency should be discouraged by the contribution to a higher sum of movement times, instead of explicitly applying a penalty. The application of this penalty is burden on the problem, contributing to 3% of variables, and 10% of all constraints in the reduced problem.

- The crane’s current position is not represented. As a result, the distance between the end point of one move and the start point of the next is unquantified and cannot feature in the objective. In a realistic container block (which can be 10 rows by 37 columns), it would clearly be much more efficient to perform a series of moves in one location.

4.3 Variables

All variables in the CPPT model are binary, and are structured around the assumption that only one Gantry crane is present. Many ports operate with two Gantry cranes per container block which have synchronised operations in an efficient schedule. The underlying CPPT model structure does not allow for the extension into incorporating multiple cranes.

4.4 Constraints

Although the CPPT model contains many more constraints than variables, another important physical constraint has been overlooked in the model. This is caused by the fact that the CPPT model represents occurrences or incidences, rather than events. As a result, it was possible to place a container at a position and pick up another container at that position simultaneously (shown in Figure 4). This constraint omission was exploited in many of the ‘optimal’ solutions presented by Sibbesen, making the schedules infeasible. In order to prevent this a constraint was formulated, which represents a significant proportion of total full model constraints (10%).



Figure 4: An Unconstrained Violation

5 Conclusions

The original aim of this research had been to reimplement the CPPT optimisation model and demonstrate that it could be solved more efficiently than via a full direct solve. This objective is considered to be achieved, and it is proposed that a constraint omission and prediction approach may be useful in other applications where a constraint is largely inactive and optimisation is being used.

The secondary goal of this project was to investigate the CPPT formulation itself to check its validity as a representation of the actual problem. A number of deviations from reality and the omission of an important constraint were identified. These changes, along with the many other model deficiencies unmentioned by Sibbesen, show that the model was not intended for rigorous comparison against the heuristic method presented.

Given the two preceding conclusions, it can be said that optimisation techniques for the Container Positioning Problem have not been thoroughly tested. Sibbesen's work identified the need for the LIFO constraint to be incorporated in CPP models, however no considerable attempt to create an optimisation model with this feature has been made. Therefore I propose future work of a new CPP model which takes advantage of the problem structure more than the CPPT, and potentially a rolling time-horizon approach.

6 Acknowledgements

I would like to sincerely thank my supervisors Professor David Ryan and Associate Professor Matthias Ehrgott for their assistance and guidance throughout this research.

7 References

Coronado, D., Acosta, M., del Mar Cerban, M., del Pilar Lopez, M. (2008). Container Traffic from an Economic Perspective: Analysis, Trends and Prospects. In, Economic Impact Of The Container Traffic At The Port Of Algeciras Bay (pp. 5-26). New York: Springer-Verlag New York

Murty, K. G., Liu, J., Wan, Y., Linn, R. (2005). A decision support system for operations in a container terminal. *Decision Support Systems*, 39(3), 309-332.

Sibbesen, L. K. (2008). Mathematical models and heuristic solutions for container positioning problems in port terminals. Copenhagen, Denmark.

The World Trade Organisation. (2008). World Trade Report 2008: Trade in a Globalizing World. Website: http://www.wto.org/english/res_e/reser_e/wtr08_e.htm

United Nations. (2007). Regional Shipping and Port Development: Container Traffic Forecast 2007 Update. Website: <http://www.unescap.org/ttdw/PubsDetail.asp?IDNO=196>

An Integrated Collaboration Platform for Sustainable Development: Project Proposal and Initial Exploration

Max Erik Rohde
Department of Information Systems and Operations Management
University of Auckland
New Zealand
maxrohde@mac.com

A

Sustainability is well-established as an essential driving goal to guide organisations in a world of fierce competition. A key direction in business innovation today is sustainable development and sustainable business management that aspires to deliver integrated and balanced performances in the three sustainability dimensions: social, economic and environmental.

Small and Medium Size Enterprises (SMEs) in New Zealand face unique challenges to adapt to the best business practices on the global markets as they have limited access to the opportunities created by new technology and improved practices. Nonetheless, New Zealand SMEs are highly innovative and can learn from each other rather than relying on external knowledge sources. According to Ministry of Economic Development (2005), this is an often unrealised opportunity.

In our research, we want to explore, conceptualise, design, implement and evaluate a collaboration platform that can support New Zealand SMEs in their journey towards sustainable business models. Our collaboration platform aims at providing systematic mechanisms to address technological, pragmatic, strategic, legal, and cultural barriers to collaboration. The initial focus of our exploration aims at addressing technological barriers by proposing a system to represent, author and navigate complex multi-paradigm, multi-domain, and multi-dimensional data in network representations.

Knowledge Management, Collaboration, Sustainability, SMEs

In the light of globalisation and ever increasing competitive pressure, New Zealand organisations must be able to learn and innovate their processes quickly (Chetty & Campbell-Hunt, 2004). The driving force of process innovation in modern organisation is knowledge (Srivardhana & Pawlowski, 2007): without knowledge of what improvements are possible and how they can be implemented, complex practices cannot be improved.

A key direction in business innovation today is sustainable development and sustainable business management that aspires to deliver integrated and balanced performances in the three sustainability dimensions: social, economic and environmental. What has become known as sustainable business transformation is employed by many leading

multinational organisations (BP, 2002). Small and Medium Size Enterprises (SMEs) in New Zealand face unique challenges to adapt to the best business practices on the global markets. New Zealand organisations often have difficulties to seize the opportunities created by new technology or improved practices (World Economic Forum, 2006; Ministry of Economic Development, 2005). Notwithstanding, New Zealand SMEs are highly innovative and have formed specialised clusters in many areas that can compete on the global market and grow significantly. According to the report *SMEs in New Zealand: Structure and Dynamics 2009*, companies with 20-49 employees account for 60% of the High Growth organizations in New Zealand measured in terms of creating new jobs (Ministry of Economic Development, 2009).

In their journey towards sustainable business models, we see enabling collaboration as a key factor. New Zealand SMEs have limited access to global knowledge sources but can, nonetheless, learn from each other; an opportunity they often do not realise (Ministry of Economic Development, 2005). This research aims at exploring, designing, implementing, and evaluating a collaboration platform that New Zealand SMEs could leverage to share knowledge and innovate.

S C

Sustainability is well-established as an essential driving goal to guide organisations in a world of fierce competition (Hart, 1997). The establishment of a sustainable business model confronts organisations with multi-dimensional challenges such as a transparent governance process and transparent communication (Ahmed & Sundaram, 2009; GRI, 2002). Roadmaps can provide useful guidance for organisations in the process of transformation into a sustainable business (see Figure 1).

Although roadmaps as shown in Figure 1 provide useful guidance for organisations in achieving a sustainable business, the enactment of such roadmaps is a very difficult and complex task. The key ingredient here, as in any innovation driven transformation process, is knowledge. In the New Zealand context, we see the key challenge for organisations in acquiring the necessary knowledge they need to make their business sustainable. As already argued, New Zealand SMEs are confronted with unique challenges in acquiring such expert knowledge. One opportunity could be the strengthening of collaboration and thereby allowing dissemination of essential knowledge between businesses, universities, government, and independent agencies.

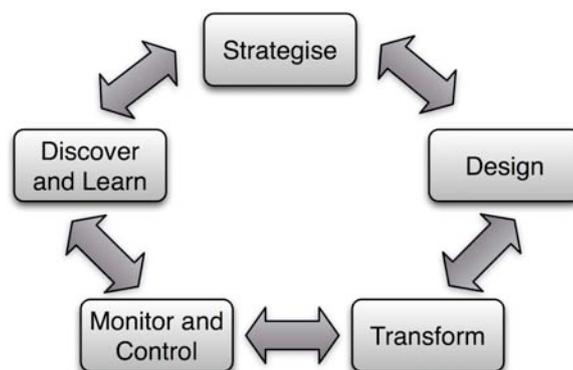


Figure 1. Sustainable Business Transformation Roadmap (adapted from Ahmed & Sundaram, 2009)

C B

However, achieving sustainability by inter-organisational collaboration is subject to many barriers:

L : In an economy with a free market, companies are confronted with many legal barriers that can complicate collaborations. Competition law, especially in New Zealand, has strict rules that govern how organisations are allowed to collaborate. Furthermore, in the context of sharing information, businesses are concerned with their Intellectual Property, which they do not want to give freely to the hands of their competitors. Especially technological infrastructures to support collaboration raise many intellectual property issues (Kolko, 1998).

C : Cultural issues can play an important role in inter-organisational collaboration. One important factor, for instance, is the establishment of one common identity (Dyer & Nobeoka, 2000) without which collaboration is constrained.

S : Besides being concerned with the legal issues that collaboration on Intellectual Property might bring, businesses have a strong interest to only collaborate where it will bring them competitive advantage. Collaboration might be too costly and thereby outweighing the benefits. Organisations also do not want to give information to their competitors, which might strengthen competitors' position in a way that weakens their own.

P : Pragmatic issues that can prevent collaboration can be manifold. First and foremost, if organisations have incompatible processes it is difficult for them to collaborate (Ciborra & Andreu, 2001); if the individuals perceive the collaboration as a burden that interferes with their work practices, the collaboration might not be successful.

T : Technological barriers are concerned with problems that arise from not existing or insufficient information technology to support the collaboration. For instance, if the information systems that the collaborating parties employ are incompatible, collaboration is difficult (Ciborra & Andreu, 2001).

In the context of knowledge dissemination for sustainable business transformation, all these barriers prevent the flow of information from one organisation to the other, which might help them in their efforts to be sustainable (see Figure 2).

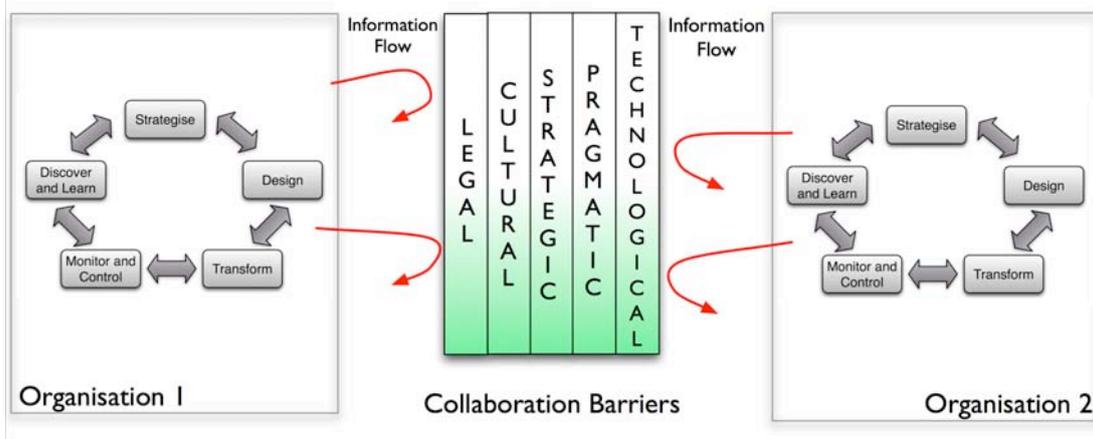


Figure 2. Collaboration Barriers Constraining Knowledge Dissemination

R

Derived from the primary objective of overcoming the collaboration barriers, this research follows three sub objectives:

1. *Explore, Design and Implement* a software **solution** to address **technological** barriers to collaboration.
2. *Explore, Design, Implement and Evaluate* an integrated technological and procedural **platform** that addresses **pragmatic** and **technological** barriers to collaboration.
3. *Explore, Design, Implement and Evaluate* an integrated technological and procedural **platform** that addresses **legal, strategic, pragmatic** and **technological** barriers to collaboration.

A problem in the field of information systems is that systems, methodologies, and processes developed by researchers are difficult to transfer into the domain of practical application (Baskerville & Myers, 2004; Orlikowski & Iacono, 2001). The industry is often ahead of researchers, who should be the ones pointing the directions (Baskerville & Myers, forthcoming). One approach to address this problem that continues to become more and more popular among IS researchers is design science (Baskerville, 2008) in the spirit of Hevner, March, Jinsoo, & Ram (2004) and Nunamaker, Chen & Purdin (1991).

In our research, we want to lay the foundation for a system that New Zealand SMEs can adapt to foster their innovation and transformation into sustainable businesses using information technology.

We base our research process on Nunamaker et al.'s (1991) multi-methodological approach, which is comprised of the phases theory building, systems development, experimentation, and observation. Our adapted methodology suggests three iterations as shown below.

Iteration	Methodology	Data Collection	Data Analysis
1	Design Science	Scenario Tests	Scenario Tests
2	Design Science & Experiments	I. Disseminate prototype through Social Media mechanisms and gather usage data II. Conduct experiments with developed prototype	Analyse data with quantitative methods
3	Design Science & Action Research	Field Notes, Interviews, and Log files	Analyse data with quantitative and qualitative methods

Table 1. Three Planned Iterations

C

In order to guide our research, we have proposed a preliminary artefact in form of a framework for a collaboration platform. In the later stages of our project, this framework, of course, requires further specification and refinement. However, we want to highlight at this stage that any platform that aspires to enable collaboration towards innovation and sustainable development must provide mechanisms to address each of the collaboration barriers discussed above. Figure 3 illustrates our rough framework for

a collaboration platform that provides mechanisms to overcome, account for, or govern the challenges of the collaboration barriers and thereby foster the information flow between individual companies involved in sustainable business transformation processes.

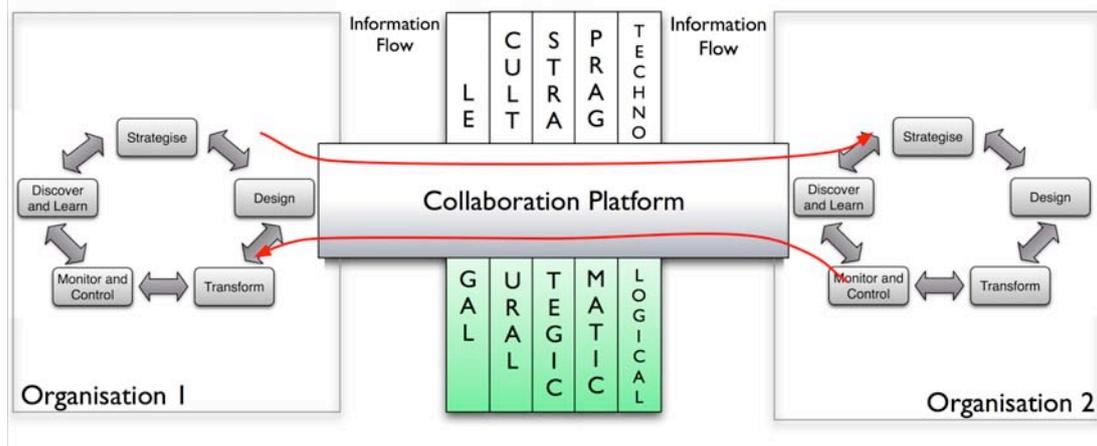


Figure 3. Preliminary Proposed Collaboration Platform

The collaboration platform aims at advancing current research and technologies on the following dimensions in order to overcome the collaboration barriers:

: Although it is a natural way to think of and represent complex and distributed information as a network, modern organisations do not employ systems (people, processes, and technology) that allow the composition or navigation of such information networks. Our platform aims at providing a technological and procedural infrastructure that enables organisations to represent complex multi-paradigm, multi-domain, and multi-dimensional data in network representations.

: The information in the networks is embedded in fine grained and flexible ownership concepts. Information can physically and conceptually belong in the ownership of individual organisations, or by choice, certain parts can be shared and opened up for cross-organisational collaboration.

A : The platform further aims at providing technological and procedural mechanisms that enable organisations to have a clearer overview of their information assets by attributing the networks with discrete monetary and non-monetary value allocations.

: Many tools for knowledge management are designed for a particular domain. Our research is based on the understanding that any approach to capture knowledge in a specific domain leads to knowledge being detached from its context as even knowledge of a particular domain is contextualised with knowledge from an unpredictable set of other domains.

C : Many Knowledge Management initiatives do not account for the context dependent nature of complex information and knowledge (Thompson and Walsham, 2004). The focus of the platform is to provide information with a rich context comprised of related concepts and other information pieces.

: The platform aims at seamlessly integrating into the way individuals would naturally conduct their work. Therefore, a strong focus is set on making the collaboration facilities compatible, integrated, and non-invasive with the business processes in today's businesses (Raghu & Vinze, 2005).

C B T

Following the research objectives and methodology outlined in section 5, our initial exploration aims at the exploration, design and implementation a software solution to address technological barriers to collaboration. We have implanted a prototype of an environment that allows the composition and navigation of heterogeneous information pieces in an integrated semantic network. The prototype can be downloaded at <http://www.linnk.de>.

Figure 4 shows a screenshot of the application. The application is implemented in Java Swing and can thereby be used on most common operation systems. The screenshot shows the document “Demo” in the databases of the application. This document is comprised of a number of items such as the item “Graph Structures”. Some of these items (e.g. the item “Graph Structures”) are linked with further documents. These linkages enable the composition of complex graph or network structures of documents and items.

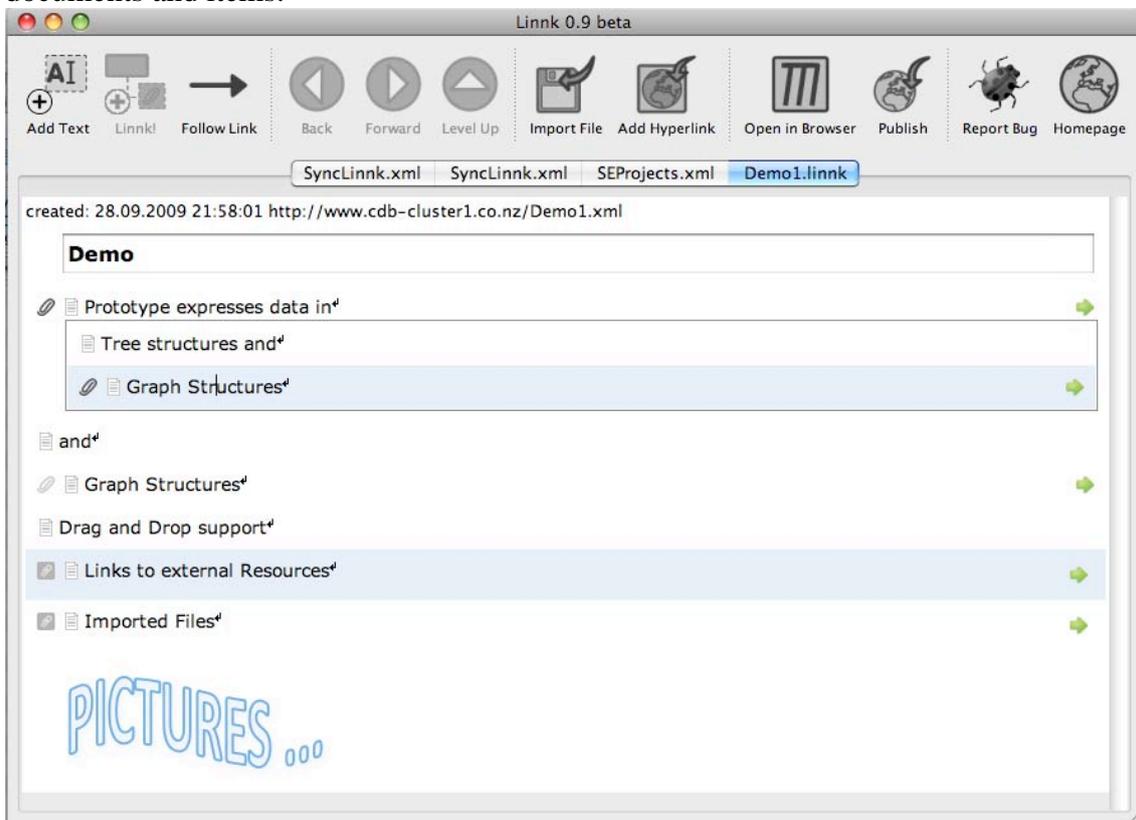


Figure 4. Screenshot of Prototype

Using simple file synchronisation, for instance enabled by standard technologies such as Subversion, or shared network folders using technologies such as WebDav, these documents can be edited collaboratively by a team, inside an organisation, or in inter-organisational collaboration. The software has been tested with databases comprised of thousands of documents.

The central theme driving the implementation is the paradigm that information must not be created or written within the application (as for instance when writing a report with Microsoft® Word) but can be composed by reusing information already created in other contexts. Key ingredients for this embedded in the application are (1) the easy way of adding files edited by external application using drag and drop to the documents and (2) adding hyperlinks to external resources as items to the documents. The software

is thereby designed to work with data originating from diverse applications that can be used to support work following different paradigms and data from different domains. The dynamic graph structure underlying the database allows to model data in simple uni-dimensional structures or as multidimensional data sets enfolding complex information.

C

Achieving sustainability is a complex and challenging task: becoming a profitable organisation, becoming an organisation considering the environment, and becoming an organisation with a positive impact on society are by themselves tasks not easy to achieve. Balancing these dimensions of sustainability requires extensive knowledge of procedural and technological issues spanning many domains.

Our proposed preliminary platform aims at increasing the efficiency of New Zealand SMEs in achieving these tasks by proposing integrated legal, cultural, strategic, pragmatic and technological frameworks and artefacts.

Our initial implemented prototype highlights some of the key directions our project aims to achieve in the course of the next three years. First and foremost, the prototype enables the composition of heterogeneous information pieces from various applications and data sources into one integrated information network. We see this integrated network as an essential stepping stone towards achieving a better understanding of ownership, value, and contextualisation of information that we aspire to explore in the second iteration of our project.

Our research is still on the first steps towards being the foundation for a solution for New Zealand SMEs. However, following the research philosophy of “discovery-through-design” (Baskerville, 2008), we see our initial results as good foundations to guide the further refinement of the technological collaboration platform and development of frameworks to address pragmatic and strategic collaboration barriers.

A

I thank my supervisor David Sundaram for his invaluable contributions to this article in form of ideas, guidance and revision.

R

Baskerville, R. (2008). What design science is not. *European Journal of Information Systems* , 17, 441-443.

Ahmed, M. D., & Sundaram, D. (2009). A roadmap for sustainable business transformation. *The International Journal of Environmental, Cultural, Economic and Social Sustainability* , 5 (2), 165-182.

Baskerville, R. (2008). What design science is not. *European Journal of Information Systems* , 17, 441-443.

Baskerville, R., & Myers, M. D. (2004). Special issue on action research in information systems: Making is research relevant to practice-foreword. *MIS Quarterly* , 28 (3).

BP (2002). *Sustainable development reporting case study* . Auckland: BP New Zealand and New Zealand Business Council for Sustainable Development.

Baskerville, R. L., & Myers, M. D. (forthcoming). Fashion waves in information systems research and practice. *MIS Quarterly* .

- Chetty, S., & Campbell-Hunt, C. (2004). A strategic approach to internationalization: A traditional versus a "born-global" approach. *Journal of International Marketing* , 12 (1), 57-81.
- Ciborra, C. U., & Andreu, R. (2001). Sharing knowledge across boundaries. *Journal of Information Technology*, 16 (2).
- Cole, R., Puro, S., Rossi, M., & Sein, M. (2005). Being proactive: Where action research meets design research. In *ICIS 2005 Proceedings* .
- Dyer, J. H., & Nobeoka, K. (2000). Creating and managing a high-performance knowledge-sharing network: the toyota case. *Strategic Management Journal* , 21 (3), 345-367+.
- Hart, L. S. (1997). Beyond greening: Strategies for a sustainable world. *Harvard Business Review* .
- Hevner, A. R., March, S. T., Jinsoo, P., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly* , 28 (1), 75-105+.
- GR Initiative (2002). Sustainability reporting guidelines. Tech. rep., GR Initiative, Boston, MA.
- Kolko, B. (1998). Intellectual property in synchronous and collaborative virtual space. *Computers and Composition* , 15 (2), 163-183.
- Ministry of Economic Development (2005). *SMALL BUSINESS PORTFOLIO BRIEFING* . New Zealand Ministry of Economic Development.
- Ministry of Economic Development (2009). *SMEs in New Zealand: Structure and Dynamics 2009* .
- Nunamaker, J. F., Chen, M., & Purdin, T. D. M. (1991). Systems development in information systems research. *Journal of Management Information Systems* , 7 (3).
- Orlikowski, W. J., & Iacono, S. C. (2001). Research commentary: Desperately seeking the "it" in it research-a call to theorizing the it artifact. *INFORMATION SYSTEMS RESEARCH* , 12 (2), 121-134.
- Raghu, T., & Vinze, A. (2007). A business process context for knowledge management. *Decision Support Systems* , 43 (3), 1062-1079.
- Srivardhana, T., & Pawlowski, S. (2007). Erp systems as an enabler of sustained business process innovation: A knowledge-based view. *The Journal of Strategic Information Systems* , 16 (1), 51-69.
- Thompson, M. P. A., & Walsham, G. (2004). Placing knowledge management in context. *Journal of Management Studies* , 41 (5), 725-747.
- World Economic Forum (2006). *The global competitiveness report 2006-2007* . World Economic Forum.

A Simulation of the Student Health Centre at the University of Auckland

Kathryn J. Trevor

Department of Operations Research

University of Auckland

New Zealand

ktre013@aucklanduni.ac.nz

Abstract

The aim of this project was to create a realistic model of the Student Health Centre (SHC) at the University of Auckland using the simulation software package, Arena. Data collected regarding patient inter-arrival times and service times as well as employee scheduling and roles have been used as inputs for the realistic model.

Some sub-models are based on concepts developed during previous work on the project however most sub-models are completely unique to the realistic model. An animation demonstrating the flow of patients and employees throughout the SHC has also been created. Due to detailed implementation and rigorous verification, the model is able to deal with all foreseeable variations in patients and employee behaviour at the SHC.

Backlogs of patients occur at the doctors section of the SHC because consultations often take longer than the allocated 15 minute interval. Backlogs of patients also occur at the nurses section of the SHC. This is likely due to some nurses being allocated to triage patients for more than one doctor.

1 Introduction

1.1 Background

The Student Health Centre (SHC) at the University of Auckland offers a selection of free healthcare and counselling services to current students and staff of the university who enrol as patients with the centre. The SHC healthcare facilities often operate at full capacity because of the free services offered. This means the staff employed at the SHC are constantly busy meeting the demands of their patients. This creates several problems

including limited or no lunch breaks for staff, long wait times for patients and unplanned, extended working hours for staff. The SHC situation offers an interesting optimisation opportunity because a realistic model of SHC processes could improve the efficiency of SHC services by providing a tool to determine current problems with their service procedures and be used to test and compare alternative solutions.

1.2 Project Outline

The purpose of this project was to complete the initial steps of a simulation study of the SHC at the University of Auckland. The overall aim of the project was to make a realistic simulation model of the SHC processes using operations research techniques. Some previous work completed by Engineers Without Borders (EWB) (a volunteer engineering society at the University of Auckland) was used for the conceptual basis of some sub-models.

1.3 Arena

The simulation model was developed in Arena which is a simulation software package developed by Rockwell software. In Arena, a flowchart approach is taken to designing simulation models. For more information on specific Arena logic modules refer to the Arena Basic Edition User’s Guide.

1.4 Model Overview

An overview of the SHC model created for this project.

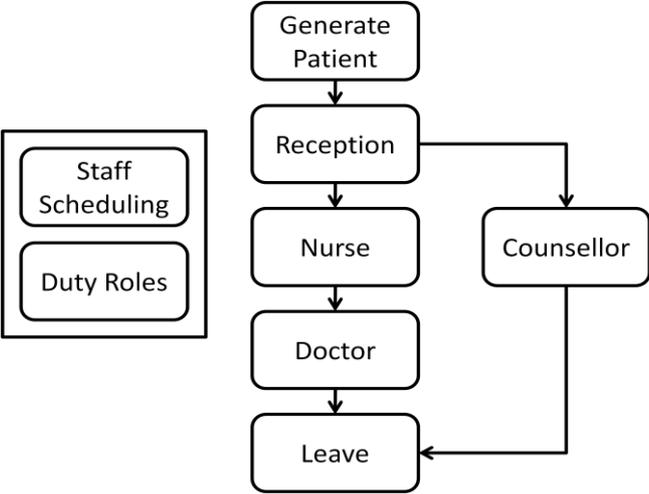


Figure 1: Model Overview

2 Implementation

2.1 Data Analysis and Preparation

Data analysis regarding service times, waiting times and inter arrival times for patients at the SHC was completed by EWB prior to this project. Further data analysis has been completed throughout this project as more data was required to create a realistic model. In particular, data has been collected regarding staff rostering at the SHC – this includes the number and type of staff, their respective roles, how their roles change throughout the day and the days they are scheduled to work. This data is given as inputs to the model to drive patients and employee behaviour.

2.2 Patient Generation

There are eight different types of patients served by the SHC. These are:

1. A patient with a pre-booked doctor's appointment.
2. A patient to see a doctor without an appointment. These patients will make an appointment with the duty doctor if one is available that day.
3. A patient who only visits to make an appointment at reception then leaves the SHC.
4. A patient who only visits to pay a bill at reception then leaves the SHC.
5. A patient who requires an aegrotat note from a doctor. These patients will wait to see the duty doctor.
6. A patient who is in a critical condition has the highest priority at the SHC.
7. A patient to see a counsellor with a pre-booked appointment.
8. A patient to see a counsellor without an appointment. These patients will only see the duty counsellor if there is an available time slot.

Appointments with doctors are 15 minutes in duration while appointments with counsellors are 60 minutes for pre-booked appointments and 30 minutes for walk-in appointments. The SHC services are in high demand and as a result are usually fully booked on the current day. Patients often arrive up to five minutes early or 10 minutes late for appointments and 10% of patients do not arrive for scheduled appointments.

The patient arrival processes was split up into 'create modules' and 'delay modules'. The 'create modules' make a generator for each type of patient. The generators are created at the beginning of the simulation and create patients according to a logical loop that deals with the SHC operating hours. The 'delay modules' delay a new patient according to the respective inter-arrival time for that particular type of patient based on data collected for the project. Once patients are created they are transferred to the reception sub-model where they wait in a queue to be processed.

The model also deals with the special case of creating patients with pre-booked doctor's appointments. The model ensures doctor's appointments are always fully booked on the current day and that 10% of patients do not arrive for scheduled appointments. This accurately represents the SHC situation.

2.3 Employee Sub-models

The sub-models for receptionists, nurses, doctors and counsellors are conceptually based on previous work completed on the project by EWB. Each of the employee sub-models processes patients according to 'seize delay release' logic. This means the entity (i.e. patient) 'seizes' the resource (e.g. nurse), 'delays' the resource for the time required to process the entity then 'releases' the resource. Processing times are based on data collected from SHC for this project. Triangular distributions were used to model processing times because data was scarce, however maximum, minimum and average values were acquired. Critical patients are effectively prioritised over other types of patients throughout the model. The number of employees working in each section of the SHC model at any given time is determined by scheduling for staff sub-model.

The role of reception is to facilitate communication between the different areas of the SHC and to deal with patient queries. The reception sub-model effectively captures the role of reception by transferring patients to their appropriate areas of the SHC. From SHC information it is known that 20% of walk-in patients for doctors do not see the duty doctor and instead book an appointment for a later date. This situation is captured in the reception sub-model. The reception sub-model also effectively distinguishes between the two different types of receptionists at the SHC. Desk receptionists' role is to serve patients while admin receptions have other administrative duties. However, when the SHC becomes busy the admin receptionists assist the desk receptionists. In the reception sub-model, admin receptionists assist desk receptionists when the queue at reception becomes longer than 10 patients.

All patients who are at the SHC to see a doctor are triaged by a nurse prior to their consultation. There are two different types of nurses. A duty nurse triages patients for the duty doctor while an appointment nurse triages patients for an appointment doctor. The nurses' sub-model effectively models the triage process and facilitates the different nurse roles however these roles are determined in the nurse roles sub-model and the duty roles sub-model. Once triage processes are complete, patients join a queue for their respective doctor and wait for their consultation. There are two types of doctors. A duty doctor only consults walk-in patients who do not have a pre-booked appointment and patients who require an aegrotat. There is usually one duty doctor available at all times during the day

however the duty doctor role may be fulfilled by different doctors at different times. If a doctor is not the duty doctor then they are an appointment doctor for patients who have pre-booked appointments. The doctors' sub-model effectively models the consultation process and facilitates the different roles of doctors however the duty role is determined in the duty roles sub-model.

Counselling is another service offered by the SHC. The SHC does not believe the counselling section of the SHC contributes to service issues. Therefore the counselling component of the SHC has been modelled for realistic purposes but has not been investigated regarding service issues for the SHC and hence will not be discussed further.

2.4 Employee Roles and Scheduling Sub-models

There are several types of staff members at the SHC:

- The receptionist team consists of seven receptionists. Two of which are admin receptionists while the remaining five are desk receptionists. Receptionists work full time.
- There are 11 counsellors employed at the SHC that work both part time and full time.
- The nursing team has seven members who work both part time and full time.
- There are six doctors at the SHC who work full time.

Each team of employees adheres to a roster that dictates which days an employee works each week and what times they start and finish their shifts. The roster for the doctors also dictates when (if at all) a doctor has the duty doctor role during the week. The nurses and doctors have another roster that allocates the pairing of nurses and doctors (Note: these pairings may change throughout the day). This particular roster is also used to establish the nurse roles. It is SHC policy that at least one receptionist must remain working until all counsellors, nurses and doctors have finished consulting their patients. This policy is implemented in the model.

The scheduling for staff sub-model creates each employee as well as an individual scheduler for each employee. These individual schedulers are sent through logical flow that determines the behaviour of each individual employee. The scheduler first checks if that employee is working on the current day. The scheduler for that employee is then delayed until the employee starts work. The scheduler is again delayed until the employee leaves for lunch and is delayed for a third time until the employee finishes their shift for that day. Counters were implemented to determine how many patients are in the system for particular employees. The counters are used to determine specific conditions that permit specific employee behaviour such as an employee finishing their shift for the day or taking their lunch break. Several realistic situations have been dealt with in the scheduling for staff

sub-model. In regard to lunch breaks, receptionists can only go for lunch if there is no one in the queue for reception or if the current receptionist has no current patients and another receptionist will be at the reception desk while the current receptionist takes their lunch break. Counsellors, doctors and nurses are permitted to take their lunch break when they are idle and have no patients in the system according to their respective counters.

To implement duty role changes a duty role controller was created. This controller assigns the duty doctor role or duty counsellor role to an appropriate employee. The controller then checks for any duty role changes later that day. If there is another duty change that day, the controller waits and updates the change accordingly. If the last change for the day has been completed, the controller is delayed until the following day.

Nurses may be allocated to complete the triage process for a different doctor in the afternoon than in the morning. Consequently, nurse roles were implemented in the nurse roles sub-model. A controller is created that loops through all nurses employed at the SHC, generating a role scheduler for each one. The role scheduler allocates a nurse's first role for the day based on data stored in the scheduling for staff sub-model. The role scheduler is delayed until the duration of the first role is complete then updates the nurse's role to their second role for the day. (Note: some nurses will be assigned to the same doctor for the entire day and hence their second role is identical to their first). Once the role scheduler for a particular nurse has updated all the role changes for that nurse, the role scheduler waits until the following day.

3 Animation

The animation section simulates the physical behaviour of patients and employees at the SHC. The animation was a useful communication tool with SHC employees because employees could easily understand the model using the animation. The animation also assisted in finding and fixing problems in the model.

3.1 Practical Description

There are several pathways a patient may follow when they enter the SHC. All patients check in with reception. All types of patients for doctors or counsellors then wait in the first waiting room to be called upon by a counsellor or to be triaged by a nurse. Patients who visit a counsellor may return to reception after their consultation to pay or book another appointment and then leave the SHC. Patients who are to see a doctor are triaged then wait in a second waiting room to be called upon by a doctor. Once they have seen the doctor they may return to reception to make a payment or book a future appointment then leave the SHC. Some patients only interact with reception because they only need to make a

payment or book an appointment. There is also a flow of employees through the centre. Employees arrive for their shifts and depart once their shift is complete. Employees also leave their work station when they have lunch.

3.2 Arena Animation

The Arena animation shows the flow of patients through the SHC (Refer to **Figure 2: Animation**). Patients enter the SHC and queue at reception. Patients who need to see a doctor will then join the triage queue for their respective nurse. Once triage is complete, patients join the queue for their respective doctor. Once the consultation with the doctor is complete some patients may reappear at the queue for reception because they need to make a payment or book a future appointment. Patients who need to see a counsellor join the queue for their respective counsellor after checking in with reception. The animation shows individual patients being processed by receptionists, nurses, receptionists and counsellors respectively (Refer to **Figure 2: Animation**). The flow of employees is shown instantaneously in the animation. All employees appear when they are at work and disappear when their shift is finished or they are at lunch (Refer to **Figure 2: Animation**).

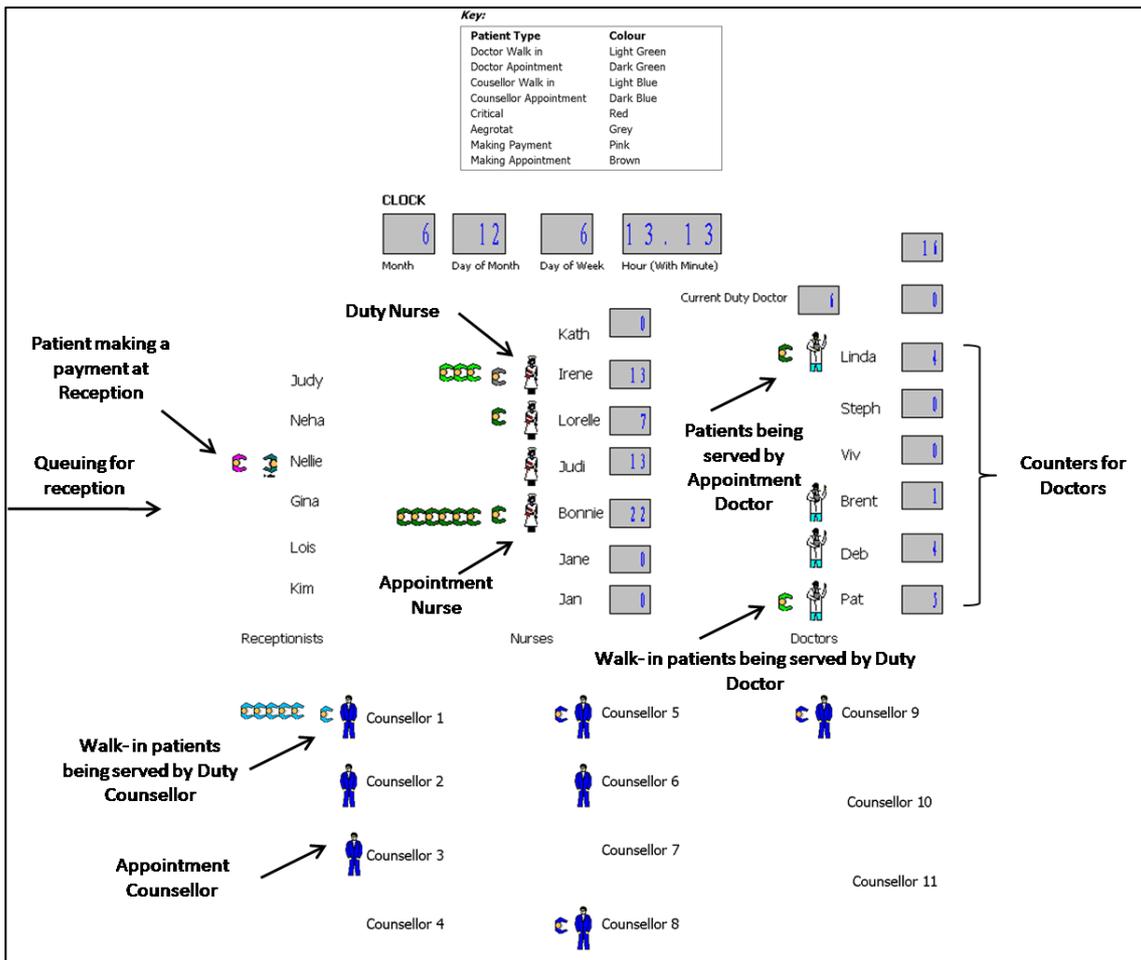


Figure 2: Animation

4 Model Verification

4.1 Debugging Approach

The aim of the verification process for this project was to ensure the simulation accurately represented the practical situation at the SHC. The debugging process involved tracking entities (both patients and employees) to determine that their behaviour within the model reflected the realistic situation at the SHC. Error checking was also completed to ensure all patients were effectively dealt with by the model. When an error was detected the model logic was updated to account for the error situation. Changes made to the model included:

- Checking whether a walk-in patient for the duty doctor had instead decided to book an appointment to see a doctor at a future later date in the Nurses Sub-model.
- Checking if a critical patient arrived while the duty nurse was unavailable. In this situation the model ensures the critical patient is attended to by another nurse.
- Checking if a critical patient arrived while the duty doctor was unavailable. In this situation the model ensures the critical patient is attended to by another doctor.

Using the initial data relating to service times for triaging and doctor consultations caused the model to ‘explode’. Therefore artificial service times that were less than the actual service times were temporarily introduced to test model behaviour when service times were within the model’s capacity.

5 Results

5.1 A Realistic Model

A realistic model of the medical services at the SHC at the University of Auckland has successfully been developed using Arena. The model is realistic because it uses data collected for the project to drive the behaviour of patients and employees within the model. An extensive model verification process was completed and resulted in the model being able to deal with all foreseeable changes in employee and patient behaviour that may occur at the SHC. The counsellor sub-model is a simplification of the real world situation because it was the focus of this project to make a realistic simulation of the medical facilities at the SHC. The model is dynamic because new rosters for employees can easily be added, and changes in employee numbers can be implemented.

5.2 Observations from Simulation

In the simulation a large backlog of patients waiting for the doctors was observed. This observation was expected because the service times for a doctor’s consultation usually

exceed the given 15 minute interval. Consequently, all doctors at the SHC are overworked (Refer to Figure 3: Overtime for Doctors). Furthermore, output regarding the average proportion of time an employee is busy completing a task when they were working at the SHC shows that doctors are busy at least 87% of the time. Table 1: Utilization of Doctors shows the confidence intervals relating to this statistic for each doctor. The output suggests that the overtime completed by doctors is due to an excessive workload at the SHC (and not due to, for example, doctors working long hours but not being busy for most of their shift).

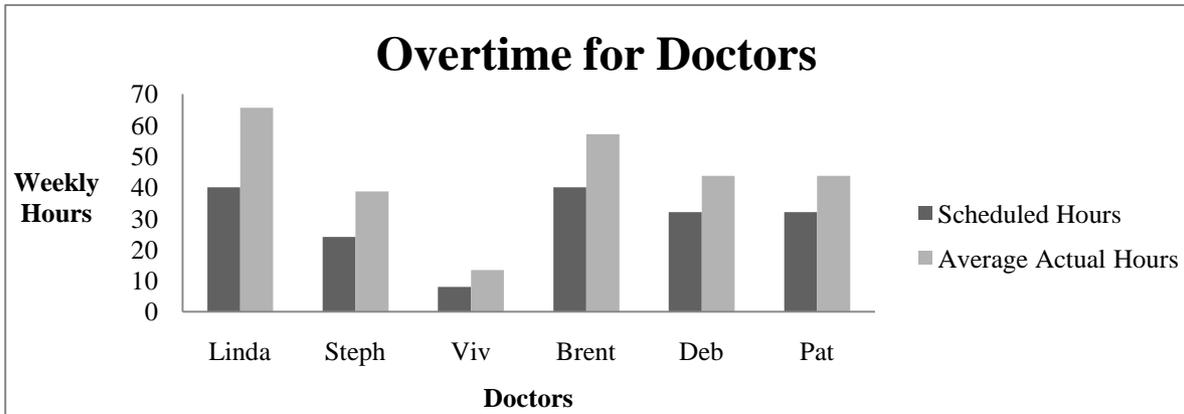


Figure 3: Overtime for Doctors

Utilization of Doctors		
Doctor	Lower Limit	Upper Limit
Linda	87%	90%
Steph	87%	92%
Viv	85%	90%
Brent	91%	95%
Deb	91%	95%
Pat	94%	97%

Table 1: Utilization of Doctors

A significant backlog of patients also occurs at the nurses. This was not expected from observations of the data for triage service times. Figure 4 shows that some of the nurses at the SHC are overworked. Nurses do not appear to be as busy as doctors but are still busy for a significant proportion of their shift (Refer to Table 2: Utilization of Nurses). Output provided by Figure 4 combined with information about nurse scheduling and the utilization of nurses suggests that the backlog of patients in the triage queues is likely to be due to nurses completing triage of patients for more than one doctor.

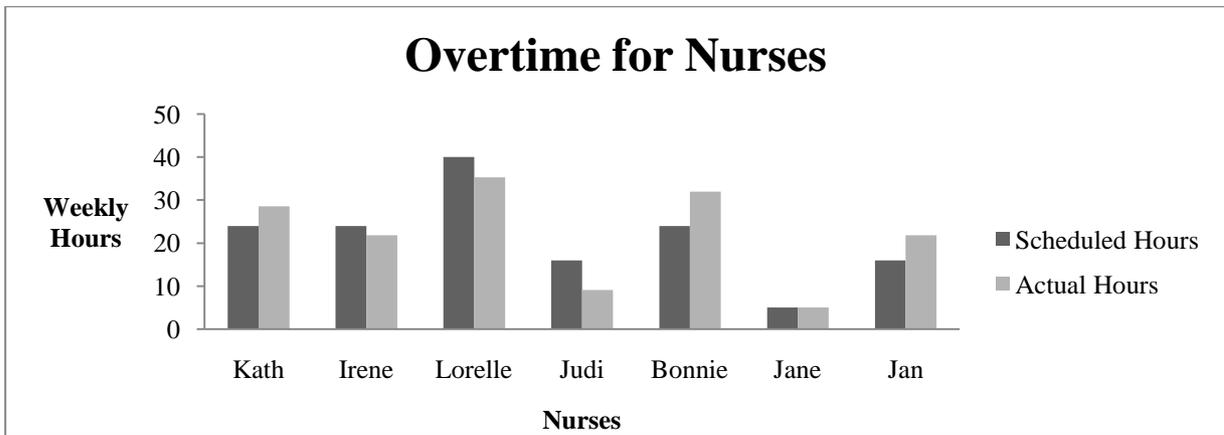


Figure 4: Overtime for Nurses

Utilization of Nurses		
Nurse	Lower Limit	Upper Limit
Kath	69%	74%
Irene	58%	65%
Lorelle	63%	65%
Judi	24%	31%
Bonnie	77%	80%
Jane	37%	45%
Jan	58%	64%

Table 2: Utilization of Nurses

Acknowledgements

Special thanks to Dr Mike O’Sullivan and Dr Cameron Walker for their consistent support throughout the project. Thanks also to the staff at the SHC and the EWB team for their previous work on the project.

6 References

Kelton, W.D., R.P. Sadowski. And D.T. Sturrock. 2004. *Simulation with Arena*, 3rd Edition, McGraw Hill Co., Inc., New York.

Improve Efficiency of OR Applications with ILOG CPLEX

Liu Huan Dong
ILOG Technical Account Manager
lhong@cn.ibm.com

Abstract

Discover how you can improve efficiency, quickly implement strategies and increase profitability with ILOG CPLEX. ILOG CPLEX's mathematical optimization technology enables better decision-making for efficient resource utilisation. Learn how advanced optimization algorithms can allow you to rapidly find solutions to complex business problems which can be represented as mathematical programming models.

Operating Theatre Optimisation

Oliver Weide

The Optima Corporation

o.weide@theoptimacorporation.com

Abstract

The aim of Optima's theatre optimisation is to allow hospitals to plan the best long-term allocation of surgical procedures and surgeons to operating theatres. The optimisation creates a cyclical theatre roster specifying the surgeons and procedures that should be allocated to each session in each theatre. The roster is then repeated over the entire planning period.

The optimisation objective is usually chosen to minimise waiting list length or to achieve government contracts which specify the number of each type of procedure to be carried out.

The feasible solution space is limited by business and logical rules. A procedure may require the patient to stay in a ward pre- or post-surgery and the procedure can only be scheduled if a bed is available. The theatre sessions during which procedures can be performed have fixed start and end times and the type of procedure that can be performed in each theatre is limited by the theatre's equipment. Likewise, surgeons can only perform procedures they are qualified for and they can only be scheduled when available.

We present an optimisation model that is used to obtain optimal solutions to practical problems via a branch and cut solution process.

MIP-based Heuristics for the Capacitated Lot-Sizing Problem with Startup Times

Shaorui Li

Department of Logistics Management
School of Business Administration
Southwestern University of Finance and
Economics, Chengdu, 610074, China
lisr@swufe.edu.cn

Cheng Peng

Department of Computer Science
and Information Engineering
Tourism and Cultural College
of Yunnan University
Lijiang, 674100, China

Abstract

This paper deals with the multi-item capacitated lot-sizing problem with startup costs and startup times. We propose an MIP-based fix-and-relax approach that combines the branch-and-bound method and several strategies of variable-specific horizon decomposition to achieve greater efficiency and to produce good solutions. In addition, a tight formulation for building submodels is obtained to speed up the branch-and-bound strategy. The solution quality and computational performance of the developed approach are tested with reference to various problem instances of differing characteristics and sizes.

Key words: production planning, lot-sizing and scheduling, startup times, MIP-based heuristics

1 Introduction

Applications of lot-sizing problems often arise in practices such as production planning, supply chain management, etc. In this study, we address the multi-item capacitated lot-sizing problem with additional startup costs and startup times if a setup is incurred in a period but not in its preceding period. The objective aims at finding the sizes and timing of orders under the restriction of resource capacities so as to minimize the startup, setup, production and inventory holding costs while satisfying all demands without shortages. The problem considered in this study is even harder to solve to optimality than the general multi-item capacitated lot-sizing problem, which belongs to a class of \mathcal{NP} -hard (Florian, Lenstra, and Kan 1980).

Since the seminal paper by (Wagner and Whitin 1958) on the uncapacitated problem, various forms of lot-sizing problems have been investigated for decades. A recent overview on models and algorithms can be found in (Karimi, Ghomi, and Wilson 2003), researches on modeling industrial extensions are reviewed by (Jans and Degraeve 2008), and different formulations, extensions and the problem complexity of the single-item problem are discussed in (Brahimi et al. 2006).

Despite extensive work on general problems (see, (Wolsey 2002) and (Eppen and Martin 1987) among others), MIP-based heuristic approaches on big-bucket capacitated lot-sizing problems with startup times are limited, see, e.g., (Karimi, Ghomi, and Wilson 2003). However, for large scale real-world problems the ability to efficiently construct high-quality schedules is crucial for companies in order to be competitive, see, e.g., (Vanderbeck 1998).

Some works motivate interests to better understand the multi-item capacitated lot-sizing problem with additional startup costs and startup times. Wolsey (Wolsey 1989) considers the uncapacitated lot-sizing problem with startup costs, and shows how equivalent, or possibly stronger, formulations are obtained by introducing auxiliary variables. A model with changeover costs presented by (Karmarkar, Kekre, and Kekre 1987) is also tackled by (Wolsey 1989). Vanderbeck (Vanderbeck 1998) discusses the models and their complexity for the multi-item lot-sizing problem with continuous setup and sequence independent startup times. Good solutions on average less than 2% are obtained by combining a column generation approach and a branch-and-cut approach. However, both capacities and startup times are assumed to be constant in his models.

On the other hand, since real-world lot-sizing problems are often computationally very difficult, the exact solution techniques presented previously are only effective for small or medium sized instances (Akartunali and Miller 2008). Therefore, heuristic approaches for lot sizing problems remain an interesting research area (van den Heuvel and Wagelmans 2005). Lot-sizing is well-suited for fix-and-relax heuristics since in its structure early decisions affect later decisions, for recent literature, see, e.g., (Suerie and Stadtler 2003), (Federgruen, Meissner, and Tzur 2007), and (Absi and Kedad-Sidhoum 2007). Moreover, as the key issue in tackle mixed-integer programming problems is the difficulty of solving large-scale instances requires too many binary variables, in the fix-and-relax subproblem, a shorter interval for binary variables and a longer one for continuous variables will lead to higher computational speed.

Therefore, different with many previous studies, in our fix-and-relax approach, a strategy of variable-specific horizon decomposition is proposed to combine with mathematical programming to achieve greater efficiency and to obtain good solutions for the multi-item capacitated lot-sizing problem with startup times. In addition, a tight formulation for building submodels is obtained to speed up the branch-and-bound strategy.

The outline of this paper is as follows: Section 2 introduces the assumptions, notation, problem formulation and a strong reformulation. Section 3 proposes a fix-and-relax heuristic that combines with the obtained tight formulation for building submodels. Section 4 describes the results of computational experiments. Section 5 gives a summary account.

2 Model Formulation

2.1 Problem definition

The capacitated lot-sizing problem with startup costs and startup times considered here is based on the following assumptions that will hold throughout this paper:

- All time periods are of equal length. Time-varying demands appear in every period are prespecified by the higher-level decisions.

- A startup occurs when the production line is setup for a product in a period, but is not setup in its previous period. A setup can be carried over to the next period if production of the same product is continued.
- In addition to the production time the resource capacity limit may be further reduced by setup and startup times.
- A charge is incurred for maintaining a unit of item in inventory. This charge is referred to as the inventory holding cost, which is computed based on the end-of-period inventory.
- Backlogging is not allowed, so demand in a period must be met either by production in the period or by carrying inventories from earlier periods.
- Lead times are negligible.
- To simplify the notation and without loss of generality the initial physical inventory of every item is set to zero.

Our goal is to determine a production schedule that minimizes the startup, setup, production and inventory holding costs under the resource restriction.

2.2 The model

The following notation will be used throughout the rest of the paper, other symbols will be introduced as need.

Indices

t = period;
 p = product.

Sets

\mathcal{P} = set of products/items.

Parameters

T = length of the planning horizon;
 f_t^p = setup cost refers to direct labor and material costs that are required each time the production line is setup to make the indicated product;
 c_t^p = variable production cost of item p in period t , *i.e.*, the unit cost of item p ;
 h_t^p = positive unit cost of holding stock or inventory for item p at the end of period t ;
 s_t^p = startup cost item p in period t ; g_t^p = setup time for item p in period t ;
 u_t^p = startup time for item p in period t ;
 a_t^p = production time for manufacture item p ;
 d_t^p = prespecified demand of item p in period t ;
 M_t = resource capacity in period t .

Decision variables

I_t^p = continuous variable which represents the inventory or stock of item p held over at the end of period t . Its nonnegativity ensures that backorders are not permitted;

x_t^p = continuous nonnegative variable which represents the amount of item p produced in period t ;

y_t^p = binary variable indicates whether or not we produce item p in period t ;

w_t^p = startup variable indicates whether or not we startup the production line for item p in period t .

Therefore, the capacitated lot-sizing problem with startup costs and startup times under consideration can be formulated as follows:

Model MCLSS

$$\text{Minimize } \sum_{t=1}^T \sum_{p \in \mathcal{P}} (f_t^p y_t^p + c_t^p x_t^p + h_t^p I_t^p + s_t^p w_t^p) \quad (1)$$

$$\text{subject to } I_{t-1}^p + x_t^p = d_t^p + I_t^p \quad \forall p \in \mathcal{P}, t = 1, \dots, T, \quad (2)$$

$$\sum_{p \in \mathcal{P}} (a_t^p x_t^p + g_t^p y_t^p + u_t^p w_t^p) \leq M_t \quad \forall t = 1, \dots, T, \quad (3)$$

$$a_t^p x_t^p \leq (M_t - g_t^p) y_t^p - u_t^p w_t^p \quad \forall p \in \mathcal{P}, t = 1, \dots, T, \quad (4)$$

$$\min\{y_t^p, 1 - y_{t-1}^p\} \geq w_t^p \geq y_t^p - y_{t-1}^p \quad \forall p \in \mathcal{P}, t = 1, \dots, T, \quad (5)$$

$$I_0^p = 0 \quad \forall p \in \mathcal{P}, \quad (6)$$

$$x, I \in \mathbb{R}_+^{|\mathcal{P}| \times T}, \quad (7)$$

$$y, w \in \mathbb{Z}^{|\mathcal{P}| \times T}.$$

where, the initial setup state of item p , say y_0^p , is known.

Explanation The objective function (1) attempts to minimize the total cost, which consists of the setup, production, inventory holding and startup costs. The inventory balancing constraint (2) together with (7) ensures that demands are met without backlogging, so demands must be satisfied either by producing in the same period or by carrying inventories from early periods. From constraint (3) one can see that the sum of setup and production times in each time period is enforced by the corresponding capacity restriction. Constraint (4) indicates that an appropriate setup cost is paid whenever a production occurs. Constraint (5) establish connections between setup and startup indicators: a start up occurs when the machine is set up for an item for which it was not set up in the previous period. Finally, constraint (6) gives the initial inventories.

2.3 A facility location representation

A general formulation for the capacitated production planning problem with startup costs and startup times has now been introduced, we then try to strengthen model MCLSS by employing a facility location representation, see, e.g., (Brahimi et al. 2006). The basic idea of the facility location representation is that an item-period combination can be regarded as a facility at a location. Only if the facility (means item) at the location (means period) is open (means setup), may the demand of the current period for a particular item be produced.

Therefore, the capacitated production planning problem with startup costs and startup times can be reformulated as follows:

Model MCLSSFL

$$\text{Minimize} \quad \sum_p \sum_i \sum_j \left(\sum_{\ell=i}^{j-1} h_\ell^p + c_i^p \right) d_j^p \theta_{ij}^p + \sum_p \sum_i (f_i^p y_i^p + s_i^p w_i^p) \quad (8)$$

$$\text{subject to} \quad \sum_i \theta_{ij}^p = 1 \quad \forall p \in \mathcal{P}, j = 1, \dots, T, \quad (9)$$

$$\sum_p \sum_j a_i^p d_j^p \theta_{ij}^p + \sum_p (g_i^p y_i^p + u_i^p w_i^p) \leq M_i \quad \forall i = 1, \dots, T, \quad (10)$$

$$\theta_{ij}^p \leq y_i^p \quad \forall p \in \mathcal{P}, i, j = 1, \dots, T, \quad (11)$$

$$\sum_{\ell=i}^j \theta_{\ell j}^p \leq y_i^p + \sum_{\ell=i+1}^j w_\ell^p \quad \forall p \in \mathcal{P}, i, j = 1, \dots, T, \quad (12)$$

$$\min\{y_t^p, 1 - y_{t-1}^p\} \geq w_t^p \geq y_t^p - y_{t-1}^p \quad \forall p \in \mathcal{P}, t = 1, \dots, T, \quad (13)$$

$$0 \leq \theta_{ij}^p \leq 1, \quad \forall p \in \mathcal{P}, i, j = 1, \dots, T,$$

$$y, w \in \mathbb{Z}^{|\mathcal{P}| \times T}.$$

where θ_{ij}^p denotes the fraction of the demand of item p in period j that is produced in period i .

Explanation The objective function (8) minimizes the sum of the inventory holding, production, setup and startup costs. Constraint (9) implies that the total fraction of the demand in period j that is produced in all period i is one. Constraint (10) enforces the resource capacity restriction in each time period. Constraint (11) together with (9) ensures that for every item p at least one of the y_i^p will be nonzero. Constraint (12) together with (13) enforces that a set up can be carried over to the next period if production of the same product is continued.

We expect that the above model formulation MCLSSFL will work well when it is incorporated into a fix-and-relax heuristic approach.

3 Heuristic Approaches

Due to the high complexity of the problem we are studying, a heuristic approach will be the topic for our investigation. In our proposal a branch-and-bound based fix-and-relax heuristic is developed to address the problem.

3.1 Basic idea

The heuristic approach developed in this study is different from many common sense heuristics, which start with an initial solution and try to improve it, it starts with no solution and try to find one. Such a fix-and-relax strategy partitions the planning horizon into three time windows (see, e.g., (Federgruen, Meissner, and Tzur 2007) and (Absi and Kedad-Sidhoum 2007)): the frozen, operation and relaxing windows.

This study constructs the operation window with variable-specific horizon decompositions, then solve a problem over the progressively operation window to minimize

the total cost incurred up to the end of the current window. The basic idea of our fix-and-relax heuristic approach is based on twofold:

- As has been mentioned previously, in the structure of lot-sizing early decisions affect later decisions, so lot-sizing is well-suited for fix-and-relax heuristics.
- To cope with the difficulty of solving large-scale mixed-integer program requires too many binary variables, a shorter interval for binary variables and a longer one for continuous variables can be designed to speed up the computational performance.

We always start with period one, freeze the selected variables of the progressively time window after their local optimal values are obtained at each iterations, while relaxing the other variables. Therefore, when the time window of frozen variables is expanded iteratively, the number of computing variables remains constant.

3.2 Definition of submodels

The developed formulation MCLSSFL is tight as its linear programming relaxation has an optimal solution in which the setup variables are integer. It is employed to be combined with the heuristic procedure for building submodels.

We further reduce the complexity of the overall model by building a smaller submodel within each operation window. Each subproblem of a time window is specified to include decisions which adequately complement the solutions obtained from prior windows. The selected decision is implemented only for the first several periods of a operation window. The operation window is then progressively go forward to the first future interval not covered by these production decisions. At each iteration, we only consider decisions in the operation window, solve a submodel of reduced size, thereby the computational burden is limited.

3.3 Heuristics

The optimum seeking mathematical programming methodology is employed by our horizon decomposition strategy to solve a problem over a progressively time window. In each step, the decision of selected variables is implemented only for the first several periods of an interval, therefore two consecutive operation windows are overlapped. Selected decisions in an overlapping window will be resolved at the next iteration. In our implementation, shorter operation intervals for the setup variables (heuristic I) and for both the startup and setup variables (heuristic II), but longer intervals for other variables are designed in submodels.

Note at each iteration, we only consider the selected decisions in the operation interval, solve a submodel (by means of a default branch-and-bound methodology) of reduced size, thereby the computational burden is limited. At every iteration, the full flow balance and capacity constraints are considered in the model formulation.

4 Numerical Experiments

We now report some computational experiments. All runs have been carried out on an Intel's Core 2 Quad Q6600 (under Windows xp) with 2.4GHz processor and 2GB of RAM. with a commercial programming software CPLEX, version 7.1.

4.1 Test instances

We systematically generate benchmark instances to test that capability of the developed approach, a total of 45 instances were generated as follows: Five sets of nine problem instances of $|\mathcal{P}| = 12, 24, 36, 48, 60$, and $T = 60$ are obtained using data with s_t^p, c_t^p, h_t^p integers uniformly distributed in $[75, 125]$, $[3, 5]$, $[1, 5]$, respectively. Analogously, the setup, startup times g_t^p, u_t^p integers are chosen out of interval $[2, 10]$, $[1, 4]$, respectively, with uniform distribution. The production time a^p is chosen out of interval $[0.8, 1.2]$ with uniform distribution. All demands d_t^p are independently generated from a normal distribution with mean 100 and standard deviation of 10.

Consider the time between orders (abbreviated TBO) for lot sizing without startup costs, the setup cost and inventory holding cost are two major cost parameters that are varied to insure a range of order cycle lengths. Assuming that the demand is constant, and recalling that the unit inventory carrying cost is h , we can change the setup costs to have different TBO. The fixed setup cost of a product p is indirectly determined by first choosing the EOQ-cycle time $TBO = (2f^p/h^p\bar{d})^{\frac{1}{2}}$ from a uniform distribution on interval $[1, 3]$, when considering low TBO values; interval $[2, 6]$ for medium TBO values; $[5, 10]$ for high TBO values.

We systematically varied the utilization of capacity using low, medium and high levels of the capacity density $\rho = \frac{\sum_{t=1}^T C_t}{\sum_p a^p \sum_{t=1}^T d_t^p}$ that respectively correspond to 50%, 75% and 90% of resource utilization. In this way we can obtain the values of capacity. Note here the capacity utilization is an estimate only, because startup times are not considered. Hence, a value of 90% actually means that the resource utilization by startup actions is greater than 90% on average.

4.2 Results

For each problem instance that is designed in the experiment, we run the solution heuristic for the lot-sizing model as defined previously. Five sets of test problems each consists of nine combinations which arise when combining the three production densities and three TBO values.

Table 1 reports an excerpt of gaps and CPU seconds of the developed heuristics for the test instances. We observe that the solution gaps differ with capacity utilization in lower and medium levels, but increase with high level capacity utilization. CPU times increase with capacity utilization, particularly these increases are more apparent for high level capacity utilization.

We can see that problems with higher TBO values are more difficult to solve with respect to CPU time. The effect on the gap of a low and medium TBO seems minor, whereas the effect of high TBO values is more apparent.

The computational results indicate that the developed relax-and-fix heuristics appear to perform very well, the duality gap is never in excess of 3% and on average within 0.37%, and the CPU time of all 45 problem instances does not exceed 335 seconds on average. Summarizing this section we conclude that the developed heuristic approach generates superior solutions within a reasonable amount of time for different types and sizes of data.

Table 1: An Excerpt of Computational Results

D ^a	T ^b	LB(opt)	Heuristic I			Heuristic II		
			Time	Best	Gap	Time	Best	Gap
L	L	2884580.00	95.79	2884599.00	0.00%	89.15	2884599.00	0.00%
L	M	4245220.00	104.42	4246879.00	0.04%	87.09	4248526.00	0.08%
L	H	7041937.10	110.59	7100234.00	0.83%	92.06	7103144.30	0.87%
M	L	2793291.00	97.31	2793447.00	0.01%	89.32	2793319.00	0.00%
M	M	4294195.20	134.31	4295785.81	0.04%	90.01	4296918.18	0.06%
M	H	8234849.40	449.28	8294875.00	0.73%	225.56	8305724.00	0.86%
H	L	2553763.04	169.34	2555099.52	0.05%	93.78	2555394.03	0.06%
H	M	4428656.67	302.71	4434963.33	0.14%	106.43	4436892.57	0.19%
H	H	8185908.94	919.73	8252372.15	0.81%	337.50	8257280.24	0.87%

a. Levels of low, medium and high correspond respectively to 50%, 75% and 90% of resource utilization.

b. Times between orders that is randomly chosen out of the interval [1, 3] with uniform distribution for low levels, [2, 6] for medium levels, and [5, 10] for high levels.

5 Conclusion

In this paper, we have presented two MIP-based fix-and-relax heuristics to solve the multi-item capacitated lot-sizing problem with startup costs and startup times. This hybrid approach is based on an optimum seeking branch-and-bound strategy and two strategies of so called variable-specific decomposition of planning horizon. To be combined with the heuristic algorithm a tight formulation is obtained for building submodels. The solution quality and computational performance of the developed approach have been tested with reference to various problem instances of differing characteristics and sizes, and the capability and good solution quality are proved. As a future work we are planning to carry out an industrial extension to some real-world production planning problems.

Acknowledgments

This research was partly supported by grants from the National 211 Project Foundation for Southwestern University of Finance and Economics.

References

- Absi, N., and S. Kedad-Sidhoum. 2007. "MIP-based heuristics for multi-item capacitated lot-sizing problem with setup times and shortage costs." *RAIRO-Operations Research* 41:171–192.
- Akartunali, K., and A.J. Miller. 2008. "A heuristic approach for big bucket multi-level production planning problems." *European Journal of Operational Research*, p. doi:10.1016/j.ejor.2007.11.033.
- Brahimi, N., S. Dauzere-Peres, N.M. Najid, and A. Nordli. 2006. "Single item lot sizing problems." *European Journal of Operational Research* 168(1):1–16.

- Eppen, C.D., and R.K. Martin. 1987. "Solving multi-item capacitated lot-sizing problems using variable redefinition." *Operations Research* 35(6):832–848.
- Federgruen, A., J. Meissner, and M. Tzur. 2007. "Progressive interval heuristics for multi-item capacitated lot sizing problems." *Operations Research* 55:490–502.
- Florian, M., J.K. Lenstra, and A. H.G. Rinnooy Kan. 1980. "Deterministic production planning: Algorithms and complexity." *Management Science* 26(7):669–679.
- Jans, R., and Z. Degraeve. 2008. "Modeling industrial lot sizing problems: A review." *International Journal of Production Research* 46(6):1619–1643.
- Karimi, B., S.M.T. Fatemi Ghomi, and J.M. Wilson. 2003. "The capacitated lot sizing problem: a review of models and algorithms." *Omega* 31:365–378.
- Karmarkar, U.S., Sham Kekre, and Sunder Kekre. 1987. "The dynamic lot sizing problem with startup and reservation costs." *Operations Research* 35(3):389–398.
- Suerie, C., and H. Stadtler. 2003. "The capacitated lot-sizing problem with linked lot sizes." *Management Science* 49(8):1039–1054.
- van den Heuvel, W., and A.P.M. Wagelmans. 2005. "A comparison of methods for lot-sizing in a rolling horizon environment." *Operations Research Letters* 33:486–496.
- Vanderbeck, F. 1998. "Lot-sizing with start-up times." *Management Science* 44:1409–1425.
- Wagner, H.M., and T.M. Whitin. 1958. "Dynamic version of the economic lot size model." *Management Science* 5:89–96.
- Wolsey, L.A. 1989. "Uncapacitated lot-sizing problems with start-up costs." *Operations Research* 37(5):741–747.
- . 2002. "Solving multi-item lot-sizing problems with an MIP solver using classification and reformulation." *Management Science* 48(12):1587–1602.

Area restricted forest harvesting with adjacency branches

Alastair McNaughton
Department of Mathematics
University of Auckland
Private Bag 92019
New Zealand
a.mcnaughton@auckland.ac.nz

David Ryan
Department of Engineering Science
University of Auckland
Private Bag 92019
New Zealand
d.ryan@auckland.ac.nz

Abstract

Aspects of constraint branching and column generation will be discussed in relation to forest harvesting applications. The emphasis will be on recent technical advances with supportive performance evidence.

Key words: forest harvesting, adjacency branches, area restriction model.

1 Introduction

The optimization of forest harvesting is a classic problem in operations research. Because of the long planning horizon and the immense amount of data involved the number of decision variables is extremely large. Moreover these decision variables are integer. Sustainability and conservation issues necessitate a very large number of constraints, many of which are site-specific. We deal here with the complex issue of area-restricted clearfell. As this is another chapter in a continuing line of research, I assume readers are already familiar with the basic ideas involved in the optimization of forest harvesting. The main progress achieved since my last talk has been the publication of a major paper jointly with David Ryan in Forest Science, McNaughton and Ryan (2008). This present paper is largely a comment on some of the technical aspects of the Forest Science paper.

2 The Model

$$\text{Minimize} \quad -\sum (c_{jn} - c'_{jn})x_{jn} \quad (1)$$

subject to

$$\sum_{ckx_{jn} \in H_{jk[t,t]}} a_k x_{jn} + s_t - s'_t = A_t, \quad t = 1, t^* \quad (2)$$

$$\sum_{ckx_{jn} \in H_{jk[t,t]}} y_{kt} a_k x_{jn} + {}^*s_t - {}^*s'_t = Y_t, \quad t = 1, t^* \quad (3)$$

$$\sum_n x_{jn} + x_j^* = 1, \quad j = 1, R \quad (4)$$

where x_{jn} is the 0/1 decision variable associated with harvest plan n on road j ,
 c_{jn} is the present net worth associated with plan n on road j excluding set-up costs
 c'_{jn} is the discounted set-up costs associated with plan n on road j

a_k is the area in hectares of unit k ,

A_t is the area, in hectares, to be harvested in time period t ,

y_{kt} is the yield per hectare if unit k is harvested in time period t ,

Y_t is the target yield for time period t ,

s_t , *s_t , s'_t and ${}^*s'_t$ are appropriately bounded slack and surplus variables,

x_j^* is a 0/1 variable representing a null harvest on road j ,

R is the number of roads,

t^* is the planning horizon,

T is the length of the green-up period,

and $H_{jk[a,b]}$ is the set of all harvest plans on road j in which unit k is harvested in $[a, b]$.

Each unit lies on a particular road segment. Each variable, x_{jn} , represents a harvest plan for all the units on a certain road and encompasses all the time periods in the planning horizon. The solution shown in Figure 2 illustrates several of these road harvest plan variables.

The objective coefficient corresponding to x_{jn} is $c_{jn} - c'_{jn}$. Here c_{jn} represents the yield and revenue to be obtained from this plan minus the related harvesting costs. These costs and revenues, are separable to individual units. The component c'_{jn} refers to the one-off set-up costs needed before any harvesting can begin on this road. These costs are not separable to units. The full set-up cost is due if 1 or more units are harvested. Typical set-up costs are the cost of constructing skid sites and spur roads. Both c_{jn} and c'_{jn} are discounted to represent the present net worth.

The *area constraints* given in Equation 2 bound the area harvested in each time period. If a separate strategic plan has already been developed, then the areas to be harvested each time period are the A_t coefficients. The *yield constraints* are given in Equation 3. Y_t , the yield for time period t , may also be obtained from the strategic planning. Strict implementation of area and yield constraints would impede the solution algorithm. An elastic constraint, included in Equations 2 and 3, solves these problems. Refer Vielma et al. (2003).

The *road plan constraints* as given in Equation 4 ensure that only one road harvesting plan be included in the solution for each road in the forest. This ensures each unit is harvested at most once during the planning horizon.

3 A nuclear set

The concept of a *nuclear set* is almost the same as the concept used by Gunn and Richards (2005) for their central stand, except that there is a plurality of units in the central stand and every perimeter unit is required to be individually of sufficient area to form an adjacency violation in association with the central stand. A nuclear set has two parts: a nucleus and a perimeter. The *nucleus* is a contiguous block of units with total area less than or equal to the maximum clearfell area. The *perimeter* consists of the surrounding units which are adjacent to the nucleus such that the total area of nucleus plus each individual perimeter unit exceeds the maximum clearfell area.

An example of a nuclear set is given in Figure 1. This detail taken from the 400 unit forest is circled in Figure 2. The maximum clearfell area is 30 hectares. Units 75 and 95 form the nucleus, and units 55, 56, 74, 94 and 114 are the perimeter. Unit 96 is not part of the perimeter since its area is too small. A nuclear set is not restricted to the units along 1 road. The example given in Figure 1 spans 4 roads. The nucleus has similarities with the Generalized Management Unit of McDill et al. (2002).

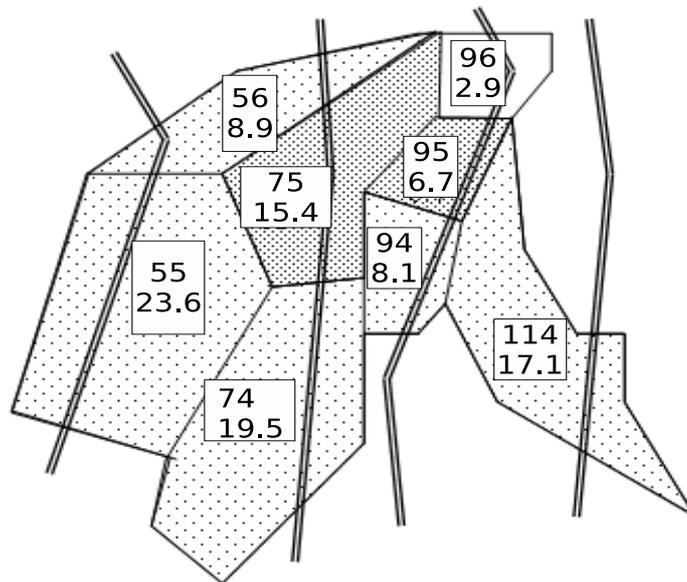


Figure 1: A nuclear set with nucleus (dark shaded) and perimeter (light shaded).

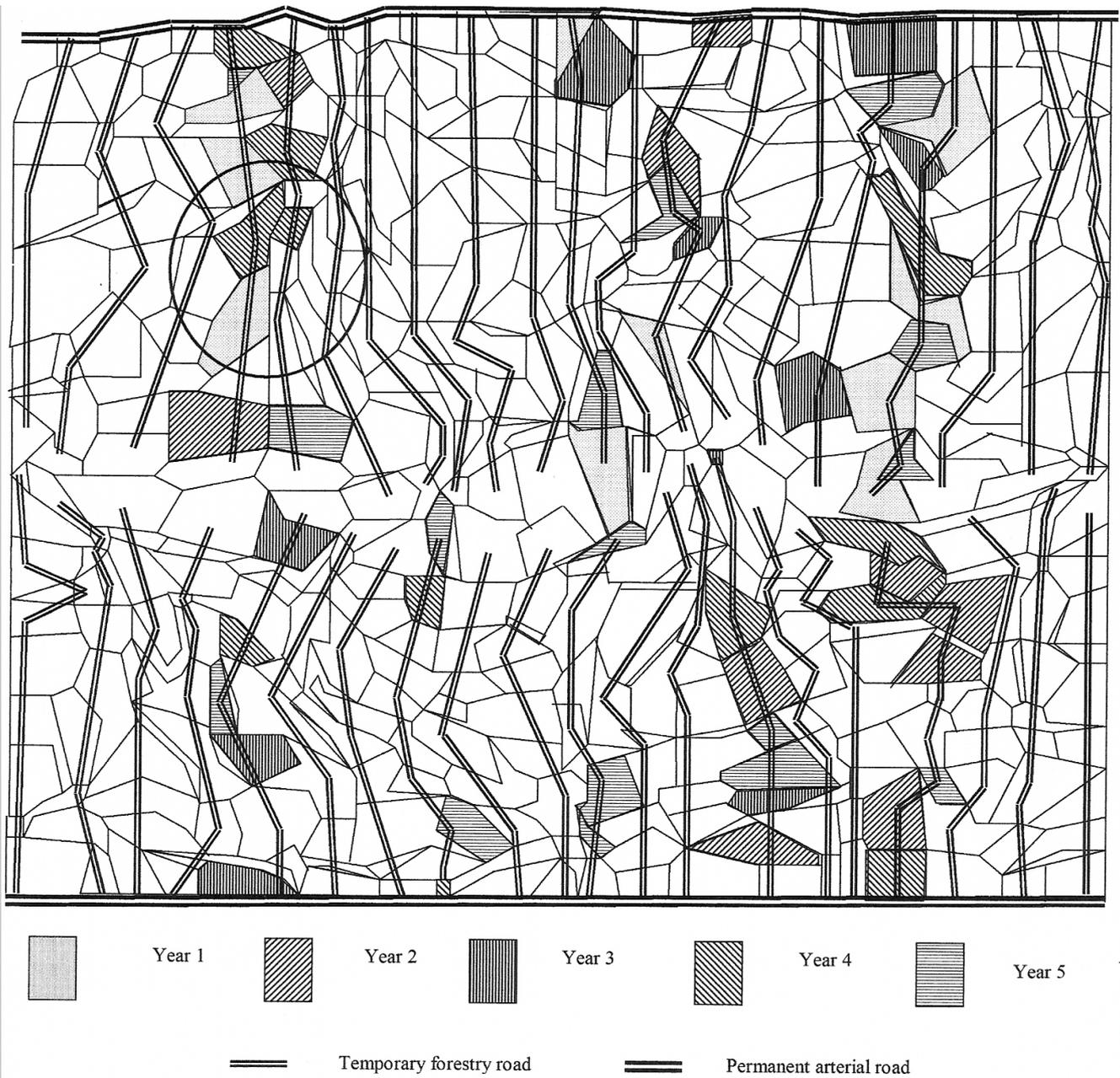


Figure 2: A 400 unit poorly regulated forest. The solution for the trial using a planning horizon of 5 time periods with a green-up of 2 time periods is shown.

4 The algorithm

The solution algorithm deals with the planning horizon as a single optimization. The term *relaxed LP* refers to the current LP at the given stage of the solution pro-

cess including all columns added by column generation and all nuclear constraints (given in Equation 5 below) added by constraint generation along with all adjacency and integer branches currently active. The strategy is as follows. A list is made of all possible nuclear sets. This is done as that described by McDill et al. (2002) in relation to their path algorithm. Murray et al. (2004) have shown that such a list can be compiled for quite large applications. We then wait to see which of these will be required. For example, the nuclear set in Figure 1 would be included in our list at this stage. However, it would only be used in the solution algorithm if at some stage a relaxed LP solution happened to contain it as part of an adjacency violation.

Figure 3 presents a flow diagram. The forest harvesting problem is initially formulated and solved as a relaxed LP, with no adjacency or integer requirements. A phase of column generation follows with a number of composite road harvest plan variables being added to the model. Optimization (2) follows. The solution obtained is searched so as to find any cases of adjacency violation. This is done by a simple scanning process in which each nuclear set on the list is checked sequentially time-wise. An infringement is detected in the form of an identification of a nuclear set with a time interval $[a, b]$, where all the units in the nucleus are harvested within the interval $[a, b]$, with $b - a < T$, and at least 1 unit in the perimeter harvested during the time interval $[b - T + 1, a + T - 1]$. Each iteration only 1 infringement is dealt with and an appropriate nuclear constraint of the form given in Equation 5 below is added.

An adjacency branch is associated with each nuclear constraint. In the 1-branch all units in the nucleus are felled within the time interval $[a, b]$. Concurrently with this, all the units in the perimeter are left unharvested during the appropriate time interval, $[b - T + 1, a + T - 1]$. After each nuclear constraint and the associated adjacency branch has been implemented, the modified linear programme is re-solved, with more column generation as required. During optimization (1) the new branch is enforced by an artificial penalty on the decision variables which are to be removed by the branch. Column generation then produces several new columns. After this the decision variables which are being removed have their upper bounds set at 0 and optimization (2) follows. If the solution obtained is unacceptable perhaps due to fathoming or to infeasibility, then we back-track. This involves replacing the 1-branch with the 0-branch. The process is repeated until no adjacency violations can be detected. At this stage the solution may still contain fractional values of the decision variables. So integer branches are then used until an integer solution with an acceptable objective value has been obtained. Then the problem is re-optimized with more column generation after every branch. During this integer branching, regular checks are made to detect any further adjacency violations with further adjacency branches are implemented as required. A feasible integer solution is obtained once no adjacency violations and no fractional values of decision variables are left. For small applications the process is continued until the entire branch and bound tree has been searched. For larger applications a near-optimal solution is found by choosing the best out of the first few integer solutions, following Gunn and Richards (2005).

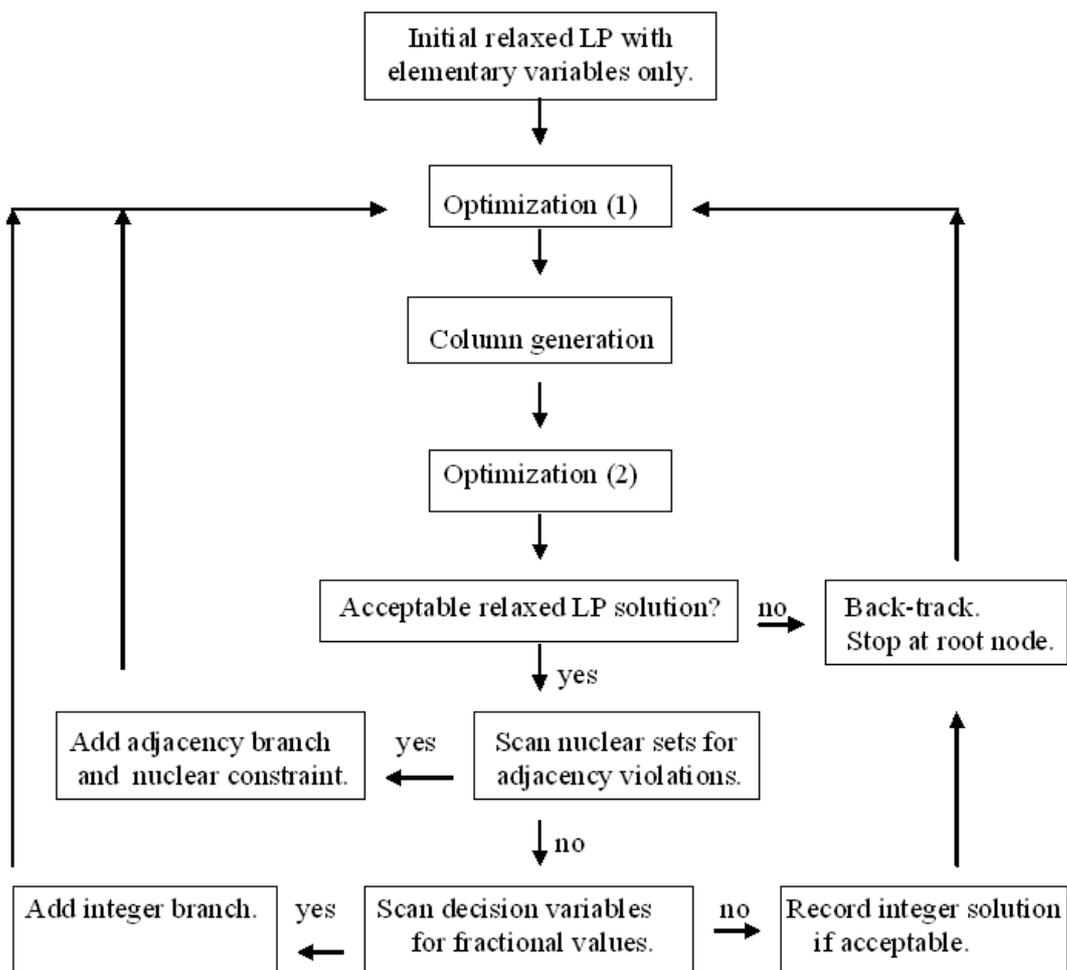


Figure 3: A flow chart of the solution algorithm.

5 Adjacency branches

After each episode of column generation, each nuclear set on the previously prepared list is scanned. Let us suppose an adjacency violation is found with the nucleus

felled during a time interval $[a, b]$. If there are several we select the one with the nucleus of greatest yield. We impose the 1-branch which forces the nucleus units to be harvested within the time interval $[a, b]$. A new constraint specific to this particular branch is now added. This *nuclear constraint* is

$$\sum_{ck \in N, x_{jn} \in H_{jk[a,b]}} x_{jn} + s_{N[a,b]} = N^* \quad (5)$$

where N is the nucleus of the n -th nuclear set on the previously prepared list, N^* is the order of the n -th nuclear set, $[a, b]$ is the time interval within which the nucleus units are harvested, with $b - a < T$, and $s_{N[a,b]}$ is a slack variable associated with the adjacency branch on the n -th nuclear set relative to the time interval $[a, b]$.

Nuclear constraints are not included at the initial stage but are added to the LP one at a time, as required during the solution process in response to specific nuclear set violations. These are not adjacency constraints in the usual sense of the term, but serve an important role in the adjacency branching.

To implement we impose an upper bound of 0 on every road harvest plan variable that represents the harvesting of any of the nucleus units *outside* the interval $[a, b]$, including the possibility of a null harvest. Also we require there to be no harvesting of any of the perimeter units during the period $[b - T + 1, a + T - 1]$, where T is the number of time periods in the green-up. To do this we impose an upper bound of 0 on all road harvest plan variables that represent the harvesting of any of the perimeter units *inside* the interval $[b - T + 1, a + T - 1]$. During branching steps, if an upper bound of 0 is imposed on certain composite road harvest plans, say on the set $H_{jk[a,b]}$, then the corresponding elementary variables, x_{jkt} with $t \in [a, b]$, are flagged. If the branch is removed then the flag also is removed. An elementary variable is not used in the column generation process when flagged.

As an example of an adjacency branch, consider the nuclear set shown in Figure 1 and identifiable in the solution shown in Figure 2, where it is circled. The green-up is 2 time periods, that is $T = 2$, and the planning horizon is 5 time periods. The solution indicates units 75 and 95 are to be harvested in time period 4. The time interval for the nucleus is $[a, b] = [4, 4]$. The corresponding forbidden time interval for the perimeter is $[b - T + 1, a + T - 1] = [3, 5]$. During the solution generation the relaxed LP solution had unit 74 in the perimeter being harvested within this forbidden interval, thus causing an adjacency violation. So an adjacency branch was implemented forcing the nucleus units to be harvested *in* the interval $[4, 4]$, and forcing the perimeter units to be harvested *outside* the interval $[3, 5]$. In the final solution we see unit 74 is harvested in year 1.

During back-tracking it may be required to explore the associated 0-branch. If the nucleus contains N^* units, the 0-branch is imposed, requiring that at most $N^* - 1$ of these units are harvested within the time interval $[a, b]$. This is done by placing a lower bound of 1 on the slack variable $s_{N[a,b]}$ in Equation 5. At the same time all the upper bounds of 0 imposed during the 1-branch are removed.

The adjacency branches also tend to remove fractions from the relaxed LP and so work harmoniously with the integer branches. In the trials the algorithm gave precedence to adjacency branches, followed by integer branches. Nuclear sets

corresponding to parts of the forest desirable for harvesting take integer values at an early stage during the algorithm. As a consequence they make adjacency branches which occur early in the branch and bound tree. In this way the branching is prioritized in a good way.

The following aspects of the process are significant. No explicit adjacency constraints are used, thus avoiding issues concerning the preferred manner of construction of these constraints, and the associated tightness of the formulation. There is no requirement that all the units in a nucleus, and/or all the units in the associated perimeter must belong to the same road harvest plan. There is no requirement that all the units in a nucleus be harvested during the same time period. Because the model only uses branches which correspond to actual adjacency violation, the number of adjacency branches is relatively small.

6 Integer branches

Integer branches remove fractional values from the variables, x_{jn} , in the relaxed LP, so as to obtain an integer solution. First, the required harvesting time for each unit is obtained from the current relaxed LP solution. If the unit is completely harvested in one particular year then no branch is required. However, for some units the result will be a series of fractions, relative to the time periods, which sum to 1. These are spread over a time interval, say $[a, b]$. We find the smallest fraction associated with either a or b across all the possible units. If the smallest fraction is associated with unit k on road j in time period a , then the 1-branch requires unit k to be harvested after time a . Any variable x_{jn} in which unit k is harvested up to time period a is removed from the problem by having its upper bound set at 0. The associated 0-branch requires unit k to be harvested no later than time period a . The algorithm may require a large number of integer branches. However, because each integer branch produces only a tiny change in the objective value, the quality of the integer solution tends to be very high.

7 Some results

The algorithm has been tested with a variety of data set simulations comprising forests of 400 and 1600 units. Various forest types have been used so as to test the robustness of the model. The treatment of green-up spanning up to 5 time periods is very significant. Table 1 presents some typical output in this case with respect to a large forest comprising 1600 units.

horizon (time)	green-up periods)	RLP objective	objective (million \$ US)	upper bound	optimality gap	time (seconds)	nuclear constraints
forest type: poorly regulated							
25		71.11					
	0		71.11				
	1		70.79	71.06	0.27	429	361
	2		70.65	70.82	0.17	405	490
	3		70.21	70.39	0.18	467	577
	4		69.47	69.81	0.34	674	698
	5		68.97	69.30	0.33	383	744
forest type: well regulated							
25		62.15					
	0		62.10				
	1		61.82	62.06	0.26	444	442
	2		61.67	61.80	0.13	105	521
	3		61.12	61.40	0.28	357	604
	4		60.69	60.77	0.08	177	602
	5		60.22	60.33	0.11	170	627
forest type: over mature							
25		81.97					
	0		81.90				
	1		81.74	81.91	0.17	513	249
	2		81.47	81.66	0.19	606	370
	3		81.08	81.24	0.16	814	522
	4		80.68	80.75	0.07	664	597
	5		80.14	80.19	0.05	316	695

Table 1 : Results from trials with a forest of 1600 units.

8 Conclusions

The level of resolution in the trials indicates that this approach compares favourably with other models in terms of fine detail, with regard to both time and area. The number of time intervals used, 25, is larger than those in the literature. The occurrence of nuclear sets order 7 is a level of area detail far beyond that dealt with elsewhere. As a consequence, the amount of time or area aggregation required in any practical application will be limited. This enables the adjacency requirements to be met in a precise way.

It is possible the methods presented here may be useful in some other applications involving massive combinatorial complexity.

References

- [1] Gunn, E.A. and E.W. Richards. 2005. *Solving the adjacency problem with stand-centered constraints*, Can. J. For. Res. 35, pp 832-842.
- [2] McDill, M.E., S.A. Rebaun and J. Braze. 2002. *Harvest Scheduling with Area-Based Adjacency Constraints*, Forest Science 48, pp 631 - 642.
- [3] McNaughton, A.J., M. Rönqvist and D.M. Ryan. 2000. *A Model which Integrates Strategic and Tactical Aspects of Forest Harvesting*. In System Modelling and Optimization, Methods, Theory and Applications, Edited by M.J.D. Powell and S. Scholtes, Kluwer Academic Publishers Boston, pp 189-208.
- [4] McNaughton, A.J., G.D. Page and D.M. Ryan. 2001. *Adjacency Constraints in Forest Harvesting*, proceedings of the ORSNZ, 2001, pp 9-15.
- [5] McNaughton, A.J. 2002. *Optimisation of Forest Harvesting Subject to Area Restrictions on Clearfell*, proceedings of the ORSNZ, 2002, pp 307-313.
- [6] McNaughton, A.J. 2003. *Adjacency constraints and adjacency branches*, proceedings of the ORSNZ, 2003, pp .
- [7] McNaughton, A.J. 2004. *Recent Progress on the Area Restriction Problem of Forest Harvesting*, proceedings of the ORSNZ, 2004, pp .
- [8] McNaughton, A.J. and D.M. Ryan. 2007. *Area restricted forest harvesting with adjacency branches* , proceedings of the ORSNZ, 2007, pp
- [9] McNaughton, A.J. and D.M. Ryan. 2008. *Adjacency branches used to optimize forest harvesting subject to area restrictions on clearfell* , Forest Science 54(4), 2008, pp 442 - 454.
- [10] Murray, A. 1999. *Spatial Restrictions in Forest Scheduling*, Forest Science 45(1), pp 45-52.
- [11] Murray, A.T. and A. Weintraub. 2002. *Scale and Unit Specification Influences in Harvest Scheduling with Maximum Area Restrictions*, Forest Science 48, pp 779-789.
- [12] Murray, A.T., M. Goycoolea and A. Weintraub. 2004. *Incorporating average and maximum area restrictions in harvest scheduling models*, Can. J. For. Res. 34, pp 456-464.
- [13] Vielma, J.P., A.T. Murray, D. Ryan and A. Weintraub. 2003. *Improved Solution Techniques for Multiperiod Area-based Harvest Scheduling Problems*, Systems Analysis in Forest Resources: Proceedings of the 2003 Symposium, pp 285-290.

Initial use of discrete event simulation for New Zealand military workforce analysis

Nebojsa Djorovic
New Zealand Defence Force
Wellington, New Zealand
nebojsa.djorovic@nzdf.mil.nz

Michelle Gosse
Ministry of Defence
Wellington, New Zealand
michell.gosse@nzdf.mil.nz

Jason Markham
Royal New Zealand Air Force
Wellington, New Zealand
jason.markham@nzdf.mil.nz

Jay Ta'ala
New Zealand Army
Wellington, New Zealand
jay.ta'ala@nzdf.mil.nz

Abstract

Discrete event simulation ('DES') has recently been adopted by NZ Defence Force workforce analysts. The paper reviews the use of general purpose DES software for the case of the Air Force engineering officer trade and asks how this method was able to satisfy the client's requirements. As was anticipated, this initial project produced the degree of realism expected by the client while also providing the model builders with greater programming flexibility than existing methods. A potential role for steady-state workforce models is suggested in the early stages of DES analysis and some implications of the use of a generic DES model are raised. Overall, the initial success of the generic DES workforce model clears the way for further development of this method by the NZ Defence Force.

Key words: Air Force, Arena, discrete event simulation, engineering officer, military, workforce planning

1 Background

There are a broad range of operational research methods available for analyzing workforce problems (Wang, 2005), but as there is “no best model” for workforce planning (Purkiss, 1981: 321) workforce planners must select the methods that satisfy the needs of their clients. Workforce analysts of the NZ Defence Force, in collaboration with their foreign counterparts, have recently starting applying discrete event simulation (‘DES’) to workforce planning problems. With the development and initial use of a prototype DES workforce model this paper asks how this tool has been able to satisfy the requirements of NZ military workforce planning clients.

Most large organisations undertake some form of workforce planning and mathematical models can be a useful decision aid in this field of planning (Gass, 1991). Early workforce planning models developed from the use of actuarial life tables (Vajda, 1970) before analysts had the use of desktop computers and these early models were usually of an aggregate form (Bennison & Casson, 1984). Possibly the first problem analysed with a computerized entity-based workforce simulation model was the training of US Air Force pilots (Mooz, 1969), while non-military applications have typically been for large government organisations, such as the UK Civil Service (Wishart, 1976). DES has generally not been applied to the problems of small workforce groups.

The NZ Defence Force has traditionally employed spreadsheet models that use aggregate quantities and Markov rate calculations to estimate the flows through workforce categories. Attempts were made to introduce system dynamics models, also an aggregate method, by replicating the Markov technique using the aging chain process (Sterman, 2000). However, at the present time most workforce analysis is undertaken using spreadsheets. McLucas (2002) has offered several reasons for this lack of success including the scarcity of aggregated data, expectations of detailed results, and a tendency for military clients to interpret results as predictive. Another possible explanation is that NZ military workforce groups are typically small, whereas successful system dynamics applications tend to be macro-level problems involving qualitative variables and feedback loops (e.g. Forrester, 1973). For these reasons it is perceived that NZ workforce analytical needs exceed the limits of aggregate simulation methods.

NZ Defence Force interest in entity based methods springs from several sources. The complex flow networks and variables that constitute the workforce suggest the need for a commensurate level of modelling complexity. It has been found that workforce models require frequent modification, either due to changes in the problem being analyzed or changes in the variables considered necessary. DES is a common method of analysis for inventory systems (e.g. Law, 2007) and that is how workforce problems have traditionally been conceptualized. Furthermore, the stochastic variables of DES better suit the analysis of small populations (e.g. Bartholomew, et al., 1980) so it is on good grounds that NZ Defence Force workforce analysts expect some success with DES.

2 Engineering Officer Trade

There are three main segments within the workforce of the NZ Defence Force: regular force, civilian and reserves. The regular force segment presents the greatest challenge to workforce planners because it is managed primarily through internal replacement. This constraint imposes long delays in responding to skill shortfalls due to long recruitment times, demanding training requirements and slow upward progression through numerous ranks. All regular force personnel are trained and employed within a military trade as this determines their operational utility. The point at which a workforce group because 'small' is generally taken to be 100 (e.g. Edwards, 1983). In the case of the Air Force regular force workforce, trades smaller than 100 people account for 50% of all personnel and 90% of all trades, and this statistic is representative of trade size distributions in the Navy and Army.

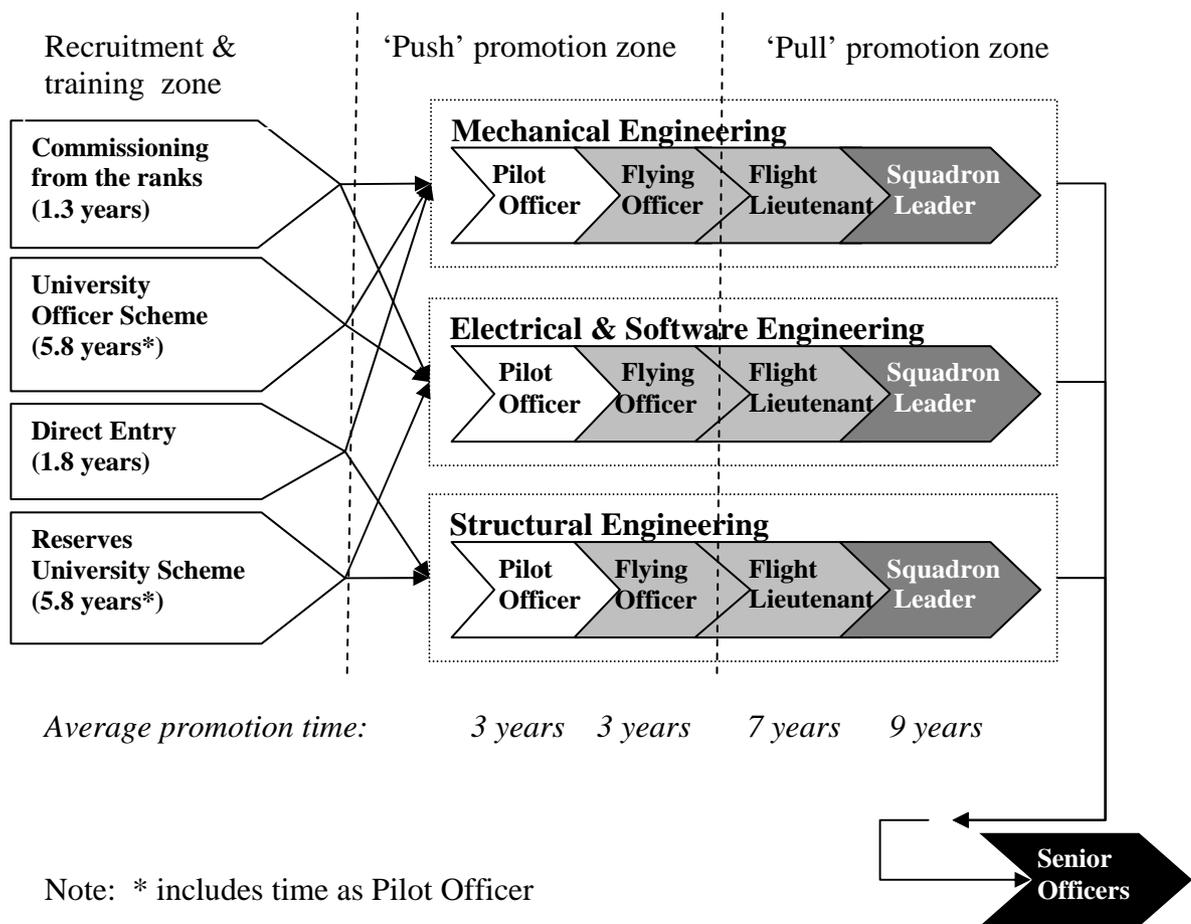


Figure 1. Engineering officer trade structure and flows

The Air Force employs regular force engineering officers to design and manage the maintenance and modification of military aircraft. There are four recruitment streams and three employment streams for this trade, as shown in Figure 1 above. For the last five years the shortfall of mid-rank engineering has ranged between 20% and 30% of

posts. Over the last 10 years the total number of engineering officers, including trainees, has averaged 102 and ranged between 90 and 116. This is a small workforce group given the complexity of its personnel flows, as the number of individuals within most rank and stream categories is typically less than 10. While the potential for recruiting sufficient replacements is apparently high due to the four separate recruitment streams, this however introduces a degree of planning complexity and risk.

In addition to the structural complexity in Figure 1, analysis of the workforce must also account for the distinct promotion zones within the workforce. Until promotion to Flight Lieutenant, promotion occurs automatically according to time served, while promotion beyond Flight Lieutenant is determined by vacancies within that stream. As spreadsheet and system dynamics models can both handle this combination of promotion rules it has been assumed that DES would simulate this behaviour in a realistic manner.

3 Methodology

Three models were built in the process of analysis, as shown in Table 1. A prototype model was first built to verify the ability of DES to replicate existing spreadsheet models. In stage 2, a preliminary spreadsheet model of the engineering officer trade was developed to estimate the ideal recruitment and training through-flow. The full DES model was developed by adding a number of features to the prototype model to satisfy the client requirement for realism. As the generic DES model becomes more sophisticated it is anticipated that the need modifications will reduce significantly. Staging of the model-building process helped build confidence in the analytical results; through the gradual clarification of business rules, refinement of data and debugging of the models.

Table 1. Stages of model building

Stage	Details
1. Prototype DES model	A model was developed to satisfy the first stage of generic user requirements of Defence Force workforce analysts. The prototype was tested using dummy data and reviewed by peer model-builders.
2. Steady state spreadsheet model	The ideal personnel flows for the engineering officer trade were estimated for a hypothetical steady state situation in which all rank targets are achieved and all personnel flows are in equilibrium. A bespoke spreadsheet model was developed and an optimum throughflow solution obtained using an ‘add-in’ macro.
3. Full DES model	The prototype model was modified to include the four entry streams, three employment streams and push promotion policy.

While a steady state spreadsheet model was only developed in this case at the client's request, it proved to be worthwhile for several reasons. The spreadsheet model was simple enough for the client to understand without specialist knowledge and so they were able to verify input data and reach a common understanding with the model builder at an early stage. Furthermore, the model was simple enough for the client to operate without support and consequently they were able to experiment with simple workforce variables and so improve the mutual understanding of system behaviour.

A 'general purpose simulation package' (see Law, 2007: 187-213) was chosen for this application so as to collaborate with counterpart analysts in foreign forces. The Arena package (Rockwell 2005), currently offered by Rockwell Automation, is a commonly reviewed simulation package (e.g. Banks et al., 2001 and Law, 2007) and links to Microsoft Office applications for desktop use. The DES model was developed in version 12 using a monthly time step and a simulation run time of 20 years.

Table 2. Summary of model testing

Model	Tests	
1. Steady state spreadsheet model	1.1	Conservation of flows
	1.2	Historical comparison of key data
	1.3	Direction and degree of input-to-output effects
	1.4	Behaviour under extreme input values
2. Full DES model	2.1	Step-wise switching-on of the main input data groups in the sequence: starting population, entries, exits and promotions
	2.2	Replication of spreadsheet results for t=20 years, i.e. end of run
	2.3	Behaviour under extreme input values
	2.4	Observing animations and tracing individual entities

A summary of the formal modelling tests is included in Table 2. Tests 1.1 and 1.2 were programmed into the spreadsheet model for the client to check whenever they used the model. Test 2.2 provided compelling verification of the DES model because it showed whether that model tended towards a steady-state situation given the same recruitment inflows and attrition rates.

4 Results

The outputs of a simulation run from the full DES model are shown in Figure 2 on the following page for the structural stream of the engineering officer trade. The time series displayed are the mean of 100 replications and include the target for their respective rank. Flight Lieutenants and Flying Officers have been combined because the positions for both ranks are bracketed for personnel of either rank.

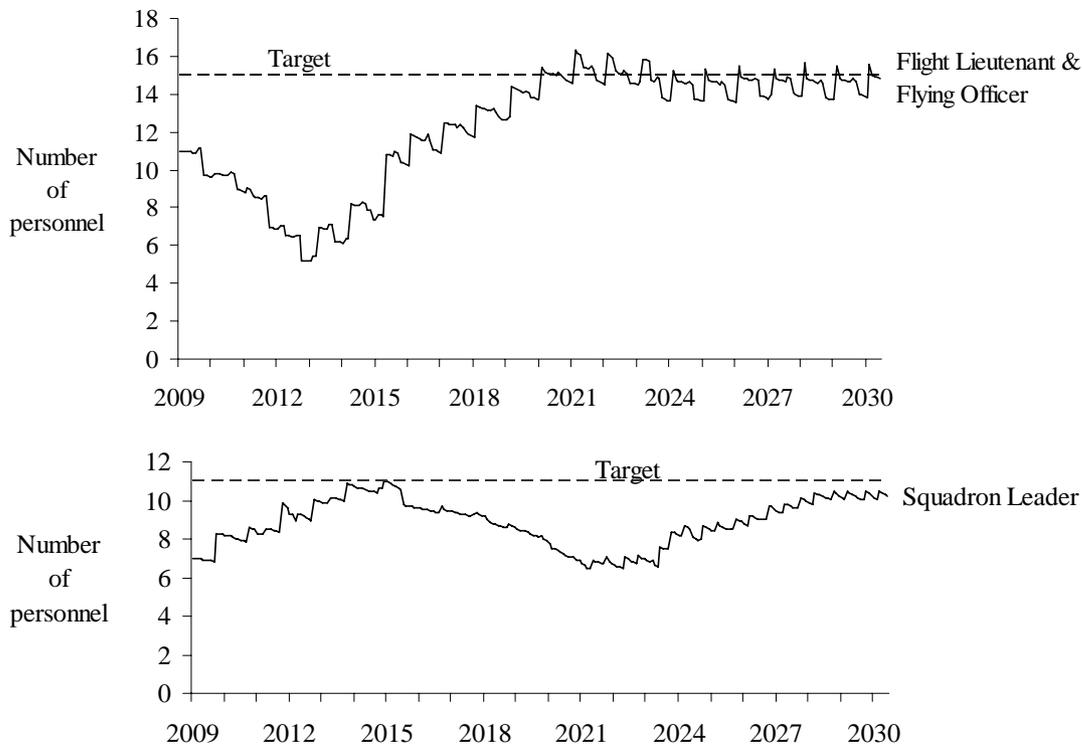


Figure 2. Structural engineering officer simulation: mean of 100 replications

These results illustrate the jaggedness of the time series in monthly steps, despite the average over 100 replications and this variability imitates the cyclical behaviour of recruitment and promotion queues in the real world. However the mean statistic do not portray some important discrete properties of these small groups and some refinement of presentation methods might need to occur.

The changes of phase in the underlying trend of the time series demonstrate the long delay times before recruitment shifts take effect at higher levels. Initially there is a nett transfer of Flight Lieutenants through to Squadron Leader who cannot be replaced by Flying Officers. After about three years this shift eases and the increased recruitment inflow begins to swell the number of Flight Lieutenants and Flying Officers. After 12 years this levels off at an equilibrium level of 15 as foreshadowed by the steady-state model, noting this occurs under a push promotion rule so it is not goal seeking. Only after about 20 years does the number of Squadron Leaders come within 10% of its goal

by pulling staff through from the rank below. The underlying trend behaviour is realistic for a military workforce.

5 Discussion

As the previous section demonstrates, the full DES model of this small and complex military workforce group produced realistic results and contained the key features of the workforce flow dynamics. The client for this project expressed confidence in the results and endorsed the recommended changes to recruitment targets. It was noteworthy in this case that the client declined to observe DES animations, preferring to see the summarised results.

The development of a steady state spreadsheet model is not specifically recommended in DES models of inventory systems, but in this case it helped to build client confidence, verify modelling data and provide a reference point for testing the DES model. Law (2007: 245) recommends creating an ‘assumptions document’ as the precursor to a simulation model. Similarly, Banks et al. (2001: 16) recommend the early step of ‘model conceptualization’ but suggest there can be no prescription for how to do this. A steady-state model can be an effective first step of analysis if it achieves the objectives of problem structuring and client communication, and provided that it is supported by adequate documentation.

As anticipated, the entity-based environment of the DES method simplified the process of converting the prototype model into the full model by adding attributes for the additional entry streams and employment streams. Although this resulted in some additional de-bugging work it was not onerous. It is further hoped that the flexibility of the DES package will allow multiple model-builders to collaborate in the development of the generic model and assist each other when applying the model to specific problems.

The prototype DES model was significantly modified to satisfy client requirements for realism and this amounted to significant new work and testing. The aim is to produce a generic DES model that is used to analyze all military trades in a similar manner. This approach implies some modest overhead of effort to update the model as the workforce system changes however this initial project suggests additional work may result from clients seeking inclusion of particular aspects of the real world. The temptation to build-in many real world features is neither practical nor sensible because ‘there is no such thing as absolute model validity’ (Law, 2007: 244). A possible outcome is that the generic model is routinely modified to satisfy client requirements but the challenge will then be to decide which features to leave out of the generic model.

6 Further work

The main objective of the next stage of DES modelling in the NZ Defence Force will be to produce a generic DES workforce model that can be applied to any trade. A number of issues in this initial stage which are likely to be the subject of work in the subsequent stages:

Training process. Add a training sector which determines promotion eligibility based on skill pathways, training schedules and resources. This will also allow personnel under training to be treated separately from strength targets.

Inputs management. Counterpart analysts have found that ‘scenario management’ software is useful for handling the different datasets associated with input variables and the different model versions linked to multiple model-builders and problems under analysis.

Outputs management. DES outputs require more advanced methods of analysis than previous models and ways should be sought to encourage clients to consider the implications of the uncertain behaviour inherent in DES models.

Testing. Further testing of the generic model is advised, particularly to exploit the wealth of historical data held by the NZ Defence Force.

Acknowledgments

The authors gratefully acknowledge the support of the Royal NZ Air Force for granting permission to publish data from the engineering officer simulation.

References

- Banks, J., Carson, J.S. II, Nelson, B.L., & Nicol, D.M. (2001) *Discrete-event system simulation* (3rd ed.), Prentice Hall: Upper Saddle River, NJ.
- Bartholomew, D.J., Hopes, R.F.A., & Smith, A.R. (1980) “Manpower planning in the face of uncertainty”, In A.R. Smith (Editor) *Corporate Manpower Planning*, Epping, UK: Gower Press.
- Bennison, M., & Casson, J. (1984). *The manpower planning handbook*. Maidenhead, UK: McGraw-Hill.
- Edwards, J.S. (1983) “A survey of manpower planning models and their application.” *Journal of the operational research society* **34(11)**: 1031-1040.
- Forrester, J.W. (1973) *World dynamics*, (2nd ed.), Pegasus Communications: Waltham, MA.
- Gass, S.I. (1991). “Military manpower planning models.” *Computers and operations research*, **18(1)**: 65-73.
- Law, A.M. (2007) *Simulation modeling and analysis* (4th ed.), McGraw-Hill: New York.

- McLucas, A.C. (2002). "Dynamic modeling to aid management of military capability." *Journal of battlefield technology* **5(1)**: 37-46.
- Mooz, W.E. (1969) *The pilot training study: personnel flow and the pilot model*. RAND.
- Purkiss, C. (1981) "Corporate manpower planning: a review of models" *European Journal of Operational Research* **8**: 315-323.
- Rockwell Automation (2005). *Arena user's guide*, Version 12.0, Sewickley, PA.
- Sterman, J.D. (2000). *Business dynamics: Systems thinking modelling for a complex world*. Irwin McGraw-Hill, Boston, MA.
- Vajda, S. (1970). "An historical survey." In A.R. Smith (Editor) *Models of manpower systems*, (7-10), London: The English Universities Press.
- Wang, J. (2005) *A review of operations research applications in workforce planning and potential modeling of military training*, Australian Defence Science and Technology Organisation report TR-1688.
- Wishart, D. (1976), "Manpower supply models IV: the MANSIM model." In *Manpower planning in the civil service*, A.R. Smith (Editor) Her Majesty's Stationery Office: London.

MIP Models for Scheduling the Operations for a Coal Loading Facility

R. Clement

School of Mathematical and Physical Sciences
University of Newcastle, Australia
Riley.Clement@studentmail.newcastle.edu.au

Abstract

Australia is the world's largest exporter of coal and is also home to Port Waratah Coal Services (PWCS), who export more coal by volume than any other export coal terminal in the world. PWCS service the Hunter Valley Coal Chain, which consists of 35 mines spread over 350km in the Hunter Valley. Managing the supply chain requires the planning and coordination of train and ship movements, subject to a number of operational constraints. This task is critical as increasing coal export is limited by the capacity of the supply chain. Two mixed integer programming (MIP) models are developed for producing a detailed rail and port schedule — a time indexed formulation, and a positional date and assignment formulation. The performance of these models is compared using historical data provided by PWCS, and Gurobi, a commercially available solver for MIP optimisation problems.

1 Introduction

Australia has an abundance of coal and is the world's largest exporter of the mineral, supplying markets in over 35 countries. The 70% of Australia's coal production which is exported accounts for 10% of Australian export, by volume, and was worth more than AU\$24 billion in the 2007–08 financial year. The industry is supported by 9 coal loading terminals located in north eastern Australia, servicing 6 major coal chains across the nation (ACA 2009). Australia is also home to Port Waratah Coal Services (PWCS), which exports more coal by volume than any other export coal terminal in the world. PWCS services the Hunter Valley Coal Chain (HVCC), which consists of 35 mines spread over 350km in the Hunter Valley (PWCS 2009). The HVCC is expected to export approximately 95 million tonnes of coal in 2009, exceeding AU\$12 billion in value (HVCCLT 2009).

The demand for Australian coal has risen significantly in recent times and will continue to do so, largely driven by the Asian markets. The abundance of Hunter Valley coal and the large number of coal producers means the ability to increase coal export is limited by the capacity of the supply chain.

The Hunter Valley Coal Chain Logistics Team (HVCCLT) was founded in 2003 with the aim of providing efficient supply chain management for the HVCC by centralising the planning and scheduling operations. The members of the logistics team

include PWCS, Newcastle Port Corporation (responsible for vessel movements), rail owners and rail operators.

We present two mixed integer programming (MIP) models for producing a detailed rail and port schedule — a time indexed formulation, and a positional date and assignment formulation. The performance of these models is compared using historical data provided by PWCS, and Gurobi, a commercially available solver for MIP optimisation problems.

2 Problem Description

The scheduling and planning task faced by the HVCCLT can be categorised into the following sub-tasks: vessel scheduling, rail scheduling, and stockyard management. The stockyard acts as a buffer between rail and vessel operations, allowing stockpiles of coal to be stored until its destination vessel is ready for loading. The management of the stockyard includes planning the position of stockpiles and the movements of large mobile machines which stack and reclaim the piles of coal. Stockyard management is not considered in this paper, and so it is assumed that the yard is of infinite capacity and coal can be moved to and from the yard without delay.

2.1 Vessel Scheduling

A ship arriving at the Hunter Port must typically wait in a queue before proceeding to a berth at a coal terminal. The HVCC includes 2 coal terminals in the port however only one is modelled in this paper. Each ship has a release date, due date and a processing time. The release date coincides with the arrival of the ship at the port and the ship cannot, of course, be loaded prior to this time. If the processing of the ship completes after the due date, which is determined by the release date alone, then demurrage (a financial penalty) is paid by the HVCC. The processing time includes the time required to load the ship and a buffer period to account for the movement of the ship in and out of the port. The loading time is dependent on the size of the cargo while the buffer period is dependent on the size of the ship as vessel movements may be restricted by the tide.

With over 1000 vessels expected to utilise the HVCC in 2009, minimising the cost of *demurrage* is the primary objective of this scheduling operation. The cargo of a vessel may range from 20,000 to over 200,000 tonnes (PWCS 2009) and may be made up of several *components* of varying size, sourced from different mines.

Although the mathematical modelling of ship scheduling problems found in container terminals has received increasing attention (Steenken, Voß, and Stahlbock 2004), there are no obvious investigations using similar approaches for coal terminals (or bulk handling terminals in general).

2.2 Rail Scheduling

A train is loaded with coal from a mine at a *loadpoint*, where it incurs a loading time which is dependent on the loadpoint and the volume of coal to be loaded onto the train. There is a set of standard sizes for the fleet of trains, and each loadpoint has a preferred train size. Once loaded a train then proceeds to the *dump station* at the coal terminal where it unloads its cargo.

A train *path* describes the movement of a train spatially and temporally. Many of the rail scheduling problems presented in literature involve creating paths, i.e. determining the routing and frequency of trains subject to rail network constraints. Such investigations are often solved with heuristic approaches e.g. (Burdett and Kozan 2008), (Pudney and Wardrop 2008), or with discrete simulation (Dorfman and Medanic 2004), as a large number of logic variables makes the problem computationally intractable for exact methods (Dessouky et al. 2006).

The rail scheduling task faced by the HVCCLT is not as flexible as these problems since a large section of track, referred to as the *main corridor* is shared with domestic rail operations. The trains are restricted to moving through the main corridor along *base paths* whose timing is pre-determined and fixed by the rail operators. Outside of this main corridor, the movement of a train is not as constrained and it may wait or travel slower than it otherwise could. Paths are divided into up-paths, which travel from a loadpoint to coal terminal, and down-paths which travel in the opposite direction. Since a train is either heading to, or returning from, a loadpoint, a set of *extended paths* is generated by pairing base paths with those loadpoints available to it. A base path using the full extent of the main corridor is shown in figure 1a), while figure 1b) shows a possible extended path which could be generated from the base path — timing data is omitted.

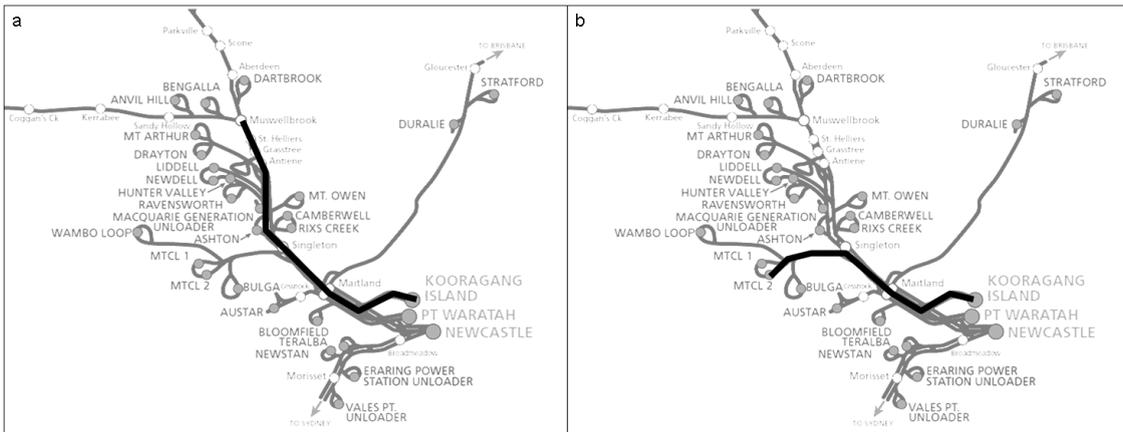


Figure 1: a) a base path (main corridor), b) an extended path

In order to simplify the problem it is assumed the trains travel to and from the main corridor at their maximum speed, allowing the extended paths to have a static structure. It is assumed that the loading time at the loadpoint is dependent on the train size (rather than the size of its cargo) and that a loadpoint will only receive trains of its preferred size. Under these assumptions a train size and load (unload) time can be associated with each extended down-path (up-path), since a loadpoint is unique to each extended path.

The loadpoints restrict possible schedules by having a maximum daily throughput which is imposed as a maximum number of trains per day using the assumptions above. Additionally, a loadpoint can only accommodate one train at a time. Similarly there is a maximum number of trains that can unload at the dump station at any one time. *Dwell*, which refers to the time a train spends idle at a loadpoint, is limited to 3 hours which restricts the possible pairings of down-paths and up-paths. It is also preferable to have a schedule with a small amount of total dwell, and this is considered as a secondary objective.

Furthermore a number of operational constraints are imposed on the rail schedules:

- Precedence constraints: a train must arrive prior to the commencement of vessel loading.
- Stockpile window constraints: a train must not arrive prior to a particular date, which is dependent on the time at which the vessel commences loading, and the mines from which the ship cargo is sourced.
- Headway constraints: *headway* is the term used to describe the time between consecutive trains. There are several sections of track in which a lower bound on the headway is imposed.
- A section of *capacitated track* has a restriction on the number of trains allowed to pass through it in any 24 hr period.

A visualisation of a schedule for a single ship is shown in figure 2.

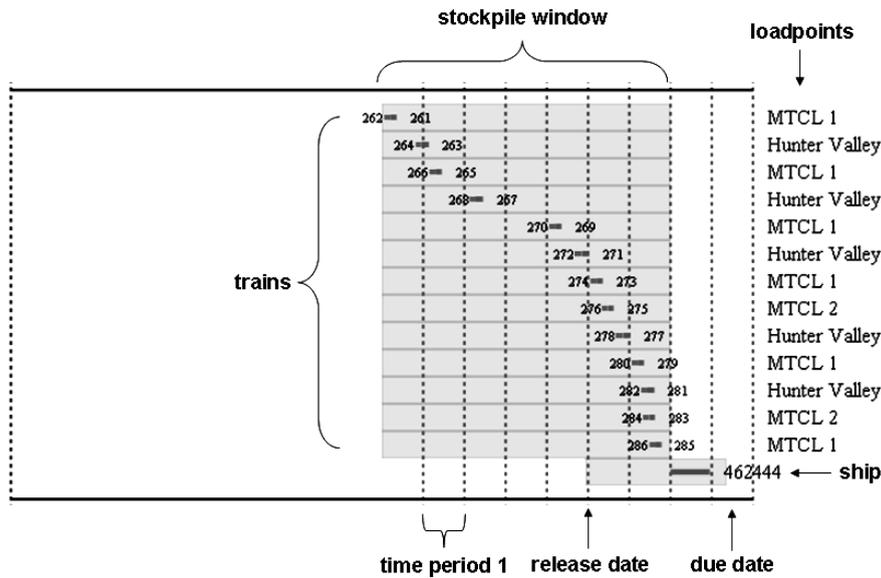


Figure 2: Example schedule for a ship

3 Problem Formulation

We present two mixed integer programming models for producing a detailed rail and port schedule: one based on a time indexed formulation, and the other based on a positional date and assignment formulation. A discussion of these types of models can be found in (Queyranne and Schulz 1994). Our two models predominantly differ in the vessel scheduling, and share a large number of variables and constraints used to schedule the rail operations.

Indices

- $j \in J$ = a ship (job) to be loaded
- $l \in L$ = a loadpoint; The set L_j contains those loadpoints which will supply ship j .
- $b \in B_j \subseteq B$ = a cargo component for ship j

- $s \in S =$ a train size
- $f \in F =$ a berth
- $m \in M$ ($m' \in M'$) = a base up-path (down-path)
- $i \in P$ ($i' \in P'$) = an extended up-path (down-path); $P_{jl} \subseteq P =$ extended up-paths originating at loadpoint $l \in L_j$; $P_j = \bigcup_{l \in L_j} P_{jl}$; P_l (P'_l) = extended up-paths (down-paths) associated with loadpoint l ; P_m ($P_{m'}$) is the set of extended up-paths (down-paths) formed from the corridor path m (m'); $P'_i =$ extended down-paths which may be paired with extended up-path i ; P_s ($P_{s'}$) = extended up-paths (down-paths) utilising a train of size s
- $\bar{P} \subseteq 2^P$ ($\bar{P}' \subseteq 2^{P'}$) = a collection of sets of extended up-paths (down-paths) which are pairwise incompatible due to headway requirements imposed on the rail network
- $\tilde{P}'_l \subseteq 2^{P'_l}$ = a collection of sets of extended down-paths to loadpoint l which all conclude on the same day
- $\check{P} \subseteq 2^P =$ a collection of sets of extended up-paths which overlap during dump station unloading
- $\hat{P} \subseteq 2^P =$ a collection of sets of extended up-paths utilising capacitated sections which all begin within 24hrs of each other

Parameters

- $l_b =$ the loadpoint associated with component b
- $n_b =$ the number of trains required to satisfy component $b \in B$
- $e_j =$ the length of the stockpile window for ship j
- $a_s =$ the number of trains of size s available in the fleet
- $s_i =$ the train size of extended path i
- $q_l =$ the number of trains which can be served by loadpoint l in a day
- $t_{i'i} =$ the idle time spent at a loadpoint by a train using extended down-path i' and extended down-path i
- $G =$ the number of trains allowed through the section of capacitated track in a 24hr period
- $D =$ the number of trains which may simultaneously unload in dump station
- $start(i), end(i) =$ time at which extended path i begins and concludes

3.1 Common Rail Element

The variables and constraints are common to both the time indexed model and positional assignment and date model.

Variables

- $y_{ij} \in \mathbf{B}$ is defined for all $j \in J, i \in P_j$. $y_{ij} = 1$ iff up-path i is used by ship j
- $w_{i'} \in \mathbf{B}$ is defined for all $i' \in P'$. $w_{i'} = 1$ iff down-path i' is used
- $z_{i'i} \in \mathbf{B}$ is defined for all $i \in P, i' \in P'_i$. $z_{i'i} = 1$ iff down-path i' is linked with up-path i

Dwell objective

$$\min \sum_{i' \in P'} \sum_{i \in P} t_{i'i} z_{i'i}$$

Constraints

$$\sum_{j \in J: i \in P_j} y_{ij} \leq 1 \quad \forall i \in P \quad (1)$$

$$\sum_{j \in J} \sum_{i \in P_j \cap P_m} y_{ij} \leq 1 \quad \forall m \in M, \quad \sum_{i' \in P'_{m'}} w_{i'} \leq 1 \quad \forall m' \in M' \quad (2)$$

$$\sum_{i \in P_{jl}} y_{ij} = \sum_{b \in B_j: l_b = l} n_b \quad \forall j \in J, \quad \forall l \in L_j \quad (3)$$

$$\sum_{i' \in P'_{s_{p'}}: \text{start}(i') \leq \text{start}(p')} \sum_{i \in P_{s_{p'}}: \text{end}(i) > \text{start}(p')} z_{i'i} \leq a_{s_{p'}} \quad \forall p' \in P' \quad (4)$$

$$\sum_{i' \in P'_l: \text{end}(i') \leq \text{end}(p')} \sum_{i \in P_l: \text{start}(i) > \text{end}(p')} z_{i'i} \leq 1 \quad \forall l \in L, \quad \forall p' \in P'_l \quad (5)$$

$$\sum_{i' \in P^*} w_{i'} \leq q_l \quad \forall l \in L, \quad \forall P^* \in \tilde{P}'_l \quad (6)$$

$$\sum_{j \in J} \sum_{i \in P_j \cap P^*} y_{i'j} \leq 1 \quad \forall P^* \in \bar{P}, \quad \sum_{i' \in P^*} w_{i'} \leq 1 \quad \forall P^* \in \bar{P} \quad (7)$$

$$\sum_{j \in J} \sum_{i \in P_j \cap P^*} y_{ij} \leq G \quad \forall P^* \in \hat{P} \quad (8)$$

$$\sum_{j \in J} \sum_{i \in P_j \cap \check{P}} y_{ij} \leq D \quad \forall P^* \in \check{P} \quad (9)$$

$$\sum_{j \in J: i \in P_j} y_{ij} = \sum_{i' \in P'_i} z_{i'i} \quad \forall i \in P, \quad w_{i'} = \sum_{i \in P_{i'}} z_{i'i} \quad \forall i' \in P' \quad (10)$$

The *path capacity constraints* (1) are required since an extended up-path cannot be used by more than one ship. Also extended paths formed from the same base path cannot be used together, giving rise to *base path constraints* (2). The *path requirement constraints* (3) are responsible for ensuring the number of paths assigned to a ship equals the number of trains required to complete the ship's cargo. The *fleet size constraints* (4) are needed to make sure that the number of paths (sharing a common train size) being utilised at any point in time does not exceed the number of trains (of that size) available. The *loadpoint loop constraints* (5) restrict multiple trains from concurrently occupying a loadpoint. The *loadpoint capacity constraints* (6) restrict the number of trains per day through each of the loadpoints. *Headway constraints* (7) ensure consecutive trains are a safe time apart. The *capacitated track constraints* (8) prevent the number of trains utilising the section of *capacitated track* from exceeding the maximum imposed in any 24 hour period. The *dump station constraints* (9) are responsible for limiting the number of trains which can unload simultaneously. Finally *linking constraints* (10) enforce the correct relationship between y, w, z variables.

3.2 Time Indexed Formulation

In a time indexed model the job (ship) scheduling variables are indexed by (j, t) pairs where j is a job and t is a time period. In order to have a finite set of these variables

it is necessary to have a finite set of time periods, and so a planning horizon must be introduced. The problem is therefore modelled over a set of contiguous time periods $\{1, 2, \dots, T\}$, each of unit size.

Indices

$t \in \{1, 2, \dots, T\}$ = a time period. Let $[a, b] = \{a, \dots, b\} \cap \{1, \dots, T\}$.

Parameters

- $r_j, d_j \in [1, T]$ = release date and due date of ship j , respectively
- $p_j \in \mathbf{Z}_+$ = number of time periods required for loading ship j
- $c_{jt} \in \mathbf{R}_+$ = the demurrage incurred if ship j begins loading at the start of t

Variables

- $x_{jt} \in \mathbf{B}$ is defined for all $j \in J, t \in [r_j, T - p_j + 1]$ with $x_{jt} = 1$ iff ship $j \in J$ commences loading at time t

Demurrage Objective

$$\min \sum_{j \in J} \sum_{t \in [r_j, T]} c_{jt} x_{jt}$$

Constraints

$$\sum_{t \in [r_j, T - p_j + 1]} x_{jt} = 1 \quad \forall j \in J \quad (11)$$

$$\sum_{j \in J} \sum_{t' \in [t - p_j + 1, t]} x_{jt'} \leq |F| \quad \forall t \in [1, T] \quad (12)$$

$$\sum_{\substack{t \in \{1, 2, \dots, T\}: \\ \text{start}(t) < \text{end}(i)}} x_{jt} + \sum_{\substack{t \in \{1, 2, \dots, T\}: \\ \text{start}(t) > \text{end}(i) + e_j}} x_{jt} + y_{ij} \leq 1 \quad \forall j \in J, \quad \forall i \in P_j \quad (13)$$

The *job completion constraints* (11) ensure each ship is scheduled to be loaded exactly once. *berth capacity constraints* (12) ensure there are no more ships loading during any time period, than there are berths. Finally, the *path feasibility constraints* (13) are required to make sure that trains arrive prior to commencement of ship loading, and respect the (stockpile) time window in which they can arrive prior to this time.

3.3 Positional Date and Assignment Formulation

Indices

- $k \in \{1, 2, \dots, K\}$ = a position in a job sequence

Parameters

- $r_j \in \mathbf{Z}_+$ = the earliest time (in hours) loading of shipment j can commence
- $d_j \in \mathbf{Z}_+$ = the latest time (in hours) loading of shipment j can complete without incurring a penalty
- $p_j \in \mathbf{Z}_+$ = the number of hours required for ship loading
- c is the cost of demurrage per hour

Variables

- $\tau_j \in \mathbf{R}_+$ = the tardiness of ship j (in hours)
- $v_{jfk} \in \mathbf{B} = 1$ iff j is processed in the k^{th} position on berth f
- $t_{jfk} \in \mathbf{R}_+$ = the time (in hours) at ship j commences loading if $v_{jfk} = 1$, and 0 otherwise

Demurrage Objective

$$\min c \sum_{j \in J} \tau_j$$

Constraints

$$\tau_j \geq \sum_{f \in F} \sum_{k \in \{1, \dots, K\}} t_{jfk} + p_j - d_j \quad \forall j \in J \quad (14)$$

$$\sum_{f \in F} \sum_{k \in \{1, \dots, K\}} v_{jfk} = 1 \quad \forall j \in J \quad (15)$$

$$\text{end}(i)y_{ij} \leq \sum_{f \in F} \sum_{k \in \{1, \dots, K\}} t_{jfk} \quad \forall j \in J, \quad \forall i \in P_j \quad (16)$$

$$\sum_{f \in F} \sum_{k \in \{1, \dots, K\}} t_{jfk} \leq T(1 - y_{ij}) + (\text{end}(i) + e_j)y_{ij} \quad \forall j \in J, \quad \forall i \in P_j \quad (17)$$

$$\sum_{j \in J} v_{jf1} \leq 1 \quad \forall f \in F, \quad \sum_{j \in J} (v_{jfk} - v_{jf(k-1)}) \leq 0 \quad \forall f \in F, k \in [2, \dots, K] \quad (18)$$

$$r_j v_{jfk} - t_{jfk} \leq 0, t_{jfk} - T v_{jfk} \leq 0 \quad \forall j \in J, f \in F, k \in [1, \dots, K] \quad (19)$$

$$\sum_{j \in J} t_{jfk} - \sum_{j \in J} t_{jf(k-1)} - \sum_{j \in J} p_j v_{jf(k-1)} + T(1 - \sum_{j \in J} v_{jfk}) \geq 0 \quad \forall f \in F, k \in [2, \dots, K] \quad (20)$$

The *tardiness constraints* (14) are required for the τ variables to accurately reflect the tardiness in any solution. The *job completion constraints* (15) restrict each ship to be loaded exactly once. The *train-ship precedence constraints* (16) ensure that all trains for a particular ship have finished unloading prior to the ship loading while the *stockpile window constraints* (2) ensure that the trains observe the time window in which they can arrive prior to ship loading. The *job ordering constraints* (21,22) ensure there are no gaps in the berth ordering, i.e. there is no k^{th} job for a particular berth if a $k - 1^{th}$ job does not exist (unless it is the first job). The *job timing constraints* (23,24) have two roles: the first is to enforce ship release dates, and the second is to guarantee that the processing of consecutive ships using the same berth does not overlap.

3.4 Comparison of Models

Also previously noted there is a need for a finite time horizon for both models. In the time index formulation this was necessary in order to have a finite number of variables while the positional date and assignment formulation requires an upper bound to use as a “big M” value in several families of constraints. The choice of this value may not only determine the feasibility of the problem, but it is possible

that increasing the value may result in a better solution. In the positional date and assignment model a value for K , the maximum size of a job sequence for any berth, must also be chosen however it is not clear what a suitable value for K might be. The advantage of the positional date and assignment model is its accuracy in modelling the problem, allowing ships to commence loading at any time (subject to release dates), whereas a ship must commence loading at the start of a time period in the time indexed model. The processing times, release dates and due dates used in the time indexed model must also be phrased in terms of time periods, which also decreases the accuracy of this model. As the length of the time periods decreases, the accuracy of the time indexed model increases, however so does the number of variables and constraints in the model. The disadvantage of the positional date and assignment model is that several relationships needed to be modelled as “big M” constraints, which typically yield weak linear relaxations, whereas the time indexed model contains clique and cardinality constraints only.

4 Results

The above formulations were coded in Python for use with Gurobi, a commercially available solver for MIP optimisation problems, and tested with historical data provided by PWCS. The results have shown that the time indexed (TI) model, using a time period length of one day, generally solves instances to optimality in half the time required for the positional date and assignment (PDA) model to do the same. Three different objectives were trialled with the TI model: min demurrage; min dwell; min demurrage + ϵ *dwell. The results showed that using the third objective typically achieved values for demurrage and dwell which were very close to the optimal values found for solving for demurrage or dwell alone. An interesting observation is the high number of instances which were integral at the root node (after presolve) when solving for demurrage with the TI model. This was not the case when solving for dwell with the TI model, despite having similar solve times. Furthermore it was noted that introducing dwell, even with a very small weight, into the demurrage objective significantly reduced the number of instances which were integral at the root node. The positional date and assignment(PDA) model, however, had few instances solve in this way (regardless of the objective used). The PDA model also shows a tendency to explore many nodes of the branch and bound tree where the TI model does not.

Table 1:

Instance			TI Model			PDA Model		
Name	Days	Jobs	Rows	Cols	Time (s)	Rows	Cols	Time (s)
d1d4	4	7	174533	153362	197.4	206744	158320	440.2
d1d6	6	11	220328	192921	289.1	261830	197569	1924.9
d1d8	8	18	258840	225605	353.8	309737	229831	18193.6
d2d4	4	2	63613	52091	45.0	69927	52000	54.6
d2d6	6	8	134258	115259	122.2	154079	118155	263.1
d3d4	4	5	125573	109644	119.3	142955	111647	273.2
d3d6	6	9	190974	167122	281.1	218820	170155	846.8

5 Further Work

Each of the models presented above have their advantages. The positional date and assignment model achieves higher accuracy in modelling the problem but does not solve as quickly as the time indexed model. Attempts to combine the advantages of both models may include:

1. Using the solution from the time indexed model to fix a set of variables within the positional date and assignment model
2. Using the time indexed model to provide a starting solution for the positional date and assignment model
3. Developing a hybrid of the two models

Acknowledgements

This work presented in this paper is part of a collaboration involving the University of Newcastle (UoN) and the Hunter Valley Coal Chain Logistics Team (HVCCLT). Thank you to the following people for their guidance and contributions to this paper: Natashia Boland (UoN), Bhaswar Choudhury (HVCCLT), Tracey Giles (HVCCLT), Rob Oyston (HVCCLT), Martin Savelsbergh (UoN, Georgia Institute of Technology), Hamish Waterer (UoN), Palitha Welgama (HVCCLT).

References

- ACA. 2009, November. Australian Coal Association. <http://www.australiancoal.com.au>.
- Burdett, R.L., and E. Kozan. 2008. "A sequencing approach for creating new train timetables." *OR Spectrum*, pp. 1–31.
- Dessouky, M.M., Q. Lu, J. Zhao, and R.C. Leachman. 2006. "An Exact Solution Procedure for Determining the Optimal Dispatching Times for Complex Rail Networks." *IIE Trans* 38:141–152.
- Dorfman, MJ, and J. Medanic. 2004. "Scheduling trains on a railway network using a discrete event model of railway traffic." *Transportation Research Part B* 38 (1): 81–98.
- HVCCLT. 2009, November. Hunter Valley Coal Chain Logistics Team. <http://www.hvccclt.com.au>.
- Pudney, P., and A. Wardrop. 2008. "Generating Train Plans with Problem Space Search." *Lecture notes in economics and mathematical systems* 600:195.
- PWCS. 2009, November. PWCS - Port Waratah Coal Services. <http://www.pwcs.com.au>.
- Queyranne, M., and A.S. Schulz. 1994. *Polyhedral approaches to machine scheduling*. TU, Fachbereich 3.
- Steenken, D., S. Voß, and R. Stahlbock. 2004. "Container terminal operation and operations research-a classification and literature review." *OR spectrum* 26 (1): 3–49.

Optimal Pricing Decision for a Dynamic Inventory Problem with Constant Price Elasticity of Demand

Chia-Shin Chung and James Flynn

Cleveland State University

Cleveland, Ohio 44114 USA

ABSTRACT

Monahan, Petruzzi, and Zhao (2004) studied a dynamic inventory problem with pricing decision for a season consisting of T periods. At the beginning of the season, a decision is made about how much inventory to carry for the entire season. No replenishment is possible after the season starts. The demand in each period is random and price dependent. Their model utilizes a special class of demand functions, which is defined to be a product of a nonnegative continuous distribution and a deterministic function of the price. The price elasticity of the demand is constant. They show that the problem can be reduced to a sequence of state dependent, static, single variable optimization problem. This makes it possible for developing an efficient algorithm. However their model does not include holding costs, which can be significant, especially for a long season. In this paper, we study the effect of holding cost on the result of the model. We found that when the holding cost is proportional to the selling price, the same simple structure can be observed in the optimal solution. However it is no longer true for the more general case. For the general case, we characterize the properties of the optimal policies, present algorithms for their computation, and develop a heuristic.

1. INTRODUCTION

In this paper, we study an inventory model with pricing decisions in a season which consists of T periods. All orders arrive before the selling season begins. For $1 \leq t \leq T$, the demand D_t for period t equals $A_t p_t^{-b}$ where b is the absolute value of the price elasticity of demand. The price p_t is allowed to have a complex structural dependence on the inventory level, I_t , at the start of period t . For a model without holding costs, an article by Monahan, Petruzzi, and Dada (2004) finds that an optimal pricing policy satisfies the condition: $p_t^b I_t = z_t$, $1 \leq t \leq T$, where z_t is a nonnegative stocking factor for period t , which is found to be independent of I_t .

Monahan, Petruzzi, and Dada (2004) provide a detailed explanation of their model. To justify the model, the paper cites three advantages - tractability, robustness and its practical and intuitive appeal. In particular, they find that with an appropriate transformation of variables, the dynamic pricing problem can be reduced to a sequence of state-independent, static, single-variable optimization problems. As a result, determination of the optimal pricing policy avoids the curse of dimensionality associated with solving a dynamic program. The model is robust since the result can be applied using any demand distribution. MPZ also provides reasons for its use of the isoelastic form of demand. One of them is that it typically provides a good statistical fit with available sales data. Our model is in the literature on dynamic inventory models with

pricing and stochastic demand. Petruzzi and Dada (1999) provides a recent review of this literature.

We extend their model to include holding costs. We found that when the holding cost is proportional to the selling price, the optimal pricing structure is similar to the one in Monahan, Petruzzi, and Dada (2004). However it is no longer true for the more general case. We characterize the properties of the optimal policies, present algorithms for their computation, and develop heuristics.

2 BASIC RESULTS

In the usual dynamic programming formulation of inventory models with pricing decisions, the price p_t at stage t , $1 \leq t \leq T$, is allowed to have a complex structural dependence on the system state, I_t , the inventory level at the start of period t . For a model without holding costs, MPZ (Monahan, Petruzzi, and Dada 2004) found that an optimal pricing policy satisfies the condition:

$$p_t^b I_t = z_t, 1 \leq t \leq T,$$

where z_t is a nonnegative constant, which does not depend on I_t , and b is the absolute value of the price elasticity of demand. In this article, such pricing policies are said to have *simple structure*. Being able to restrict attention to pricing policies with simple structure reduces the dimensionality of the optimization problem that must be solved to find optimal prices. Thus, MPZ yields significant computational savings. Its main drawback is that it ignores holding costs.

For a model with holding costs, this article investigates pricing policies that are optimal among all policies with simple structure, *referred to as OSS policies*. For the special case where the holding costs are proportional to the selling prices, there exist OSS policies that are optimal among all policies and can be computed using techniques of MPZ. For the general case, where OSS policies need not be optimal among all policies, we characterize their properties, indicate how to compute them, and develop heuristics. A numerical study in §6 evaluates these heuristics and compares OSS policies with optimal policies.

The properties of OSS policies depend on the demand information available when decisions are made. We consider two cases: the *standard case*, where the only information about current and future demands is their probability distributions and the *deterministic case*, where one knows the value of all demands. Unless stated otherwise, our results apply to all two cases.

Our model extends MPZ. Expected profit maximization is the objective. All orders arrive before the selling season begins. The season consists of T periods, indexed so period t represents the number of periods remaining in the season, e.g., T is the first period and 1 is the last period. The notation I represents the initial stock I_T . All revenue is from sales. A holding cost $h_t I_{t-1}$ is charged to period t , where $h_t \geq 0$. The only other cost is the cost cI of obtaining initial stock, where $c > 0$. The decision variables are I and the selling prices, p_1, \dots, p_T .

For $1 \leq t \leq T$, the demand D_t for period t satisfies

$$D_t = \begin{cases} \infty, & \text{if } p_t = 0 \\ A_t p_t^{-b}, & \text{if } p_t > 0 \end{cases} \quad (2.1)$$

where $b > 1$ and $A_1, A_2 \dots$ are independent, *positive-valued*, random variables with means, $\mu_1, \mu_2 \dots$, and cumulative distribution functions, $F_1, F_2 \dots$, respectively. Note that the demand

equals a *random* quantity multiplied by a *deterministic* function of the price. Also b denotes the absolute value of the price elasticity of demand. Our assumption that the A_t s are positive-valued ensures that a positive demand is always possible for some range of price values. Given that our objective is expected profit maximization, the results obtained here apply when the A_t s have a nonnegative continuous distribution such as the gamma, since continuous random variables equal 0 with probability 0. Note that if $p_t = 0$, then both the revenue and ending inventory for period t equal 0.

Following MPZ, we express p_t , $1 \leq t \leq T$, in terms of the *period- t stocking factor* $z_t \geq 0$, i.e.,

$$p_t = \begin{cases} (z_t/I_t)^{1/b}, & \text{if } I_t > 0, \\ \text{undefined}, & \text{if } I_t = 0. \end{cases} \quad (2.2)$$

Having p_t undefined if $I_t = 0$ is acceptable since p_t is irrelevant when $I_t = 0$. under (2.2),

$$z_t = I_t p_t^b, \text{ if } I_t > 0. \quad (2.3)$$

The decision variables are the initial inventory $I \geq 0$ and the nonnegative *stocking factor vector* $\mathbf{z} = (z_1, \dots, z_T)$. (As usual \mathbf{z} is nonnegative if its components are nonnegative.) Given such I and \mathbf{z} , define

$R(I, \mathbf{z})$ = the expected value of the total revenue,

$V(I, \mathbf{z}) = R(I, \mathbf{z}) - \sum_{t=1}^T E(h_t I_{t-1})$, the expected total revenue minus holding costs,

$\pi(I, \mathbf{z}) = V(I, \mathbf{z}) - cI$, the expected total profit.

An OSS policy consists of decision variables I and \mathbf{z} that maximize $\pi(I, \mathbf{z})$.

We employ the following additional notation, where $1 \leq t \leq T$.

I_0 = the stock at the end period 1,

$$m = 1 - 1/b, \text{ where } 0 < m < 1, \quad (2.4)$$

$$S_t = I_t - I_{t-1}, \text{ sales in period } t, \quad (2.5)$$

$$B_t(z_t) = \begin{cases} 0, & \text{if } z_t \leq A_t \\ 1 - A_t/z_t, & \text{if } z_t > A_t \end{cases} \quad (2.6)$$

Note that $0 \leq B_t(z_t) \leq 1$. Also, $B_t(z_t) = [1 - A_t/z_t]^+$ for $z_t > 0$.

LEMMA 2.1. Suppose \mathbf{z} is nonnegative and $1 \leq t \leq T$.

$$I_{t-1} = B_t(z_t)I_t, \quad (2.7)$$

$$S_t = I_t(1 - B_t(z_t)), \quad (2.8)$$

$$p_t S_t = z_t^{1-m} I_t^m (1 - B_t(z_t)) \text{ if } I_t > 0. \quad (2.9)$$

PROOF. Clearly (2.7) holds if $I_t = 0$, since $I_t = 0$ implies $I_{t-1} = 0$. Let $I_t > 0$ and $p_t > 0$. Then $z_t > 0$ by (2.3). Also, $I_{t-1} = [I_t - D_t]^+ = I_t [1 - A_t p_t^{-b}/I_t]^+ = I_t [1 - A_t/z_t]^+$, proving (2.7).

Alternatively, let $I_t > 0$ and $p_t = 0$. Then $I_{t-1} = 0$ by (2.1). Also, $z_t = 0$ by (2.3), so $B_t(z_t) = 0$.

Hence, (2.7) holds. Next, $S_t = I_t - I_{t-1}$, which combined with (2.7) gives us (2.8). Finally, $m = 1 - 1/b$, (2.2), and (2.7) imply (2.9). \square

Note that (2.8) is motivated by (4) of MPZ. Extending Proposition 1 of MPZ, the lemma

below, obtains a simple recursive formulas for $V(I, \mathbf{z})$. Before stating that lemma, we require some definitions.

For nonnegative \mathbf{z} , let $r(\mathbf{z}) = r_T$ and $q(\mathbf{z}) = q_T$, where r_t and q_t are defined by

$$r_0 = 0 \text{ and } r_t = E\left[z_t^{1-m} (1 - B_t(z_t)) + r_{t-1} B_t(z_t)^m\right], \text{ for } 1 \leq t \leq T, \quad (2.10)$$

$$q_0 = 0 \text{ and } q_t = E\left[(h_t + q_{t-1}) B_t(z_t)\right], \text{ for } 1 \leq t \leq T. \quad (2.11)$$

LEMMA 2.2. Let I and \mathbf{z} be nonnegative. Then

$$V(I, \mathbf{z}) = I^m r(\mathbf{z}) - I q(\mathbf{z}), \quad (2.12)$$

$$\sum_{t=1}^T E(h_t I_{t-1}) = I q(\mathbf{z}), \quad (2.13)$$

$$R(I, \mathbf{z}) = I^m r(\mathbf{z}). \quad (2.14)$$

PROOF. One need only prove (2.13) - (2.14), since these results imply (2.12). We use induction for (2.13). If $T = 1$, $E(h_1 I_0) = I_1 E[h_1 B_1(z_1)]$ by (2.7), so by $q_0 = 0$ and (2.11), $E(h_1 I_0) = I_1 q_1$, proving (2.13). Next, let (2.13) hold for $T - 1 \geq 1$. Then, $\sum_{t=1}^T E(h_t I_{t-1}) = E(h_T I_{T-1}) + E(q_{T-1} I_{T-1})$, which by (2.7) equals $I_T E((h_T + q_{T-1}) B_T(z_T))$. Thus, by (2.11), $\sum_{t=1}^T E(h_t I_{t-1}) = I_T q_T$, finishing (2.13).

Turn to (2.14). If $I = 0$, $R(I, \mathbf{z}) = 0$, so (2.14) holds trivially. Thus, assume $I > 0$. We use induction for (2.14). Let $T = 1$. Then, $R(I, \mathbf{z}) = E(p_1 S_1)$, which by (2.9) equals

$I_1^m E(z_1^{1-m} (1 - B_1(z_1)))$, which by $r_0 = 0$ and (2.10) equals $I_1^m r_1$, proving (2.14). Next, let (2.14)

hold for $T - 1 \geq 1$. Then $R(I, \mathbf{z}) = E(p_T S_T) + E(r_{T-1} I_{T-1}^m)$, which by (2.9) and (2.7) equals

$I_T^m E(z_T^{1-m} (1 - B_T(z_T))) + I_T^m E(r_{T-1} B_T(z_T)^m)$, which by (2.10) equals $I_T^m r_T$, as required. \square

In view of Lemma 2.2, one can obtain an OSS policy by finding nonnegative I^* and \mathbf{z}^* that maximize

$$\pi(I, \mathbf{z}) = V(I, \mathbf{z}) - cI. \quad (2.15)$$

After characterizing the properties of I that are best for fixed \mathbf{z} , the theorem below provides necessary and sufficient conditions for I^* and \mathbf{z}^* to be an OSS. For any nonnegative \mathbf{z} , define

$$I(\mathbf{z}) = \left(\frac{mr(\mathbf{z})}{c + q(\mathbf{z})} \right)^b, \quad (2.16)$$

$$\pi_0(\mathbf{z}) = (1 - m) m^{b-1} (c + q(\mathbf{z}))^{1-b} r(\mathbf{z})^b. \quad (2.17)$$

THEOREM 2.1.

(a) Given any nonnegative \mathbf{z} , I maximizes $\pi(I, \mathbf{z})$ if and only if I equals $I(\mathbf{z})$; furthermore, $\pi(I(\mathbf{z}), \mathbf{z}) = \pi_0(\mathbf{z})$.

(b) Nonnegative I^* and \mathbf{z}^* constitute an OSS if and only if \mathbf{z}^* maximizes $\pi_0(\mathbf{z})$ and $I^* = I(\mathbf{z}^*)$.

PROOF. For (a), let \mathbf{z} be fixed. Clearly, $\frac{\partial}{\partial I} \pi(I, \mathbf{z}) = mI^{m-1} r(\mathbf{z}) - q(\mathbf{z}) - c$. Then,

$\frac{\partial}{\partial I} \pi(I, \mathbf{z}) = 0$ has a unique solution $I = I(\mathbf{z})$. Also, $0 < m < 1$ implies $\pi(I, \mathbf{z})$ is concave in I .

These results imply $I = I(\mathbf{z})$ is the unique value that maximizes $\pi(I, \mathbf{z})$. One can show that $\pi(I(\mathbf{z}), \mathbf{z}) = \pi_0(\mathbf{z})$ by substituting $I(\mathbf{z})$ for I in (2.15), giving us (a). Part (b) follows immediately from (a). \square

Obtaining the best I for a given \mathbf{z} using (2.16) is straightforward, while finding an OSS for the special cases covered in §3 and §4 requires the same effort as getting a solution to MPZ. In general, however, finding an OSS entails finding a T -dimensional vector \mathbf{z}^* that globally maximizes π_0 . Since π_0 need not be concave, there are the usual concerns about global versus local optima. Still one can employ standard optimization algorithms in the attempt to find \mathbf{z}^* . §5 develops a heuristic that exploits results for the model of §4.

3 A MODEL WHERE OSS POLICIES ARE OPTIMAL

This section deals with a model where the holding costs are proportional to the selling prices. (A more common assumption is that holding costs are proportional to the purchase cost.) We establish that OSS policies are optimal among all policies and show that for the standard case, computing OSS policies requires the same work as in MPZ.

Formally, assume throughout this section that

$$h_t = \eta_t p_t, \text{ for } 1 \leq t \leq T, \quad (3.1)$$

where η_t is a nonnegative constant.

The next lemma extends Lemma 2.2. For nonnegative \mathbf{z} , let $v(\mathbf{z}) = v_T$, with v_T being defined by

$$v_0 = 0 \text{ and } v_t = E \left[z_t^{1-m} (1 - (1 + \eta_t) B_t(z_t)) + B_t(z_t)^m v_{t-1} \right], \text{ for } 1 \leq t \leq T. \quad (3.2)$$

LEMMA 3.1. Let $1 \leq t \leq T$ and let I and \mathbf{z} be nonnegative. Then

$$V(I, \mathbf{z}) = I^m v(\mathbf{z}), \quad (3.3)$$

PROOF. Note that (3.1), (2.2), and $m = 1 - 1/b$ imply

$$h_t = \eta_t z_t^{1-m} I_t^{-1+m} \text{ if } I_t > 0. \quad (3.4)$$

Note also that (3.3) holds trivially if $I = 0$. Thus, assume $I > 0$. Also, let r_t and q_t , $1 \leq t \leq T$, satisfy the conditions of Lemma 2.2. We use induction. Suppose $T = 1$. Then $V(I, \mathbf{z}) = I^m r_1 - I q_1$, by Lemma 2.2. Now (2.11), (3.4), $I > 0$, and $q_0 = 0$ imply $q_1 = I^{-1+m} E(\eta_1 z_1^{1-m} B_1(z_1))$. Also, (2.10) and $r_0 = 0$ imply $r_1 = E[z_1^{1-m} (1 - B_1(z_1))]$. One can easily verify that $I^m r_1 - I q_1 = I^m v_1$, proving (3.3) for $T = 1$. Next, assume (3.3) holds for $T - 1 \geq 1$. Then, $V(I, \mathbf{z}) = E(p_T S_T) - E(h_T I_{T-1}) + E(v_{T-1} I_{T-1}^m)$. By (2.9), and $I > 0$, $E(p_T S_T) = I^m E[z_T^{1-m} (1 - B_T(z_T))]$, while by (3.4), $I > 0$, and (2.7), $E(h_T I_{T-1}) = I^m E[\eta_T z_T^{1-m} B_T(z_T)]$. Using these results and $E(I_{T-1}^m v_{T-1}) = I^m E(B_T(z_T)^m v_{T-1})$, one can verify that (3.3) holds. \square

In view of Lemma 3.1, given any nonnegative I and \mathbf{z} , the profit equals

$$\pi(I, \mathbf{z}) = I^m v(\mathbf{z}) - cI. \quad (3.5)$$

One can get an optimal policy by finding first a \mathbf{z}^* maximizing $v(\mathbf{z})$ and then an I^* maximizing

$\pi(I, \mathbf{z}^*)$. The resulting policy is an OSS policy. The arguments for Theorem 2.1 prove the next theorem.

THEOREM 3.1. If \mathbf{z}^* maximizes $v(\mathbf{z})$, then

$$I^* = \left(mv(\mathbf{z}^*)/c \right)^b \text{ and } \pi(I^*, \mathbf{z}^*) = (1-m)m^{b-1}c^{1-b}v(\mathbf{z}^*)^b. \quad (3.6)$$

The remaining results cover the problem of finding \mathbf{z}^* and $v(\mathbf{z}^*)$. The theorem below, which follows directly from Lemma 3.1, deals with the standard case, where in period t one knows only the probability distributions of A_t through A_1 . Given the procedures of this theorem, computing an optimal involves the same amount of work a computing an optimal policy for MPZ.

THEOREM 3.2. Suppose the standard case holds.

(a) One can obtain an optimal \mathbf{z}^* as follows: Set $v_0^* = 0$; for $t = 1$ to T , let

$$v_t^* = \max_{z_t \geq 0} z_t^{1-m} \left(1 - (1 + \eta_t) EB_t(z_t) \right) + E \left(B_t(z_t)^m \right) v_{t-1}^*, \quad (3.7)$$

and let z_t^* achieve the maximum in (3.7).

$$(b) v(\mathbf{z}^*) = v_T^*. \quad (3.8)$$

4. THE MODIFIED MODEL

Finding an OSS can be computationally demanding. §5 develops a heuristic algorithm, whose main subroutine is a procedure that finds an optimal \mathbf{z} for a *modified model* where in period t instead the holding cost equaling $h_t I_{t-1}$ the holding cost equals $\lambda h_t I_{t-1}^m$ for some given $\lambda \geq 0$. This section covers the modified model. We prove OSS policies are optimal among all policies and show that computing OSS policies requires the same work as in MPZ. To distinguish the modified model from the original, we employ the notation $\bar{\pi}(I, \mathbf{z})$ and $\bar{V}(I, \mathbf{z})$ in place of $\pi(I, \mathbf{z})$ and $V(I, \mathbf{z})$. This $\bar{\cdot}$ -notation suppresses the dependence of these objects on λ . Note that the modified model reduces to MPZ if $\lambda = 0$.

The revenue functions for the original and modified models are identical; however, the expected holding cost functions are different, equaling $\lambda \sum_{t=1}^T h_t EI_{t-1}^m$ for the modified model.

Lemma 4.1 below characterizes $\sum_{t=1}^T h_t EI_{t-1}^m$. For nonnegative \mathbf{z} , let $\bar{q}(\mathbf{z}) = \bar{q}_T$, where \bar{q}_T is defined by

$$\bar{q}_0 = 0 \text{ and } \bar{q}_t = (h_t + \bar{q}_{t-1}) E \left(B_t(z_t)^m \right), \text{ for } 1 \leq t \leq T; \quad (4.1)$$

LEMMA 4.1. Given nonnegative I and \mathbf{z} ,

$$\sum_{t=1}^T h_t EI_{t-1}^m = I^m \bar{q}(\mathbf{z}); \quad (4.2)$$

furthermore, $q(\mathbf{z})$ (of Lemma 2.2) and $\bar{q}(\mathbf{z})$ satisfy the following:

$$q(\mathbf{z}) = \sum_{t=1}^T h_t \prod_{j=t}^T EB_j(z_j) \text{ and } \bar{q}(\mathbf{z}) = \sum_{t=1}^T h_t \prod_{j=t}^T E \left(B_j(z_j)^m \right) \quad (4.3)$$

$$q(\mathbf{z}) \leq \bar{q}(\mathbf{z}). \quad (4.4)$$

PROOF. An induction proof like the one for Lemma 2.2 establishes (4.2). Using $\bar{q}_0 = 0$ and

(4.1), one can prove $\bar{q}_T = \sum_{t=1}^T h_t \prod_{j=t}^T E \left(B_j(z_j)^m \right)$. Similarly, using $q_0 = 0$ and (2.11), one can

prove $q_T = \sum_{t=1}^T h_t \prod_{j=t}^T EB_j(z_j)$. These results, $\bar{q}(z) = \bar{q}_T$, and $q(z) = q_T$ give us (4.3). Finally, $0 < m < 1$, the $\alpha_j(z_j)$ s being in the interval $[0,1]$, the nonnegativity of the h_t s, and (4.3) imply (4.4). \square

Results (2.14) and (4.2) imply

$$\bar{V}(I, z) = I^m r(z) - \lambda I^m \bar{q}(z), \text{ for nonnegative } I \text{ and } z. \quad (4.5)$$

Note that (4.5), $V(I, z) = I^m r(z) - Iq(z)$, and (4.4) imply that $\bar{V}(I, z)$ is a lower bound on $V(I, z)$ if $\lambda \geq I^{1-m}$ and an upper bound on $V(I, z)$ if $\lambda = 0$. The heuristic algorithm of §5 exploits this observation.

More useful than (4.5) is the following corollary to Lemma 4.1.

COROLLARY 4.1. Given nonnegative I and z ,

$$\bar{V}(I, z) = I^m \bar{v}(z), \quad (4.6)$$

where $\bar{v}(z) = \bar{v}_T$, with \bar{v}_T being determined by $\bar{v}_0 = 0$, and

$$\bar{v}_t = E \left[z_t^{1-m} (1 - B_t(z_t)) + B_t(z_t)^m (-\lambda h_t + \bar{v}_{t-1}) \right], \text{ for } 1 \leq t \leq T. \quad (4.7)$$

PROOF. Let $\bar{v}_t = r_t - \lambda \bar{q}_t$, for $0 \leq t \leq T$, where the r_t s and \bar{q}_t s satisfy the conditions of Lemmas 2.2 4.1. Using (4.5), Lemma 2.2, and Lemma 4.1, one can easily verify that (4.6) holds. \square

The situation is similar to the one in §3. Given any nonnegative I and z , the profit equals

$$\bar{\pi}(I, z) = I^m \bar{v}(z) - cI. \quad (4.8)$$

One can get an optimal policy by first finding a \bar{z}^* maximizing $\bar{v}(z)$ and then finding an \bar{I}^* maximizing $\bar{\pi}(I, \bar{z}^*)$. The resulting optimal policy is an OSS policy. The theorem below provides specifics.

THEOREM 4.1. Suppose the standard case holds.

(a) One can obtain an optimal \bar{z}^* for the modified model as follows: Set $v_0^* = 0$; for $t = 1$ to T , let

$$\bar{v}_t^* = \max_{z_t \geq 0} z_t^{1-m} (1 - EB_t(z_t)) + E(B_t(z_t)^m) (-\lambda h_t + \bar{v}_{t-1}^*), \quad (4.9)$$

and let \bar{z}_t^* achieve the maximum in (4.9).

$$(b) \bar{v}(\bar{z}^*) = \bar{v}_T^*, \bar{I}^* = (m\bar{v}_T^*/c)^b \text{ and } \bar{\pi}(\bar{I}^*, \bar{z}^*) = (1-m)m^{b-1}c^{1-b}(\bar{v}_T^*)^b. \quad (4.10)$$

PROOF. Part (a) follows from Corollary 4.1; (b) follows from the arguments for Theorem 2.1. \square

In view of Theorem 4.1, computing an optimal policy for the modified model involves the same work as computing an optimal policy for MPZ. Indeed, the modified model has the same structure as MPZ.

5. A HEURISTIC

The last two sections covered special cases where OSS policies are optimal and where obtaining them entails the same work as in MPZ. In general, however, OSS policies are suboptimal and finding them entails computing a T -dimensional vector z^* that globally maximizes the non-concave function π_0 . This section develops a heuristic, which exploit the necessary conditions

of Corollary 5.1 below.

Under the MPZ policy, which ignores holding costs, $\mathbf{z} = \mathbf{z}^{MPZ}$ and $I = I^{MPZ}$, where \mathbf{z}^{MPZ} maximizes $r(\mathbf{z})/c$ and $I^{MPZ} = \left(mr(\mathbf{z}^{MPZ})/c \right)^b$.

$$(5.1)$$

This policy tends to make I too high. In particular, the nonnegativity of $q(\mathbf{z})$, (2.16), and (5.1) imply

$$I(\mathbf{z}) \leq I^{MPZ}, \text{ for all nonnegative } \mathbf{z}. \quad (5.2)$$

One can improve on MPZ with MPZX, which selects \mathbf{z}^{MPZ} and $I(\mathbf{z}^{MPZ}) \equiv \left(mr(\mathbf{z}^{MPZ}) / (c + q(\mathbf{z}^{MPZ})) \right)^b$. Furthermore, one can do even better with heuristics, based on the corollary below.

COROLLARY 5.1. Nonnegative I^* and \mathbf{z}^* constitute an OSS only if

$$I^* = I(\mathbf{z}^*) \text{ and } \mathbf{z}^* \text{ maximizes } r(\mathbf{z}) - \lambda^* q(\mathbf{z}) \text{ for } \lambda^* = (I^*)^{1/b}. \quad (5.3)$$

PROOF. Let I^* and \mathbf{z}^* constitute an OSS. By Theorem 2.1(b), $I^* = I(\mathbf{z}^*)$. Also, \mathbf{z}^* must maximize $V(I^*, \mathbf{z})$. These results imply (5.3), since $V(I^*, \mathbf{z}) = (I^*)^m \left(r(\mathbf{z}) - (I^*)^{1/b} q(\mathbf{z}) \right)$. \square

Corollary 5.1 could lead to a search algorithm for an OSS. Consider the following naïve procedure.

Step 1. Let $\mathbf{z}^{(1)} = \mathbf{z}^{MPZ}$ and let $I^{(1)} = I(\mathbf{z}^{MPZ})$. Set $n = 1$.

Step 2. Let $\mathbf{z}^{(n+1)}$ maximize $\pi(I^{(n)}, \mathbf{z})$, i.e., $\mathbf{z}^{(n+1)}$ maximizes $r(\mathbf{z}) - \lambda^{(n)} q(\mathbf{z})$ for $\lambda^{(n)} = (I^{(n)})^{1/b}$.

Step 3. Let $I^{(n+1)} = I(\mathbf{z}^{(n+1)})$. Compute $\pi(I^{(n+1)}, \mathbf{z}^{(n+1)})$. Stop if a given stopping criterion is satisfied else set $n = n + 1$ and repeat Step 2.

Note that $\pi(I^{(n)}, \mathbf{z}^{(n+1)}) \geq \pi(I^{(n)}, \mathbf{z}^{(n)})$, implying, $\pi(I^{(n+1)}, \mathbf{z}^{(n+1)}) \geq \pi(I^{(n)}, \mathbf{z}^{(n)})$, for $n \geq 1$. Hence, $\pi(I^{(n)}, \mathbf{z}^{(n)}) \uparrow \pi^*$, where π^* is finite. Clearly, $\pi^* \leq \max_{I, \mathbf{z}} \pi(I, \mathbf{z})$. Even if equality fails to hold, π^* might be a worthwhile heuristic value.

The sticking point is Step 2, which seeks a \mathbf{z} maximizing $r(\mathbf{z}) - \lambda q(\mathbf{z})$. Finding such a \mathbf{z} is harder than the task, covered in Theorem 4.1, of finding a \mathbf{z} maximizing $r(\mathbf{z}) - \lambda \bar{q}(\mathbf{z})$: The approach of Theorem 4.1 does not apply, since one cannot get simple recursions relating $r_t - \lambda q_t$ to $r_{t-1} - \lambda q_{t-1}$, for $t \geq 2$: If one defines $v_t = r_t - \lambda q_t$, for $0 \leq t \leq T$, where the r_t s and q_t s satisfy Lemma 2.1, then $v_0 = 0$ and

$$v_t = z_t^{1-m} - \left(z_t^{1-m} + \lambda h_t \right) E B_t(z_t) + v_{t-1} E \left(B_t(z_t)^m \right) + \lambda \left(E \left(B_t(z_t)^m - B_t(z_t) \right) \right) q_{t-1}, \text{ for } 1 \leq t \leq T. \quad (5.4)$$

The appearance of q_{t-1} in (5.4), which itself satisfies recursion (2.14), prevents one from finding a maximizing \mathbf{z} by first, finding a maximizing z_1 , second, finding a maximizing z_2 , and so forth. Indeed, Example 5.1 below illustrates a situation where (z_1, z_2, z_3) is optimal for a 3 period problem, without (z_1, z_2) being optimal for a 2 period problem. (This situation cannot occur for

the MPZ model.) One could, of course, determine $\mathbf{z}^*(\lambda)$ by employing a standard optimization; however, if one is going to resort to such an algorithm, one might be better off applying it directly to the task of maximizing π_0 , thus avoiding the use of a heuristic.

EXAMPLE 5.1. Demands are deterministic. For $1 \leq t \leq T$, let $\mu_t = 1000$ and $h_t = 0.3$. Also, let $b = 2$ and $c = 1$. Our computations for an OSS yield the following results (two decimal place accuracy). First, if $T = 2$, then $I^* = 410.00$, $\mathbf{z}^* = (1000.00, 1640.00)$, $\pi(I^*, \mathbf{z}^*) = 450.00$.

Second, if $T = 3$, then $I^* = 512.11$, $\mathbf{z}^* = (1000, 1694.44, 2084.44)$, $\pi(I^*, \mathbf{z}^*) = 616.67$. Thus z_2^* changes when one changes T from 2 to 3.

Our heuristics exploit §4, working with $r(\mathbf{z}) - \lambda \bar{q}(\mathbf{z})$ instead of $r(\mathbf{z}) - \lambda q(\mathbf{z})$. For $\lambda \geq 0$, let

$$\mathbf{z}(\lambda) \text{ maximize } r(\mathbf{z}) - \lambda \bar{q}(\mathbf{z}). \quad (5.5)$$

The quantity $\mathbf{z}(\lambda)$ is identical to $\bar{\mathbf{z}}^*$ of §4. (The notation $\mathbf{z}(\lambda)$ makes the dependence on λ explicit.) Using techniques of §4, evaluating $\mathbf{z}(\lambda)$ requires the same work as solving MPZ model.

Our heuristic, H0, selects $\mathbf{z}^{H0} = \mathbf{z}(\lambda^{H0})$ and $I^{H0} = I(\mathbf{z}^{H0})$, where

$$\lambda^{H0} = \left(I(\mathbf{z}^{MPZ}) \right)^{1/b}. \quad (5.6)$$

Referring to our three step procedure above, $\lambda^{H0} = \lambda^{(1)}$. Thus, $\pi(I^{H0}, \mathbf{z}^{H0})$ is our approximation to $\pi(I^{(2)}, \mathbf{z}^{(2)})$. Since $\pi(I^{(2)}, \mathbf{z}^{(2)}) \geq \pi(I^{(1)}, \mathbf{z}^{(1)})$ and $\pi(I^{(1)}, \mathbf{z}^{(1)}) = \pi(I(\mathbf{z}^{MPZ}), \mathbf{z}^{MPZ})$, it is plausible that $\pi(I^{H0}, \mathbf{z}^{H0}) \geq \pi(I(\mathbf{z}^{MPZ}), \mathbf{z}^{MPZ})$, the profit under MPZX. This inequality holds in all our test problems.

6 REFERENCES

- Monahan G. E., N C. Petruzzi and W. Zhao. 2004, The Dynamic Pricing Problem from a Newsvendor's Perspective, MSOM, 6, 73691.
- Petruzzi, N. C., M. Dada. 1999. Pricing and the newsvendor problem: A review with extensions. Oper. Res. 47, 1836194.
- Petruzzi, N. C., M. Dada, 2002. Dynamic pricing and inventory control with learning. Naval Res. Logist. 49, 3046325.
- Zabel, E. 1972, Multiperiod monopoly under uncertainty. J. Economic Theory 5 5246536.
- Zhao, W., Y. Zheng. 2000. Optimal dynamic pricing for perishable assets. Management Science. 46, 3756388.

Modelling Values of Lake Ellesmere

John F. Raffensperger
Dept. of Management, Univ. of Canterbury
john.raffensperger@canterbury.ac.nz

Ken Hughey
Faculty of Environment, Society and Design, Lincoln University
ken.hughey@lincoln.ac.nz

Abstract

This paper describes a model called PLOVER that is being used to help guide the process of planning openings for Lake Ellesmere. PLOVER is a deterministic simulation based on hydrological data. Water levels and weather drive other components in the model, including water quality, eel and flounder migration, and farm values. The model was able to deliver on multiple management scenarios associated with lake level management. These scenarios were used to inform community discussions about future management of the lake.

Key words: ecology, community planning, scenario modelling

1 Lake Ellesmere: a treasure of variability

Lake Ellesmere is New Zealand's fifth largest lake by area. It is a lowland semi-saline and bar-type lagoon located on the east coast of the South Island – it is also one of New Zealand's great ecological treasures. While usually closed off from the Pacific Ocean, it can fill quickly over a few months, requiring several openings per year to allow the lake to drain. The timing of these openings and the associated lake levels have significant impacts on the local farmers, recreational users, and the lake's flora and fauna, especially the eel and flounder fisheries.

For many decades, the lake has been controlled primarily on the basis of depth: it was opened to the sea whenever the lake level rose over about 1.05 meters, except from 1 April to 31 July, when the trigger level was a bit higher at 1.13 meters. Meanwhile, the community and scientists raised a number of concerns about the lake, that a range of ecological values appeared to be in poor health. The nadir was reached in 2005 when an Environment Court judge declared the lake “technically dead” (Booker 2007). Following that, the community pulled together, through a community group called the Waihora Ellesmere Trust (WET). This group organised two important community events, the Living Lakes Symposium 2007, and the Living Lakes Symposium 2 on 4 Nov 2009, in collaboration with the local regional council Environment Canterbury.

The first Living Lakes Symposium provided an important status report on the lake, covering ground water, water quality, lakeshore vegetation, fish, birds, cultural values, recreation, economic values, and some ideas about future management (WET 2007, Hughey and Taylor 2009). The name “Living Lakes,” a reaction to the judge's declaration, served to clarify that Te Waihora/Lake Ellesmere is very much alive. Among other things, the community recognized the need for a wide range of additional scientific research. Part of this additional need was for a model that would synthesize the lake's key values, allowing some means to make better decisions about lake

management. Lake levels were clearly the main driver, so any tool would have to take the hydrology into account.

This paper reports on a model, PLOVER 2k, developed for the Living Lakes Symposium 2. PLOVER is a deterministic Excel-based simulation which attempts to measure some changes in ecological and economic values as a function of the lake opening schedule. The purpose of PLOVER is to help choose lake opening regimes, to find a balance between farming, fishing, recreation, flora, and fauna.

2 Quality of results in PLOVER

Data in PLOVER comes from a range of sources. Wherever possible, PLOVER relies on quantitative effects rather than qualitative judgement. Most importantly, PLOVER is based heavily on a water balance model developed at NIWA (Horrell 2009). Quality of data is designated at three levels:



Strong, based on causal relationships and considerable data. Example: water flows are based on the NIWA water balance model.



Okay, based on a correlative relationships and a reasonable quantity of data. Example: salinity was calculated as a function of water depth, using Environment Canterbury temperature information.



Hazy, based on anecdotal data or speculative relationships. Example: blue-green Nodularia algae are a “threat” formula based on a few reported events and general scientific data on likelihood of algae.

3 Components in PLOVER

Components were selected during lengthy consultation. Ideally, PLOVER would contain more components, such as more types of birds and fish, and in greater detail. However, the components and their detail were limited by data availability. Further, some components may actually be misleading, due to shortcomings in PLOVER. For example, we have modelled *Ruppia* (an aquatic macrophytic plant) only in sprouting, not in growth, because we have a plausible model of sprouting, but none for growth; maximising sprouting may actually hinder growth. Figure 1 shows an influence diagram of PLOVER’s components. Figure 4 shows graphs of the main components for one lake level management scenario.



Date follows the NIWA water model, which starts 5 Jan 1970 and ends with 31 Dec 2007. PLOVER uses date arithmetic heavily. PLOVER does not attempt to predict the future.



Hours of sunlight drives seasonality. We obtained hours of sunlight from sunrise and sunset times, from http://aa.usno.navy.mil/data/docs/RS_OneYear.php, 172:45 east, 43:30 S, 12 hrs E of Greenwich.



Temperature data came from ECan (2009), a linear regression of temperature to day of year.



Lake depth is calculated directly by the NIWA water model.

Opening. An opening regime is defined with date and depth pairs at which to start an opening attempt. The opening date and depths are the only decision variables in the model. PLOVER uses three dates and three associated water levels, with a boolean to indicate whether an opening regime lasts only for 30 days starting from the input date (with an additional 1,200mm “circuit breaker” trigger is maintained year round), or whether the opening regime lasts beyond 30 days, continuing to the next date. The latter case is more of a depth-based regime, not a date based regime.

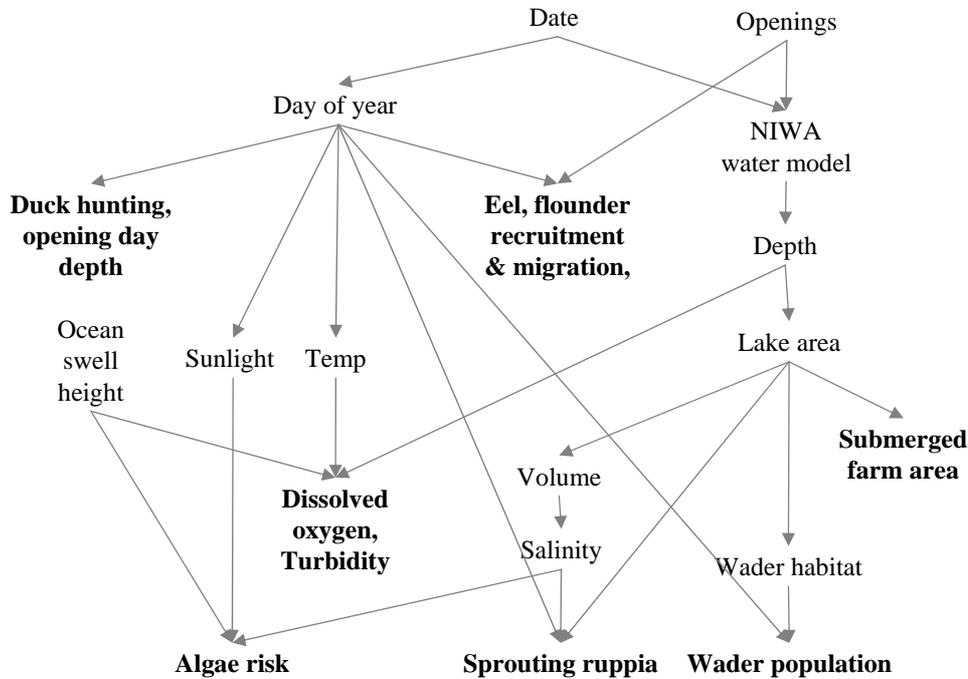


Figure 1. Influence diagram of PLOVER’s components.

PLOVER treats each water level as the depth trigger at which an opening should be attempted, for the period between the dates. The lake is inherently a chaotic system – a small change early in the planning horizon can result in large changes later in the planning horizon, and the NIWA water model reflects that chaotic behaviour. Figure 4 includes a graph of a simulated 5 years of lake level.

Because an opening attempt does not always succeed, we must define what we mean by “opening”. The definition which best matches the associated costs of opening is the number of first-day openings at least 30-days apart, i.e., a 30-day continuous attempt to open the lake. For example, if we select 1 May as an opening date, with a threshold of 1,000 mm, then PLOVER will not open the lake in May if depth is below 1,000 mm. PLOVER will try to open the lake if depth is above 1,000 mm. If the lake does come below 1,000 mm, PLOVER relies on the NIWA water model to decide when the opening stops. If during May the lake comes below 1,000, and the NIWA water model closes the lake, but the depth again exceeds 1,000 mm, PLOVER will try to open the lake again, for 30 days. PLOVER may try to open the lake on the last of these days, and the NIWA water model may allow the lake to stay open until after that. An opening may take several days to initiate due to rough weather, so opening dates may need to be set somewhat earlier, and the decision depths may need to be set somewhat lower, to achieve the desired policy. Figure 2 shows an example of roughness in the model.

The baseline values in the model correspond to the opening regime as modelled by the NIWA water balance model.



Area of the lake is calculated from detailed GIS data supplied by ECan (Hill, 2009). This GIS data is a combination of AgriBase and NIWA LIDAR data. The model is uncertain only due to wind lash, which has not been included. That is, the inundated area assumes calm weather. Except for wind lash, this is excellent data.



Volume of the lake is calculated from the ECan GIS data.

1319	27-Jul-73	27-Jul	1,249	Closed
1320	28-Jul-73	28-Jul	1,347	Closed
1321	29-Jul-73	29-Jul	1,427	Closed
1322	30-Jul-73	30-Jul	1,398	Open
1323	31-Jul-73	31-Jul	1,363	Open
1324	1-Aug-73	1-Aug	1,329	Open
1325	2-Aug-73	2-Aug	1,293	Open
1326	3-Aug-73	3-Aug	1,257	Open
1327	4-Aug-73	4-Aug	1,221	Open
1328	5-Aug-73	5-Aug	1,280	Open
1329	6-Aug-73	6-Aug	1,243	Open
1330	7-Aug-73	7-Aug	1,255	Closed
1331	8-Aug-73	8-Aug	1,267	Closed
1332	9-Aug-73	9-Aug	1,285	Closed
1333	10-Aug-73	10-Aug	1,301	Closed
1334	11-Aug-73	11-Aug	1,317	Closed
1335	12-Aug-73	12-Aug	1,287	Open
1336	13-Aug-73	13-Aug	1,255	Open
1337	14-Aug-73	14-Aug	1,224	Open
1338	15-Aug-73	15-Aug	1,191	Open
1339	16-Aug-73	16-Aug	1,159	Open
1340	17-Aug-73	17-Aug	1,125	Open
1341	18-Aug-73	18-Aug	1,092	Open
1342	19-Aug-73	19-Aug	1,058	Open
1343	20-Aug-73	20-Aug	1,025	Open
1344	21-Aug-73	21-Aug	1,043	Closed
1345	22-Aug-73	22-Aug	1,061	Closed
1346	23-Aug-73	23-Aug	1,077	Closed
1347	24-Aug-73	24-Aug	1,094	Closed

Figure 2. In the (simulated) days before 30 Jul 1973, depth is above the trigger of 1.05m, but the lake could not be opened to due rough weather. When it was opened on 30 Jul, it could be kept open only for a week.



Dissolved oxygen also came from ECan (2009), calculated as a multiple linear regression of hours sunlight, ocean swell height offshore from Lake Ellesmere (Horrell 2009), lake level, and temperature. This is probably overkill, as sunlight and temperature are correlated, so one of those could be dropped to have a more concise model, but no more predictive.



Salinity depends on openings, approximated salinity as a function of volume, based on data from ECan (2009). If we suppose the lake has a constant amount of salt S as a result only of mixing during opening, salinity may be approximated as S/volume . We examined other models of salinity, including salinity as a linear function of depth, and salinity as a power function of depth. The power function has a slightly better fit, but doesn't make as much sense physically. The function is given as salinity = $2,407/\text{volume} - 1.2$, where volume was calculated as above, in millions of cubic meters. Taylor (1996, p. 18) says "mean salinity of 8.5 ppt," and ECan (2009) indicates a mean of 6.9; PLOVER is showing about 7.0.



Nodularia algae risk is based on Taylor (1996) and Hughey & Taylor (2009), indicating that this microbe, *Nodularia spumigena*, blooms under high heat and high salinity. Taylor (1996), mentions three specific blooms, in autumn 1971 (p180), March 1981 (p180), and autumn 1990 (p180). From that, we created a pseudo-time series, and observed that ocean swell height (Horrell 2009) was never over 5,000 mm when there were algae. Taylor 1996, p180, also states that the optimum growth occurs “in the range of 5-20 ppt”. We therefore created a multiplicative function, if (wave height $\geq 5,000$, 0, $\text{MAX}(0, (\text{hours sunlight} - 8:55) * (\text{salinity} - 5)) * 10$, multiplying by 10 simply to have a nice scale. Hence, blue-green algae are likely to grown under conditions of light wind, long daylight and high salinity. This produces values between 0 and about 1.2.



Turbidity was calculated as a multiple regression, based on depth, temperature, and ocean-wave height, using data from ECan (2009) and Horrell (2009). Turbidity is not used elsewhere in the model. The macrophyte calculations could be improved by accounting for turbidity, but sufficient data is not immediately available.



Submerged macrophytes, sprouting should be interpreted as the expected hectares with sprouting macrophytes. It is a qualitative measure of the opportunity of macrophytes to sprout.

Sprouting time. Taylor 1996, Hughey & Taylor (2009), and Jellyman et al (2009) indicate that macrophytes sprout in “the spring,” but do not specify what that means. Hammer (1986, p. 347) writes of Wakaw Lake in Saskatchewan, “Growth of *Ruppia* and *P. pectinatus* began in the middle of May at temperatures of 10°C. Seeds sprouted at 8°C. Flowering began by the middle of June and produced fruit by the middle of July...By October visible plants were senescent...” He refers to *Ruppia maritima*, while Lake Ellesmere has *Ruppia megacarpa* and *Ruppia polycarpa*. We decided to use the values 8°C and 10°C. The Lake Ellesmere temperature data indicates that the lake reaches 8°C about 17 August, and it reaches 10°C around 11 September.

Sprouting conditions. Taylor (1996, p. 182) indicates that growing is best when salinity is between 0 and 8. If *Ruppia* sprouts in late August and requires low salinity, then *Ruppia* prefers **no openings** between 15 June and 15 September, because the increased salinity will slow sprouting. A paper by da Silva & Asmus (2001) develops a detailed simulation model of *Ruppia maritima* in Brazil, which uses an exponential function of control by salinity on germination velocity. We used it here: the probability that sprouting occurs depends on $e^{-0.048 * \text{salinity}}$.

We therefore created a distribution function that returns a normal density value, implying a *centre* of sprouting on 29 Aug, with a distribution of 4 days around that, so that sprouting began just when average temperature became 8°C. We then multiply this value by $e^{-0.048 * \text{salinity}}$; the function returns 1 if salinity is 0, 0.68 if salinity is 8. Next, we assumed a given fraction, e.g., 10%, of the area is of correct depth for sprouting. Finally, we multiply by the lake area. This value may be interpreted as the expected hectares with sprouting macrophytes. Available light is known to strongly affect growth, but we did not attempt to take account depth or turbidity in sprouting.



Shortfin eel is the fraction of ideal migration. It was computed with a distribution function that returns a normal density value centred at the maximum days of ingress and egress. Since both dates are considered to help the eel population, the two

values are summed and divided by 2. Thus, the function conveys the fraction of ideal migration for eel.

From Chisholm (2009), pp. 36-37:

...the SIEIA's preferred times for the lake to be open and for it not to be open are as follows:

- One or two openings during spring/summer (October – mid December)
- No further openings during the summer months (mid December – mid April), with lake levels preferably kept high
- One or more openings during late April/May

As the Chisholm report was developed by the eel fishing industry, we assume that this is what they want. Centre-of-ideal dates were therefore taken as 1 Nov, with a standard deviation of 12 days, and 1 May with a standard deviation of 7 days. We then multiplied this distribution function by 1 if the lake were actually open, else 0.

Butcher (2009) suggested that the baseline value of the eel fishery is \$430,000. However, a recent news article (Dominion 2009), which indicates the earnings (not profit) of the eel fishery at \$360,000/year. Chisholm (2009) does not give a value, and the Ministry of Fisheries website shows blanks for data. We used the lower value, and multiplied the total \$ by the change in average area from mid-December to mid-April.

Chisholm also states, "A high lake during the mid-summer months (mid-December – mid April) is desirable, as it allows eel access to the (more desirable) riparian feeding grounds. A low lake during the mid-summer is detrimental to the fisheries, as it can lead to algal blooms and high water temperatures. Both these events can lead to fish dying." None of these components were included, thus the measure given for eel may actually be detrimental to them.



Flounder is the fraction of ideal migration. It was computed with a distribution function that returns a normal density value centred at the maximum days of ingress and egress. Taylor (1996, p. 211) indicates migration is "August to November," so we have used a mean centre-of-ideal day of 1 October, with a large standard deviation of 20 days. We then multiplied this by 1 if the lake were actually open, else 0. The dollar value from Butcher (2009) is scaled to the baseline percentage. This value will go up with more openings, because flounder have a greater chance to migrate. The large standard deviation implies that 30 continuous days of opening, even at the peak migration season, would get only a fraction of the flounder. Flounder could appear to have a greater gain if this standard deviation were smaller, because a larger fraction would be modelled as entering over a shorter period. Note that this measure ignores the size and quality of the habitat.



Duck hunting is the (daily bag limit)*depth. Thus, hunters are assumed to prefer a deeper lake during hunting season. The measure will be higher to the extent that the lake is high during hunting season. This measure does not put any special weight on the May opening day.



Duck hunting, opening day avg depth is the depth on the first Saturday of each May. This puts all the weight on opening day. The Lake Ellesmere workshop participants indicated that duck shooters liked to have the lake at a depth between 800mm and 1300mm on opening day. We have assumed that higher is better.



Hectares of quality-adjusted wader habitat indicates total wader habitat over the growing season. The “raw” area was calculated as the lake area for the current depth, minus the lake area for the current depth minus 100mm. Hence, this provided the area at the margin of the lake. For each GIS land type, Ken Hughey provided a % quality of habitat. This quality was multiplied by the “raw” area of each land type at the margin of the lake, to provide the hectares of quality-adjusted habitat.



Indigenous waders is population, scaled by habitat area: (wader population on day t) = (observed population on day t)*(wader habitat area on day t)/(baseline habitat area on day t). Thus, this indicator is highest when both habitat and birds are at high levels together. Hughey & Taylor (2009) contains 1986-7 populations by month. We interpolated those values by day.



Hectares submerged by land type are look-up values based on the ECan GIS data. The most important of these is “SNB Mixed Sheep and Beef farming,” because this is the largest area. The areas were scaled by a cost per hectare per day of inundation, and also adjusted by “tolerance depths” from Hearnshaw (2009).

4 Comments

Of all components in PLOVER, it appears that eel fishing could have the greatest increase in benefit, as measured by migrations in PLOVER. The current opening regime does not seem to favour eel, because openings are timed simply on the basis of depth, not particularly with respect to eel’s preferred recruitment and migration dates. An opening regime that does favour eel would appear to improve migrations by a large amount. By either % or \$ measure of effectiveness, a regime which improves eel – *based on PLOVER in its current state* – significantly improves both measures of effectiveness for the lake overall. The reason is because eel fishing is larger financially than any other component, and appears also to have the largest % improvement potential. Lowering the opening trigger depth appears to help eel and flounder due to greater chance of migration, as the lake is open more often. Both of these measures ignore habitat, and may even be counter-productive. Chisholm (2009), for example, reports that eel fishing is easier in flood conditions. Measuring the eel fishery only on migration ignores the habitat; switching to a “maximise eel migration” scenario, as given by PLOVER, could actually harm the eel fishery.

A separate issue is whether PLOVER is to be used for short term decisions or long term decisions. According to Chisholm (2009), the Lake Ellesmere commercial eel quota were 100% caught from 2003 to 2006. In the short run, the current quota limits any gain to eel fishing. However, if the modelled gains were valid, the eel quota for Lake Ellesmere could possibly be raised.

5 Use in Living Lakes Symposium 2

To prepare for the symposium, the analyst was tasked with developing a set of scenarios. This was straightforward: for each key component in PLOVER, simply find a set of dates and depths which maximised its value. This resulted in nine different scenarios, some of which were obviously undesirable. From these nine, the project’s management team chose three scenarios. The analyst created an additional “blend” scenario. Symposium participants therefore had four scenarios to review, in comparison

with the baseline status quo scenario. Figure 3 shows model output for one scenario, and Figure 4 shows graphs of the main components for one scenario. Those two figures were combined on a single page, for each scenario.

Before the model was presented, different scientists gave presentations on key aspects of Lake Ellesmere, including the hydrology and its modelling by NIWA, the eel and flounder fisheries, and water quality. A senior manager at Environment Canterbury gave a presentation about modelling generally, why models are used, what they can do and what they cannot do. In early afternoon, PLOVER was presented. This was followed by presentations by scientists and some community representatives attesting to PLOVER’s strengths and weaknesses.

In late afternoon, participants had an opportunity to study the scenarios in small group breakout sessions, with feedback to the larger group. In the evening, additional discussion generated several interesting ideas. For example, one participant suggested that the opening day for duck hunting could be made more flexible, thus allowing more flexibility in the timing of openings, while ensuring good lake levels on the first day of hunting. Generally, participants recognized the value of the model, and understood that PLOVER enabled a more objective method of lake management.

	18-Apr	24-Jul	20-Sep	
First day of opening attempt	18-Apr	24-Jul	20-Sep	
Trigger depth, mm	600	1,200	500	Opening attempt for 30 days only, with 1,200mm circuit breaker
	Baseline	Scenario	% change	worse. better.
Lake depth, mm	841	770	-8%	
Opening cost	-\$124,106	-\$128,619	4%	
# of openings	3.8	4.3	13%	
Lake area, h	19,554	19,178	-2%	
Volume, million m ³	308	295	-4%	
Dissolved oxygen	11.0	10.9	0%	
Salinity, parts/000	6.8	7.2	7%	
Nodularia algae risk	2.7	3.4	23%	
Turbidity, NTU	87.6	91.0	4%	
Sprouting ruppia, h	398.8%	445.6%	12%	
Eel recruitment & migration	\$360,000	\$887,051	146%	
Flounder recruitment	\$200,000	\$386,362	93%	
Duck hunting, opening day depth	838	587	-30%	
Wader habitat, h	255	304	19%	
Waders, population	4,348.3	5,208.0	20%	
Farm covered	-129,978	-121,923	-6%	
Total \$000	\$305,917	\$1,022,871	234%	

Figure 3. Model output shown for one scenario from PLOVER.

6 Suggested future work.

PLOVER, and the understanding of Lake Ellesmere, could be improved by more focused science.

Eel. Because eel appears to be the largest dollar value associated with the lake, a simulation of eel growth, taking into account both openings and habitat, would be the largest-value research.

Ruppia. Silva & Asmus (2001) develop a detailed simulation model of *Ruppia maritima* in Brazil. Their model could be calibrated to the *Ruppia* in Lake Ellesmere, and should be helpful in guiding macrophyte reestablishment.

Modelling. The PLOVER model could be improved in at least three ways.

First, Lake Ellesmere could be modelled with existing stochastic reservoir optimisation techniques. This should provide improved policies, such as conditional regimes, e.g., “if the lake depth is below *D* in week *W* then don’t open”. In another

interesting paper, Estrada, Parodi, and Soledad (2009) developed a reservoir model specifically to manage algae growth and eutrophication.

Second, additional time series data, such as for water quality, would improve the strength of the regression relationships.

Third, PLOVER was improved every time another scientist reviewed it. PLOVER would benefit from additional peer review, especially for suggestions in better structures for the various components.

Fourth, Lake Ellesmere should be modelled as part of a larger area, taking in a significant portion of the catchment. This modelling should include water flows, nutrient run-off, and impervious cover, and include both economic and ecological components. This technology is available now, and probably could be done at reasonable cost.

Database. Information about Lake Ellesmere appears in many places. Future research would be considerably easier with a consolidated database, or even a web site, containing GIS data, regularly measurements, a bibliography, and models.

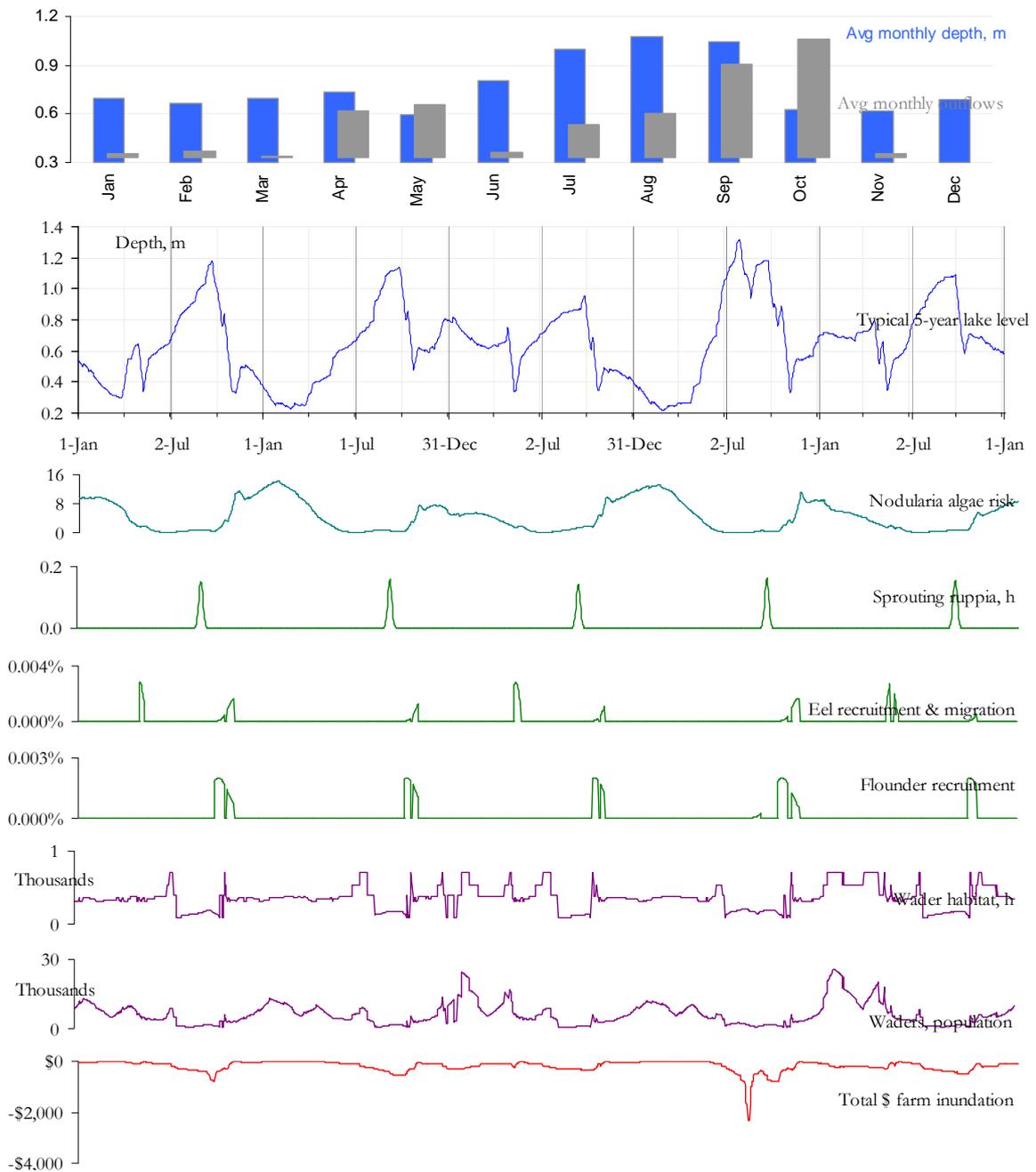


Figure 4. Graphs shown for one scenario from PLOVER.

7 Acknowledgments

Many people provided guidance on this work, especially Ian Whitehouse, Graeme Horrell, Don Jellyman, Zach Hill, Ken Taylor, and Ed Hearnshaw. Thanks also to Environment Canterbury for its considerable support in this project and the ongoing work.

8 References

- Booker, Jarrod, 2007, "There's life after death for Lake Ellesmere," *NZ Herald*, 31 Oct 2007, www.nzherald.co.nz/environment-court/news/article.cfm?o_id=262&objectid=10473013, accessed 6 Nov 2009.
- Butcher, Geoff, "Economic values," pp. 101-110, in: K.F.D. Hughey & K.J.W. Taylor, eds., *Te Waihora/Lake Ellesmere: State of the Lake and Future Management*, EOS Ecology, Christchurch, 2009.

- Chisholm Associates, "South Island Eel Industry Association, Eel Fishery Plan for the South Island," Sept 2009, an industry report.
- da Silva, Eduardo Teixeira, and Milton L. Asmus, "A dynamic simulation model of the widgeon grass *Ruppia maritima* and its epiphytes in the estuary of the Patos Lagoon, RS, Brazil," *Ecological Modelling* 137 (2001) 161–179.
- Dominion Post 2009, news article, <http://www.stuff.co.nz/dominion-post/news/2003175/Tribe-sets-lakebed-fishing-levy>, accessed 24 Oct 2009.
- ECan 2009, "ECan water quality data.xls," a spreadsheet given to me by Ed Hearnshaw, of detailed lake data from Environment Canterbury. The file's properties dialog reports the author as "ECAN".
- Estrada, Vanina, Elisa R. Parodi, and M. Soledad Diaz, "Addressing the control problem of algae growth in water reservoirs with advanced dynamic optimization approaches," *Computers & Chemical Engineering*, **33** 12, 10 Dec 2009
- Hammer, Ulrich Theodore, *Saline Lake Ecosystems of the World*, vol. 59 of *Monographiae Biologicae*, Springer, 1986.
- Hearnshaw, Ed, 1999, data on various maximum and minimum limits associated with Lake Ellesmere, which he obtained in interviews.
- Hughey Kenneth F.D. & Taylor, Kenneth, J.W. eds., *Te Waihora/Lake Ellesmere: State of the Lake and Future Management*, EOS Ecology, Christchurch, 2009.
- Hill, Zach, 2009. This is GIS data, a combination of AgriBase and NIWA LIDAR data.
- Horrell, Graham, 2009. This consists of inflow and weather roughness data, and TIDEDA algorithm files, sent by Graham Horrell, NIWA, July to Sep 2009. We partially converted the algorithms to Visual Basic; Horrell and other staff an NIWA corrected those. We then converted those algorithms exactly to Excel.
- Jellyman, Don, Jeremy Walsh, Mary de Winton, and Donna Sutherland, "A review of the potential to re-establish macrophyte beds in Te Waihora (Lake Ellesmere)," Report No. R09/38, Environment Canterbury, May 2009.
- Taylor, Kenneth, ed., *The Natural Resources of Lake Ellesmere (Te Waihora) and Its Catchment*, Canterbury Regional Council, 1996.
- WET 2007, the *Living Lakes Symposium 2007* page, Waihora Ellesmere Trust website, www.wet.org.nz/events/living-lakes-symposium-2007/, accessed 4 Nov 2009.

New models and methodologies for group decision making, rank aggregation, clustering and data mining

Dorit S. Hochbaum
Haas School of Business and Department of IE&OR
University of California, Berkeley
hochbaum@ieor.berkeley.edu

Abstract

We introduce models for problems of group decision making, aggregate ranking and clustering techniques for data mining. The problems are modeled as graph problems. One of these problems we call the equal paths problem. This problem as well as all problems studied here have convex objective function representing penalties for deviating from specified a-priori comparison/ranking beliefs. These problems are shown to be solvable in polynomial time using network flow techniques such as parametric cut and fractional multicommodity linear programming.

One application of the aggregate ranking problem is to determine the ranking of sports teams based on the outcomes of games played. Current techniques are based on finding a maximum eigenvector. Our alternative model has a number of advantages including the ability to differentiate between games based on some measure of significance. Further, the problem is stated as a combinatorial graph problem. This problem is shown to be solved in polynomial time even with a convex objective function, using flow techniques.

A closely related area that addresses various forms of rankings is data mining with applications to customer segmentation, patient diagnosis and assessment of bankruptcy risk. We demonstrate new models for these problems and how to solve them with flow techniques. Similarly, the models and solution methodology are applicable to multi-criteria decision making.

Approximation Algorithm for Firefighter Problem on Trees

Yutaka Iwaikawa, Naoyuki Kamiyama, and Tomomi Matsui
Department of Information and System Engineering,
Faculty of Science and Engineering, Chuo University,
Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.

Abstract

In this paper, we discuss the firefighter problem on rooted trees. We propose a polynomial time $(1 - \frac{k-1}{1+(k-1)e})$ -approximation algorithm in case that the root vertex has k children.

Key words: firefighter problem, approximation algorithm

1 Introduction

In this paper, we consider the *firefighter problem* on graphs which is used to model the spread of fire, infectious diseases, and computer viruses. In this problem, we are given a graph $G = (V, E)$ with a specified vertex $r \in V$ and a weight function $w: V \rightarrow \mathbb{Z}_+$, where \mathbb{Z}_+ represents the set of nonnegative integers. Then at time 0, a fire breaks out at r . At each subsequent time interval, a firefighter deploys a vertex which is not yet on fire and defends it. The fire spreads to all unprotected adjacent vertices of each burned vertex. The objective of firefighter problem is to determine posture of firefighters so as to maximize the sum of weights of saved vertices.

In this paper, we consider the firefighter problem on rooted trees in which at time 0 a fire breaks out at the root r of a given tree. It is known (Finbow et al. 2007) that this problem is \mathcal{NP} -hard even for rooted trees whose maximum degree is 3. Hartnell and Li (Hartnell and Li 2000) have proved that a simple greedy method gives a 0.5-approximation algorithm. Cai, Verbin and Yang (Cai, Verbin, and Yang 2008) proposed a polynomial time randomized $(1 - \frac{1}{e})$ -approximation algorithm. Furthermore, Anshelevich, Chakrabarty, Hate and Swamy (Anshelevich et al. 2009) showed that this result can be also derived from a reduction to the submodular function maximization problem.

In this paper, we propose an algorithm and improve the approximation ratio for the firefighter problem on trees. More precisely, for a tree in which the root vertex has k children, we propose a randomized $(1 - \frac{k-1}{1+(k-1)e})$ -approximation algorithm. For any positive integer k , the approximation ratio is greater than $1 - \frac{1}{e}$ which is attained by the algorithm in (Cai, Verbin, and Yang 2008).

2 Our Algorithm

Let $T = (V, E)$ be a rooted tree with root $r \in V$, and let v_1, \dots, v_k be the children of the root r . For each $i \in \{1, \dots, k\}$, we use the following notations (see Figure 1).

- Let T_i be a subtree of T rooted at v_i
- Let X_i be a set of vertices of T_i .
- Let \bar{T}_i be a tree obtained by contracting the roots of $T_1, \dots, T_{i-1}, T_{i+1}, \dots, T_k$.
- Let \bar{Z}_i^* be a set of saved vertices in an optimal solution for the firefighter problem on \bar{T}_i .

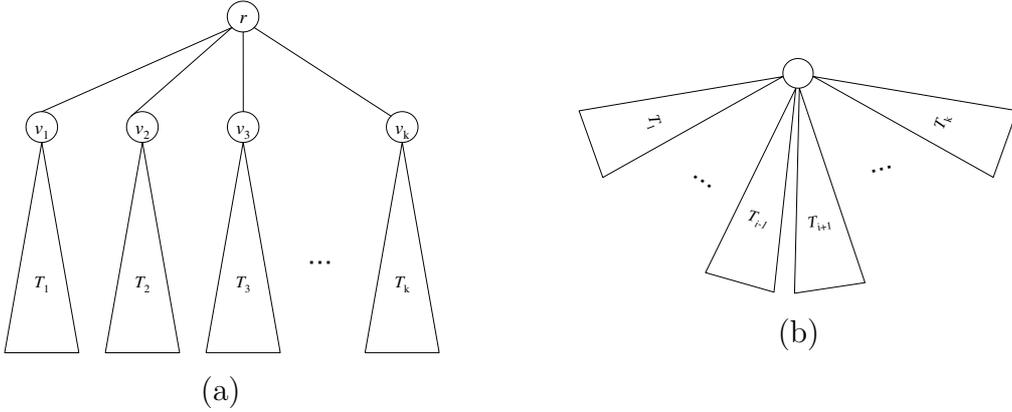


Figure 1: (a) Subtree T_i . (b) Tree \bar{T}_i .

We should note that we can obtain a feasible solution of the firefighter problem on \bar{T}_i such that a set of saved vertices \bar{Z}_i satisfies $w(\bar{Z}_i) \geq (1 - \frac{1}{e})w(\bar{Z}_i^*)$ by using the randomized algorithm in (Cai, Verbin, and Yang 2008), where we use the notation $w(W) = \sum_{v \in W} w(v)$ for each $W \subseteq V$. Then, our algorithm outputs a maximizer $X_i \cup \bar{Z}_i$ of $\max\{w(X_i) + w(\bar{Z}_i) \mid i \in \{1, \dots, k\}\}$.

Let us discuss the approximation ratio of our algorithm. We can assume without loss of generality that $X_1 \cup \bar{Z}_1^*$ is a set of saved vertices in an optimal solution for the firefighter problem on T . We will use the following lemmas.

Lemma 1.

$$\sum_{i=2}^k w(X_i) \geq w(\bar{Z}_1^*).$$

Proof. This lemma immediately follows from that \bar{T}_1 is constructed by contracting the roots of T_2, \dots, T_k . \square

Lemma 2.

$$\sum_{i=2}^k w(\bar{Z}_i^*) \geq (k-2)w(\bar{Z}_1^*).$$

Proof. For each $i \in \{1, \dots, k\}$, let X_i^- be the set of vertices of T_i except its root, i.e., $X_i^- = X_i \setminus \{v_i\}$. Notice that for each $i \in \{1, \dots, k\}$ the vertex set of \bar{T}_i consists of the vertex sets $X_1^-, \dots, X_{i-1}^-, X_{i+1}^-, \dots, X_k^-$ and the root obtained by contracting

$v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k$. For each $i \in \{2, \dots, k\}$, since the union of $\bar{Z}_1^* \cap X_j^-$ over $j \in \{2, \dots, k\} \setminus \{i\}$ is a feasible solution for the firefighter problem in \bar{T}_i ,

$$w(\bar{Z}_i^*) \geq \sum_{j \in \{2, \dots, k\} \setminus \{i\}}^k w(\bar{Z}_1^* \cap X_j^-). \quad (1)$$

Hence, by (1) and $\cup_{i=2}^k (\bar{Z}_1^* \cap X_i^-) = \bar{Z}_1^*$,

$$\begin{aligned} \sum_{i=2}^k w(\bar{Z}_i^*) &\geq \sum_{i=2}^k \sum_{j \in \{2, \dots, k\} \setminus \{i\}} w(\bar{Z}_1^* \cap X_j^-) \\ &\geq (k-2) \sum_{i=2}^k w(\bar{Z}_1^* \cap X_i^-) \\ &= (k-2)w(\bar{Z}_1^*). \end{aligned}$$

This completes the proof. □

For each real p with $0 \leq p \leq 1$,

$$\begin{aligned} &\max_{i \in \{1, \dots, k\}} (w(X_i) + w(\bar{Z}_i)) \\ &\geq \max_{i \in \{1, \dots, k\}} \left(w(X_i) + \left(1 - \frac{1}{e}\right) w(\bar{Z}_i^*) \right) \\ &\geq (1 - (k-1)p) \left(w(X_1) + \left(1 - \frac{1}{e}\right) w(\bar{Z}_1^*) \right) + \sum_{i=2}^k p \left(w(X_i) + \left(1 - \frac{1}{e}\right) w(\bar{Z}_i^*) \right) \\ &\geq (1 - (k-1)p) \left(w(X_1) + \left(1 - \frac{1}{e}\right) w(\bar{Z}_1^*) \right) + p \cdot w(\bar{Z}_1^*) + p \left(1 - \frac{1}{e}\right) (k-2) w(\bar{Z}_1^*) \\ &\geq (1 - (k-1)p) w(X_1) + \left(\left(1 - \frac{1}{e}\right) (1-p) + p \right) w(\bar{Z}_1^*), \end{aligned} \quad (2)$$

where the third inequality follows from Lemmas 1 and 2. By setting $p = \frac{1}{1+(k-1)e}$,

$$1 - (k-1)p = 1 - \frac{k-1}{1+(k-1)e}, \quad (3)$$

and

$$\begin{aligned} \left(1 - \frac{1}{e}\right) (1-p) + p &= \left(\frac{e-1}{e}\right) \left(\frac{(k-1)e}{1+(k-1)e}\right) + \frac{1}{1+(k-1)e} \\ &= \frac{1+(k-1)e - (k-1)}{1+(k-1)e} \\ &= 1 - \frac{k-1}{1+(k-1)e}. \end{aligned} \quad (4)$$

Hence, the following theorem follows from (2), (3) and (4).

Theorem 3. *Our algorithm is a polynomial time randomized $\left(1 - \frac{k-1}{1+(k-1)e}\right)$ -approximation algorithm for the firefighter problem on a tree in which the root vertex has k children.*

References

- Anshelevich, E., D. Chakrabarty, A. Hate, and C. Swamy. 2009. “Approximation Algorithms for the Firefighter Problem: Cuts over Time and Submodularity.” *Proc. the 20th International Symposium on Algorithms and Computation (ISAAC 2009)*. to appear.
- Cai, L., E. Verbin, and L. Yang. 2008. “Firefighting on Trees: $(1 - 1/e)$ -Approximation, Fixed Parameter Tractability and a Subexponential Algorithm.” *Proc. the 19th International Symposium on Algorithms and Computation (ISAAC 2008)*, Volume 5369 of *Lecture Notes in Computer Science*. Springer, 258–269.
- Finbow, S., A. King, G. MacGillivray, and R. Rizzi. 2007. “The Firefighter Problem for Graphs of Maximum Degree Three.” *Discrete Mathematics* 307 (16): 2094–2105.
- Hartnell, B., and Q. Li. 2000. “Firefighting on Trees: How Bad is the Greedy Algorithm?” *Congressus Numerantium* 145:187–192.

Approximation Algorithm for Multi-Dimensional Assignment Problem Arising from Multitarget Tracking

Yoshitaka Sugiura, Naoyuki Kamiyama, and Tomomi Matsui

Department of Information and System Engineering,
Faculty of Science and Engineering, Chuo University,
Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.

Abstract

In this paper, we discuss a multi-dimensional assignment problem arising from multitarget tracking. We propose a polynomial time 1.8-approximation algorithm.

Key words: multi-dimensional assignment problem, approximation algorithm

1 Introduction

Multiple target tracking is a subject devoted to the estimation of the trajectory of targets. The main problem in multiple target tracking is a data association problem of determining which sensor measurements emanate from which target. In this paper, we discuss a data association problem arising from multiple target tracking.

At time $t = 1$ a sensor is turned on to observe the region. At each time instance $t \in \{1, 2, \dots, k\}$, the sensor produces a report denoted by V_t . We assume that each report consists of measurements of n targets. The actual type of measurement varies with the sensor. For example, a 2-dimensional radar measures range and azimuth of each potential target, a 3-dimensional radar that measures range, azimuth, and elevation, a 3-dimensional radar with Doppler measures these and the time derivative of range. We have a set of reports $\{V_1, V_2, \dots, V_k\}$ such that each report consists of n measurements of targets without a knowledge that which measurement emanates from which target. The problem then are to determine which measurements go with which targets.

The formulation of a data association problem requires the specification of edges $\{i, j\}$ defined by a pair of measurements i and j . An edge $\{i, j\}$ represents an assignment between measurement i from a report V_t and measurement j from report $V_{t'}$ where $t \neq t'$. To each edge $e = \{i, j\}$ corresponds a weight w_e that represents the cost of the assignment. The edge weights are computed based on the dynamics of targets, which are generally modeled from physical laws of motion. In this paper, we assume that edge weights satisfy triangle inequalities. For each target, a subset of measurements corresponding to the target includes t measurements and meets every report in exactly one measurement. The data association problem is to find a partition of all the measurements such that each subset in the partition meets

every report in exactly one measurement and which minimizes the sum of weights of edges connecting pairs of measurements in a mutual subset. In the next section, we give a mathematical formulation of our problem as a multi-dimensional assignment problem.

2 Multi-dimensional Assignment Problem

Let $\mathcal{F} = \{V_1, V_2, \dots, V_k\}$ be a given set of reports. We introduce a *relation graph* \tilde{G} with a set of k vertices $\mathcal{F} = \{V_1, V_2, \dots, V_k\}$ and an edge set \tilde{E} . In this paper, we define the edge set \tilde{E} by

$$\tilde{E} = \{\{V_i, V_j\} \mid 1 \leq i < j \leq n, j - i \leq 2\}.$$

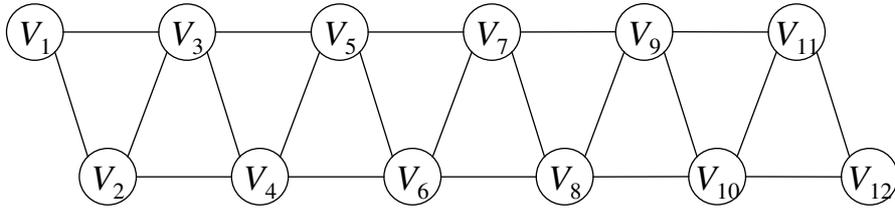


Figure 1: Relation graph \tilde{G} ($k = 12$).

We assume that each report $V_i \in \mathcal{F}$ consists of n vertices (measurements), i.e., $|V_1| = |V_2| = \dots = |V_k| = n$. A k -partite graph $G = (V_1, V_2, \dots, V_k; E)$ is defined by vertex sets V_1, V_2, \dots, V_k , and an edge set $E = \bigcup_{\{V_i, V_j\} \in \tilde{E}} \{u, v\} \mid u \in V_i, v \in V_j\}$. We introduce a non-negative edge weight $w : E \rightarrow \mathbb{Z}_+$ satisfying triangle inequalities.

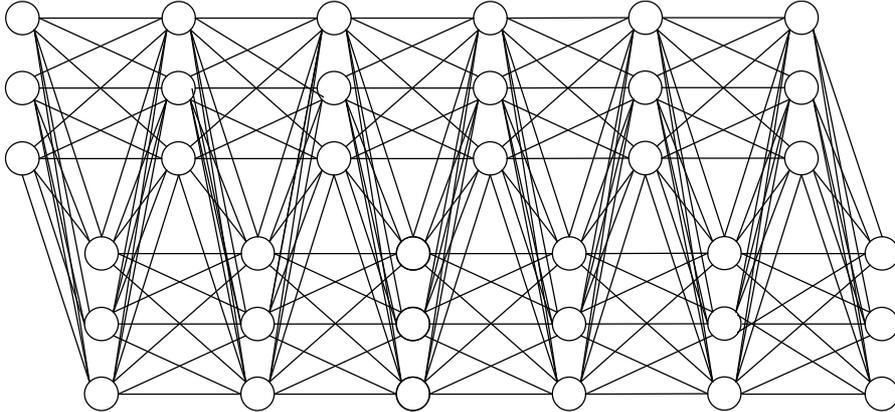


Figure 2: k -partite graph G ($k = 12, n = 3$).

We denote the vertex set $V_1 \cup V_2 \cup \dots \cup V_k$ by \hat{V} . For any vertex subset $Q \subseteq \hat{V}$, $G[Q]$ denotes a subgraph of G induced by Q and $w[Q]$ denotes the sum total of weights of edges in $G[Q]$. A vertex subset $Q \subseteq \hat{V}$ is called *feasible* if and only if Q meets every $V_i \in \mathcal{F}$ in exactly one vertex (i.e., $|Q \cap V_i| = 1$ for each $V_i \in \mathcal{F}$). When a family of feasible subsets $\mathcal{Q} = \{Q_1, Q_2, \dots, Q_n\}$ is a partition of \hat{V} (more precisely, $Q_1 \cup \dots \cup Q_n = \hat{V}$ and $Q_i \cap Q_j = \emptyset$ if $i \neq j$), we say that \mathcal{Q} is a *feasible partition*. In this paper, we discuss a *multi-dimensional assignment problem*

of finding a feasible partition $\mathcal{Q} = \{Q_1, Q_2, \dots, Q_n\}$ which minimizes the sum of weights $w[Q_1] + \dots + w[Q_n]$.

When $k = 2$, we have an ordinary (2-dimensional) assignment problem, which is efficiently solvable using the Hungarian method. When $k = 3$, Crama and Spieksma showed that the problem becomes NP-hard and proposed a simple $(4/3)$ -approximation algorithm (Crama and Spieksma 1992). When the relation graph is complete, Bandelt, Crama and Spieksma proposed a $(2 - 2/k)$ -approximation algorithm (Bandelt, Crama, and Spieksma 1994). In this paper, we propose a polynomial time 1.8-approximation problem for our problem.

3 Algorithm

First, we propose a heuristic algorithm which becomes a subprocedure in our algorithm. Let T be a spanning tree of the relation graph \tilde{G} . For each edge $e = \{V_i, V_j\}$ in T , we find a minimum weight perfect matching $M(e)$ in the bipartite induced subgraph $G[V_i \cup V_j]$ of G . We denote the sum of weights of edges in $M(e)$ by $w(M(e))$. We construct a graph $(\hat{V}, \cup_{e \in T} M(e))$ and decompose the vertex set \hat{V} into a family of connected components $\{Q_1, Q_2, \dots, Q_n\}$. We output $\{Q_1, Q_2, \dots, Q_n\}$, which is obviously a feasible partition. In the rest of this paper, we denote the heuristic algorithm described above by $\text{HT}(T)$ and the obtained feasible partition by $\mathcal{Q}(T)$. The sum of weights $w(Q_1) + \dots + w(Q_n)$ is denoted by $w(\mathcal{Q}(T))$.

We introduce an upper bound of $w(\mathcal{Q}(T))$. We construct a graph (with parallel edges) from the relation graph by substituting each edge $e \in \tilde{E} \setminus T$ for a unique path in T connecting end points of e . Denote the multiplicity of an edge e in the graph by $a_T(e)$. For any edge $e \in T$, a graph induced by $T \setminus \{e\}$ has two connected components and set of edges connecting them forms a cutset, which is denoted by $c(T, e)$. Then $a_T : \tilde{E} \rightarrow \mathbb{Z}_+$ satisfies

$$a_T(e) = \begin{cases} 0, & \text{if } e \in \tilde{E} \setminus T, \\ \text{size of cutset } c(T, e), & \text{if } e \in T. \end{cases}$$

Since the edge weights of G satisfies triangle inequalities, it is easy to show that $w(\mathcal{Q}(T)) \leq \sum_{e \in \tilde{E}} w(M(e))a_T(e)$ where $w(M(e))$ denotes the weight of perfect matching $M(e)$.

Now we introduce 4 spanning trees on \tilde{G} . In the rest of this paper, we assume that $\exists k' \in \mathbb{Z}$, $k = 6k'$. We can drop this assumption easily. Let $\tilde{E}^1 = \{\{V_1, V_2\}, \{V_2, V_3\}, \dots, \{V_{k-1}, V_k\}\}$ and $\tilde{E}^2 = \tilde{E} \setminus \tilde{E}^1$.

Spanning tree T_* : Let T_* be a spanning tree (Hamilton path) on \tilde{G} defined by \tilde{E}^1 . It is easy to see that

$$a_{T_*}(e) \leq \begin{cases} 3, & \text{if } e \in T_* = \tilde{E}^1, \\ 0, & \text{if } e \in \tilde{E} \setminus T_* = \tilde{E}^2. \end{cases}$$

Spanning tree T_i ($i = 0, 1, 2$): First, we introduce an infinite set

$$P_i = \bigcup_{j \in \mathbb{Z}} \left\{ \begin{array}{l} \{V_{2+6j+i}, V_{1+6j+i}\}, \{V_{1+6j+i}, V_{3+6j+i}\}, \{V_{3+6j+i}, V_{5+6j+i}\}, \\ \{V_{5+6j+i}, V_{4+6j+i}\}, \{V_{4+6j+i}, V_{6+6j+i}\}, \{V_{6+6j+i}, V_{8+6j+i}\} \end{array} \right\}$$

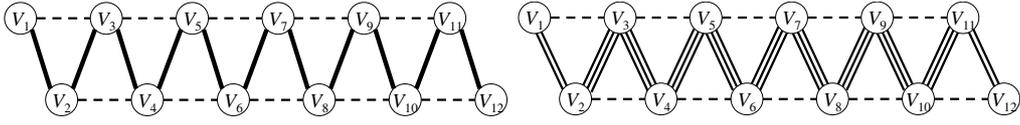


Figure 3: Spanning tree T_* and corresponding graph with parallel edges.

which forms a path in a graph with an infinite vertex set $\{V_i \mid i \in \mathbb{Z}\}$. We set $\widehat{P}_i = P_i \cap \widetilde{E}$. We define a spanning tree T_i ($i = 0, 1, 2$) by

$$T_i = \begin{cases} \widehat{P}_i, & \text{if } i = 0, 1 \\ \widehat{P}_i \cup \{\{V_1, V_2\}, \{V_{k-1}, V_k\}\}, & \text{if } i = 2. \end{cases}$$

It is easy to see that

$$a_{T_i}(e) \leq \begin{cases} 0, & \text{if } e \in \widetilde{E} \setminus T_i, \\ 3, & \text{if } e \in T_i \cap \widetilde{E}^2, \\ 5, & \text{if } e \in (T_i \cap \widetilde{E}^1) \setminus \{\{V_1, V_2\}, \{V_{k-1}, V_k\}\}, \\ 3, & \text{if } (e, i) = (\{V_1, V_2\}, 0) \text{ or } (\{V_{k-1}, V_k\}, 1), \\ 0, & \text{if } (e, i) = (\{V_1, V_2\}, 1) \text{ or } (\{V_{k-1}, V_k\}, 0), \\ 2, & \text{if } e \in \{\{V_1, V_2\}, \{V_{k-1}, V_k\}\} \text{ and } i = 2. \end{cases}$$

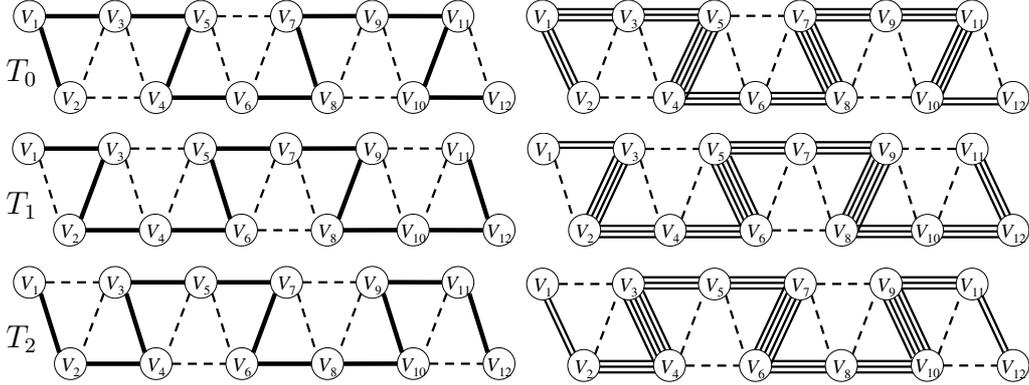


Figure 4: Spanning trees T_i ($i = 0, 1, 2$) and graphs with parallel edges.

Now, we describe our algorithm. For each spanning tree $T \in \{T_*, T_0, T_1, T_2\}$, we execute heuristic algorithm $\text{HT}(T)$ and obtain four feasible partitions $\mathcal{Q}(T_*)$, $\mathcal{Q}(T_0)$, $\mathcal{Q}(T_1)$ and $\mathcal{Q}(T_2)$. We output a feasible partition, denoted by \mathcal{Q}^* , in the set $\{\mathcal{Q}(T_*), \mathcal{Q}(T_0), \mathcal{Q}(T_1), \mathcal{Q}(T_2)\}$ which attains the value

$$\min\{w(\mathcal{Q}(T_*)), w(\mathcal{Q}(T_0)), w(\mathcal{Q}(T_1)), w(\mathcal{Q}(T_2))\}.$$

It is easy to see that the computational effort of our algorithm is bounded by a polynomial of the problem input size.

4 Approximation Ratio

In this section, we estimate the approximation ratio of our algorithm. Recall that for each edge $e = \{V_i, V_j\} \in \widetilde{E}$, $w(M(e))$ denotes the optimal value of minimum weight perfect matching problem defined on the bipartite induced subgraph $G[V_i \cup V_j]$.

Thus, it is obvious that $\sum_{e \in \tilde{E}} w(M(e))$ gives a lower bound of the optimal value of our multi-dimensional assignment problem. The feasible partition \mathcal{Q}^* obtained by our algorithm satisfies that

$$\begin{aligned} w(\mathcal{Q}^*) &= \min\{w(\mathcal{Q}(T_*)), w(\mathcal{Q}(T_0)), w(\mathcal{Q}(T_1)), w(\mathcal{Q}(T_2))\} \\ &\leq \left(\frac{1}{10}\right)w(\mathcal{Q}(T_*)) + \left(\frac{3}{10}\right)w(\mathcal{Q}(T_0)) + \left(\frac{3}{10}\right)w(\mathcal{Q}(T_1)) + \left(\frac{3}{10}\right)w(\mathcal{Q}(T_2)) \\ &\leq \sum_{e \in \tilde{E}} w(M(e)) \left(\left(\frac{1}{10}\right)a_{T_*}(e) + \left(\frac{3}{10}\right)a_{T_0}(e) + \left(\frac{3}{10}\right)a_{T_1}(e) + \left(\frac{3}{10}\right)a_{T_2}(e) \right). \end{aligned}$$

It is obvious that for each edge $e \in \tilde{E}^2$,

$$\left(\frac{1}{10}\right)a_{T_*}(e) + \left(\frac{3}{10}\right)a_{T_0}(e) + \left(\frac{3}{10}\right)a_{T_1}(e) + \left(\frac{3}{10}\right)a_{T_2}(e) \leq \left(\frac{1}{10}\right)0 + \left(\frac{3}{10}\right)(3 + 3 + 0) = 18/10.$$

When e is an edge in $\tilde{E}^1 \setminus \{\{V_1, V_2\}, \{V_{k-1}, V_k\}\}$, we can show that

$$\left(\frac{1}{10}\right)a_{T_*}(e) + \left(\frac{3}{10}\right)a_{T_0}(e) + \left(\frac{3}{10}\right)a_{T_1}(e) + \left(\frac{3}{10}\right)a_{T_2}(e) \leq \left(\frac{1}{10}\right)3 + \left(\frac{3}{10}\right)(5 + 0 + 0) = 18/10.$$

If e is either $\{V_1, V_2\}$ or $\{V_{k-1}, V_k\}$, we have that

$$\left(\frac{1}{10}\right)a_{T_*}(e) + \left(\frac{3}{10}\right)a_{T_0}(e) + \left(\frac{3}{10}\right)a_{T_1}(e) + \left(\frac{3}{10}\right)a_{T_2}(e) \leq \left(\frac{1}{10}\right)3 + \left(\frac{3}{10}\right)(2 + 3 + 0) = 18/10.$$

The above inequalities imply that

$$\begin{aligned} w(\mathcal{Q}^*) &\leq \sum_{e \in \tilde{E}} w(M(e)) \left(\left(\frac{1}{10}\right)a_{T_*}(e) + \left(\frac{3}{10}\right)a_{T_0}(e) + \left(\frac{3}{10}\right)a_{T_1}(e) + \left(\frac{3}{10}\right)a_{T_2}(e) \right) \\ &\leq \left(\frac{18}{10}\right) \sum_{e \in \tilde{E}} w(M(e)) \leq 1.8z^*, \end{aligned}$$

where z^* denotes the optimal value of our multi-dimensional assignment problem. Thus, our algorithm gives a polynomial time 1.8-approximation algorithm.

References

- Bandelt, H.-J., Y. Crama, and F. C. R. Spieksma. 1994. "Approximation algorithms for multi-dimensional assignment problems with decomposable costs." *Discrete Applied Mathematics* 49:25–50.
- Crama, Y., and F. C. R. Spieksma. 1992. "Approximation algorithms for three-dimensional assignment problems with triangle inequalities." *European Journal of Operational Research* 60:273–279.

Solving the Airline Crew Pairing Problem using Subsequence Generation

Matias Sevel Rasmussen^{*,1}, David M. Ryan², Richard M. Lusby¹,
and Jesper Larsen¹

¹Department of Management Engineering, Technical University of
Denmark, Denmark

²Department of Engineering Science, University of Auckland, New
Zealand

Abstract

Good and fast solutions to the airline crew pairing problem are highly interesting for the airline industry, as crew costs are the biggest expenditure after fuel for an airline. The crew pairing problem is typically modelled as a set partitioning problem and solved by column generation. However, the extremely large number of possible columns naturally has an impact on the solution time.

In this work in progress we severely limit the number of allowed subsequent flights, i.e. the subsequences, thereby significantly decreasing the number of possible columns. Set partitioning problems with limited subsequence counts are known to be easier to solve, resulting in a decrease in solution time.

The problem though, is that a small number of deep subsequences might be needed for an optimal or near-optimal solution and these might not have been included by the subsequence limitation. Therefore, we try to identify or generate such subsequences that potentially can improve the solution value.

1 Introduction

Crew costs are the second largest expenditure in the airline industry. Only fuel costs are larger, see [1]. Therefore, airline crew scheduling has received a lot of attention in the literature, and consequently, optimisation is heavily used by the airlines. The airline crew pairing problem which is dealt with in this work is a part of a larger series of optimisation problems that together produce the schedule for an individual crew member. In [1] a recent survey of airline crew scheduling can be found.

^{*}Corresponding author, mase@man.dtu.dk

A *pairing* or a *tour-of-duty* is a sequence of flights which can be flown by a crew member. A pairing must start and end at the same crew base and comply with several rules and regulations in order to be feasible. The *airline crew pairing problem* then finds the set of pairings that exactly covers all flights at minimum costs.

2 Solution Method

The pairing problem is modelled as a set partitioning problem. Each row corresponds to a flight and each column corresponds to a pairing. Let m be the number of rows and n be the number of columns, and let c_j be the costs of column $j \in \{1, \dots, n\}$. The entries of \mathbf{A} , a_{ij} , are one if column $j \in \{1, \dots, n\}$ covers row $i \in \{1, \dots, m\}$ and zero otherwise. The decision variables x_j for $j \in \{1, \dots, n\}$ are binary. The mathematical programme can be written as

$$\begin{array}{ll} \text{minimise} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{1} \\ & \mathbf{x} \in \{0, 1\}^n . \end{array}$$

The number of possible pairings in the set partitioning formulation is very large, so the pairings are typically only enumerated implicitly by column generation. In the present approach we will, however, not perform column generation, but *subsequence generation*.

The *subsequences* for a flight f are the set of subsequent flights that can follow f in a feasible way in a pairing. In general terms for a zero-one matrix \mathbf{A} , the *subsequence count*, $\text{SC}(s)$, for any row s is given by

$$\text{SC}(s) = |\{t : [a_{sj} = 1, a_{ij} = 0 \text{ for } s < i < t, a_{tj} = 1], j = 1, \dots, n\}| .$$

Matrices with $\text{SC}(s) \leq 1$ for all $s \in 1, \dots, m$ are said to have *unique subsequence*, and such matrices are balanced, see [2]. Exploiting results from graph theory, we know that the LP relaxation of an SPP with a balanced \mathbf{A} matrix has an integral optimal solution. Also shown in [2], the closer we get towards unique subsequence, the closer we get to naturally integral LP solutions.

Therefore, we severely limit the subsequence count for each flight when generating pairings. This results in significantly fewer possible pairings and, as mentioned, fewer fractions when solving the LP relaxation. The disadvantage, however, is that we might exclude some optimal subsequences. To remedy this, we use the information in the dual vector to identify missing subsequences. The dual vector is passed on to one or several column generators that produce negative reduced costs columns on a richer set of subsequences. These columns are analysed in order to identify potentially “good” subsequences.

The goal is, of course, to be able to, as early as possible, identify the subsequences that will end up in the optimal or near-optimal solution. Whenever a subsequence is identified as a potentially “good” subsequence, the whole set of columns which include the new subsequence are added to the LP. Furthermore, to prevent the LP from growing too big, subsequences can be removed from the LP, i.e. the set of columns containing the subsequence are removed.

3 Computational Results

When more developed, the method will be tested on real-world crew pairing problems. Meanwhile, in order to gain better understanding of the method, we have generated a set of set partitioning instances.

At the current stage the computational results are very preliminary, even on the generated instances. The results indicate that we can identify the missing subsequences in reasonable time.

4 Future Work

The results this far clearly justify further development. Firstly, as mentioned, real-world crew pairing problems will be tackled. Secondly, the subsequence identification process has room for a lot of improvement. Thirdly, the method is based on the dual vector, therefore dual stabilisation is likely to speed up the method, as dual stabilisation would make the duals more reliable. Lastly, the column generators can be run in parallel on different processors.

References

- [1] Balaji Gopalakrishnan and Ellis L. Johnson, *Airline crew scheduling: state-of-the-art*, Annals of Operations Research, 140 (2005), pp 305–337
- [2] D.M. Ryan and J.C. Falkner, *On the integer properties of scheduling set partitioning models*, European Journal of Operational Research, 35 (1988), pp 442–456

Customised Column Generation for Rostering Problems: Using Compile-time Customisation to create a Flexible C++ Engine for Staff Rostering

Andrew J Mason and David Ryan
Department of Engineering Science
School of Engineering
University of Auckland
New Zealand
a.mason@auckland.ac.nz

Anders Dohn
Department of Management Engineering
Technical University of Denmark

Abstract

This paper describes a new approach for easily creating customised staff rostering column generation programs. In previous work, we have built a large very flexible software system which is tailored at run time to meet the particular needs of a client. This system has proven to be very capable, but is difficult to maintain, and incurs the time penalties of run-time customisation. Our new approach is to customise the software at compile time, allowing compiler optimisations to be fully exploited to give faster code. The code has also proven to be easier to read and debug.

Keywords: Rostering, Column Generation

1 Introduction

For many years, staff in the Engineering Science department have been involved in developing rostering software for organisations such as NZ Customs, TabCorp and Air New Zealand (Mason, 1995; Mason, 2001). Each of these rostering problems has its own characteristics and requirements, and so has required customised software development. A long term goal has been to develop a flexible rostering engine that can be applied to rostering problems from a wide range of problem domains.

In 1995, Andrew Mason and Mark Smith, a masters student at the University of Auckland, developed a very fast column generation system in the Fortran programming language to solve a particular nurse rostering problem (Smith 1995; Mason and Smith, 1998). This system used a nested column generation approach (described below) to quickly construct entering columns for the underlying set partitioning problem. The speed of this system is perhaps best illustrated by the experimental finding that generating just one column at a time gave the best performance; we are unaware of any

other system for which is the case. Although very fast, this system was difficult to customise for different problems.

This project was followed by the work of PhD student David Nielsen, who in 2003 developed a software system whose capabilities were sufficiently general to solve a range of staffing problems (Nielsen, 2003). This system was successfully implemented by Mantrack for one of their clients, TabCorp. However, this system did not use column generation which made it most suitable for problems such as those with flexible part time staff where the rosters have little long term structure.

In 2002, Andrew Mason and masters student Faram Engineering developed a new C++ GENIE software system which generalised the key ideas in the earlier work (Engineer, 2003). This system was much more flexible, but this flexibility came at the cost of increased run times arising from a flexible, fully object oriented design. For example, the GENIE system must track ‘attributes’ which are stored in C++ classes. These classes, and their associated parameters, are generated when the software starts. This means we must incur all the overheads associated with calling arbitrary class types, and also that the compiler is very limited in the range of optimisations it can perform.

In 2009, Andrew Mason and Anders Dohn developed Genie⁺⁺, a new software framework in which the customer specific customisations are performed not at run time, but instead at compile time. This has the advantage that the problem is fully specified at compile time, and so the code optimisation techniques available in modern compilers can be fully utilised. As we will show, it also produces code that is easier to read and debug because it is written using the language of the specific rostering problem being solved.

2 Modelling Framework

We formulate the staff rostering problem (SRP) as a generalised set partitioning problem as follows. We assume there are s staff to be rostered, with these staff having different skills. Conceptually, we assume that each of these staff has n_i alternative rosters, termed *roster lines*, that they may work during the next rostering period. (A rostering period can range from one or two weeks up to five or six.) Each roster line consists of sequences of shifts separated by time off. (It is often convenient to assume a staff member works no more than one shift per day, but this is not required.) Because we formulate this problem with a minimisation objective, we assume each of these roster lines has some associated quality measure that measures the staff member’s dislike of that roster line.

To address the business needs, we assume that the requirement for staff can be specified by lower and upper bounds on the numbers of staff present across the day with different skill levels. For example, we may specify that we require “at least 5 staff of skill level 3 or higher between 10am and 2pm on Monday,” and “no more than 2 staff at skill level 6, between 10am and 2pm on Monday.” There is a set of p such requirements that define the work requirement for the roster. The j ’th roster line for person i is modelled as a column where the coefficient $a_{j[k]}^i$ defines how this roster line contributes to the k ’th work requirement. Thus, we can formulate our problem as a large generalised set partitioning problem, as follows:

$$\begin{aligned}
\text{SRP:} \quad & \text{minimise } \sum_{i=1}^s \sum_{j=1}^{n_i} c_j^i x_j^i \\
\text{s.t.} \quad & \sum_{j=1}^{n_i} x_j^i = 1 \quad \forall i=1, 2, \dots, s \quad (1) \\
\text{s.t.} \quad & \sum_{i=1}^s \sum_{j=1}^{n_i} \sum_{k=1}^p a_{j[k]}^i x_j^i \begin{cases} \geq \\ \leq \\ = \end{cases} b_k \quad \forall k=1, 2, \dots, p \quad (2) \\
& x_j^i \in \{0,1\} \quad \forall i=1, 2, \dots, s, j=1, 2, \dots, n_i \quad (3)
\end{aligned}$$

To complete the above formulation, we need to consider the construction of the columns \mathbf{a}_j^i for each staff member. These columns are generated during the solve process. The code to perform this generation forms the bulk of the Genie⁺⁺ system.

3. Column Generation

The goal of the column generator is to determine the best (or a set of good) entering columns during the linear programming and branch and bound steps of the solution process. This requires careful modelling of the quality (objective function coefficient) of the roster line, as well as any rules that define legal and illegal roster lines. For example, if we have morning “M” and night “N” shifts, then staff might prefer a sequence of four day shifts “MMMM” over a more disruptive sequence such as “DDNN”; we need to be able to reflect this in the objective function. Furthermore, union rules might make the sequence “DNNN” illegal because it contains 3 (or more) consecutive night shifts, and so such sequences need to be detected and banned during the column generation process. Examples of other quality and legality issues might include:

- Maximum number of days on in a row / week.
- Some combinations of x-on followed by y-off days prohibited.
- A minimum rest period after a shift is required.
- Specific shift transitions are not allowed.
- Split weekends (one day worked and one day off) are undesirable.
- Single days-on / days-off are undesirable.
- Staff cannot work two consecutive weekends
- Night shifts must occur in sequences of two or more consecutive night shifts.

The key concepts underpinning the column generator that enable us to efficiently model these quality and legality requirements are the ideas of *entities* and *attributes*. Figure 1 shows the relationship between these entities, while Table 1 describes the rules by which our construction scheme builds up entities by combining and extending other entities. The generator starts with *shifts*, each of which may have attributes such as “number of hours worked”, “is a weekend shift” and so on. The generator combines shifts together to give *on-stretches*, being sequences of days worked. For example, if we have morning “M” and night “N” shifts, then we might form an on-stretch “MMNN”. This on-stretch will also have attributes associated with it that are formed from simple operations on the underlying shift attributes, such as “number of hours worked” and “number of weekend shifts”. More complicated attributes can also be determined such as “number of days on” (in this case 4), and “number of day-to-night transitions” (1 in this case). The on-stretches are then combined with days off (*off-stretches*) to form *work-stretches*, which in turn are combined to form *roster-lines* that specify the sequence of activities for a staff member during the roster period.

on-stretch + shift \rightarrow on-stretch
 off-stretch + day-off \rightarrow off-stretch
 on-stretch + off-stretch \rightarrow work-stretch
 roster-line + work-stretch \rightarrow roster-line

Table 1: Construction rules for the rostering entities

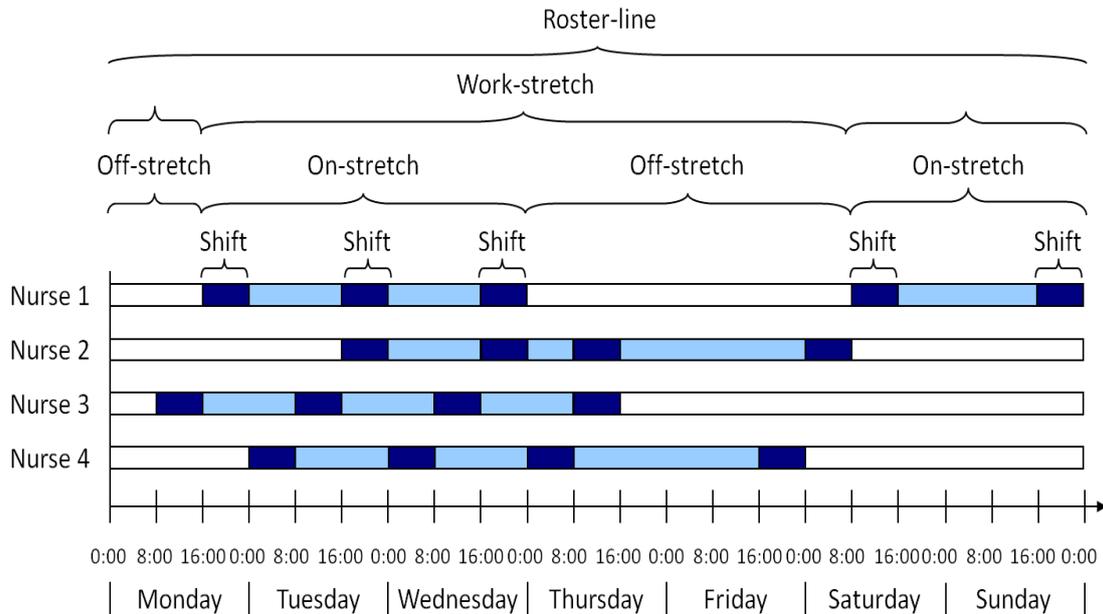


Figure 1: The entities that we use to describe a roster for a staff member.

The on-stretch, off-stretch, work-stretch and roster line entities are constructed during the column generation phase in a nested fashion using a dynamic programming approach. As in standard dynamic programming, only the best of any equivalent entities are kept. This gives us a nested column generation system that best exploits the special structure of our problem to efficiently solve the resource constrained shortest path column generation sub-problem.

Whenever a new entity is constructed using this process, the *attribute values* associated with that entity need to be calculated. These attribute values are then used to check legality and calculate quality values. The rules for calculating an attribute depend both on the entities being combined and the particular rules associated with that attribute. In the original C++ code, this multiplicity of possibilities was handled with objects and case statements, producing slow run times.

In our new design, the goal was to create customer-specific C++ code that would then be compiled to form that customer's solver engine. It was originally expected that a second program would be written to create this C++ code by interpreting some roster problem description language that we would have to create. However, after some initial experimentation, and thanks to recent advances in the C++ Boost Pre-processor Library (Karvonen and Mensonides, 2001), we realised that we could describe our rostering problems directly in the C++ pre-processor macro language. (This is the language that programmers use when writing statements such as

```
#define CURSOR(top, bottom) (((top) << 8) | (bottom)).
```

Although not a commonly used feature, C++ compilers allow the user to stop the compilation process after the pre-processor directives have been expanded, but before the code has been compiled. The output of this step is easy-to-read C++ code with variable and field names that match those from the real problem, making the code very easy to follow. The following code examples give an overview of the power of this system.

The code shown in Code Listing 1 demonstrates how the attributes are defined for a shift object. Notice that each line specifies the actual variable name (in lowercase), the variable type, and display name. The lowercase names will appear directly in the resulting C++ code.

```
// SHIFT_ATTRIBUTES must contain: starttime and endtime
# define SHIFT_ATTRIBUTES \
    ATT( (starttime      , int, "Starttime"), \
        ATT( (endtime    , int, "Endtime")  , \
            ATT( (shifttype , int, "ShiftType"), \
                ATT( (paidhours , int, "PaidHours"), \
                    ATT( (dayson   , int, "DaysOn")   , \
                        END ))))
```

Code Listing 1: Defining the attributes for a shift

The #define statement in Code Listing 1 is using a Boost pre-processor array structure to store the attribute definitions. These then get expanded by the pre-processor into a set of fields for the shift object. The code to perform this expansion, and the resulting C++ code that is generated, are shown below in Code Listing 2 and Code Listing 3. (Some code has been deleted to improve the clarity of this.)

```
class Shift {
public:
#       define   SATTR(_1, _2, i, elem)          STYPE(elem, i)
BOOST_PP_TUPLE_ELEM(3,0,elem);;
    BOOST_PP_LIST_FOR_EACH_I(SATTR, _, SHIFT_ATTRIBUTES);

```

Code Listing 2: An example of Boost pre-processor code used to expand a pre-processor array

```
class Shift {
public:
    Attribute<10, int, ... > starttime;
    Attribute<11, int, ... > endtime;
    Attribute<12, int, ... > shifttype;
    Attribute<13, int, ... > paidhours;
    Attribute<14, int, ... > dayson;
}

```

Code Listing 3: The code produced when the rostering definition in Code Listing 1 is expanded using the Boost pre-processor code in Code Listing 2.

As detailed earlier, an on-stretch is created by appending a shift (with the attributes given above) to another (possibly empty) on-stretch. The following code (Code Listing 4) illustrates how the attribute values are defined for the resulting on-stretch.

```
ATT( (paidhours, int, "paidhours", feas_all, domi_exact, cost_none,
o.paidhours + s.paidhours, s.paidhours) \

```

Code Listing 4: Definition of the 'paidhours' attribute in an on-stretch, including code for calculating the paidhours attribute value

Code Listing 4 defines a new on-stretch attribute with the name “paidhours.” This parameter is tracked during the column generation so that it can be checked in the final roster line against the target of 80 paid hours per fortnight for this staff member. The most important entries in `paidhours` definition are the last two which are actual C++ statements that will be compiled. The first of these, `o.paidhours + s.paidhours`, specifies that when an on-stretch ‘o’ and a shift ‘s’ are combined, the value for this attribute is calculated as the sum of the on-stretch’s paid hours (`o.paidhours`) and the shift’s paid hours (`s.paidhours`). The last entry handles the case when a blank on-stretch has a shift added to it.

The `o.paidhours + s.paidhours` definition gets expanded to give code such as the following (Code Listing 5):

```
T_value initialize (const Shift& s) const {
    return (o.paidhours + s.paidhours);
}
```

Code Listing 5: The code produced when the rostering definition in Code Listing 4 is expanded using the Boost pre-processor. This code is executed when a new on-stretch is formed by adding another shift to an existing on-stretch.

The other parameters in Code Listing 4 determine feasibility, dominance and cost contribution rules for `paidhours` as follows. The feasibility rule is used to discard an entity whenever it breaks some rostering rule; in this case ‘feas_all’ mean that all values are feasible. (This ‘rule’ is in fact the name of a C++ class.) Most rules, when required, can be expressed in terms of lower and upper bounds on the attribute value, and so another `feas_lbub` class is provided to handle this.

The `domi_exact` parameter details the rule for dominance. During the column generation, we can often determine that one entity dominates the other in the sense that any roster line containing the dominating entity will be better than one containing the dominated entity. (For example, an on-stretch “DND” might be dominated by “DDD” as the latter is perhaps equivalent from a rules point of view but better from a quality perspective as it avoids the day/night transitions.) The `domi_exact` term in this example says that two on-stretches must have the same attribute value if one is to be tested for dominance against the other.

Finally, the term `cost_none` describes how the value of this attribute is used to determine an associated cost value that contributes to the quality of the roster line. In this case, the paid hours of an on-stretch has no impact on the cost of a roster line. However, a commonly used option for this is to look up a table that translates the attribute value into a contribution to the objective. For example, an attribute might track the number of shift changes in an on-stretch, and penalise these in the objective (perhaps in some non-linear fashion) once each on-stretch is embedded within a work-stretch.

4. Results and Conclusions

The new GENIE⁺⁺ framework has been implemented in C++ using the branch-and-cut-and-price framework of COIN-OR (Lougee-Heimer, 2003). We use constraint branching (Ryan and Foster 1981) where branches assign a shift to a staff member; this branch is then enforced both within the masters and within the column generator for that person.

The original system developed by Faram Engineer was successfully tested on four problems including a nursing example from Middlemore Hospital, scheduling for the United States postal service, scheduling shift work at an Australian energy power plant, and rostering helicopter pilots. These tests demonstrated the ability of the system to accurately model the variety of rules found in these different examples. The new system follows the design approach of this original system, and so shares its capabilities to solve problems of this diverse nature.

The new code has proven to be much faster to both develop and debug. The new system uses the COIN-OR framework unlike the old which used the ZIP software developed by Professor David Ryan (Ryan 1980). Combined with the rapid advancement of hardware, these multiple changes make comparisons of run times difficult. However, one indication of the improvements achieved is the finding of a better solution for the Middlemore nurse rostering problem using the new system. Experiments with this Middlemore nursing instance show that on a typical desktop PC, we can prove optimality in 224 seconds (but note that this required a careful choice of branching order). A near optimal solution is found after 97 sec. The system spends 23 seconds in the root node. (The old code took 357 seconds (15.5 times longer) to solve the root node, and 1617 seconds (16.6 times longer) to find the first integer solution. Optimality was never proven.) The run times from another problem, the Rigshospitalet in Copenhagen, are not as promising, due to more flexibility in each subproblem. This flexibility means that each subproblem takes much longer to solve as there are many more possible roster lines to consider. We cannot prove optimality for this instance within 10 hours. However, we can find a solution within 1.4% of the lower bound after 15777 seconds (4 hours, 23 minutes). A better solution (gap 0.4%) is found after 23350 seconds (6 hours, 29 minutes). The root node takes 6169 seconds (1 hour, 43 minutes) to solve.

These rostering problems often have very large integrality gaps, and so, as shown in the previous results, the branch and bound process can be very time consuming. To reduce these potentially long run times, we are working on the development of heuristic techniques that can be embedded within both the column generation and the branch and bound processes. For problems such as the Rigshospitalet instance, the sub-problem is very flexible in the sense that there are many feasible columns, which should mean that heuristics can easily find good (but not necessarily optimal) entering columns, thereby significantly reducing the times to find near optimal solutions. We are confident that these improvements will, eventually, produce a system that is sufficiently flexible and reliable to meet the rigorous requirements of commercial application. We look forward to reporting results on this in the future.

Acknowledgements

The authors Andrew Mason and Anders Dohn wish to acknowledge the contribution of Professor David Ryan to this project. Not only did Professor Ryan inspire the two authors to look at complex rostering problems, but he also made it possible for Anders Dohn to visit Auckland in 2009.

References

Engineer, Faramroze G., 2003. *A solution approach to optimally solve the generalized rostering problem*, Masters thesis, University of Auckland, 2003

Karvonen, V., P. Mensonides. 2001. *Preprocessor metaprogramming. C++ library*. [Http://www.boost.org/](http://www.boost.org/) (Boost 1.36.0: 14/08/2008).

Lougee-Heimer, R., 2003. The Common Optimization INterface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development* 47(1) 57-66. [Http://www.coin-or.org/](http://www.coin-or.org/) (23/01/2009).

Mason, Andrew J., David M. Ryan, David M. Panton, 1998. Integrated Simulation, Heuristic and Optimisation Approaches to Staff Scheduling, *Operations Research*, Vol 46, Number 2, pp161-175

Mason, Andrew J., Mark C Smith, 1998. A Nested Column Generator for solving Rostering Problems with Integer Programming in *International Conference on Optimisation : Techniques and Applications*, L. Caccetta; K. L. Teo; P. F. Siew; Y. H. Leung; L. S. Jennings, and V. Rehbock (eds.), Curtin University of Technology, Perth, Australia, p827-834, April 1998

Mason, Andrew J., David Nielsen, 1999. PETRA : A Programmable Optimisation Engine and Toolbox for Personnel Rostering Applications presented at the *15th Triennial International Federation of Operational Research Societies (IFORS) Conference IFORS 99*, August 16-20, 1999, Beijing, China; available as School of Engineering Technical Report 593, University of Auckland

Mason, Andrew J., 2001. Elastic Constraint Branching, the Wedelin/Carmen Lagrangian Heuristic and Integer Programming for Personnel Scheduling" in *Annals of Operations Research*, 108(1), pp239-276

Nielsen, D., 2003. *A broad application optimisation-based rostering model*, PhD thesis, University of Auckland, 2003

Nielsen, D., Andrew Mason, 1998. Commercial development of a general application optimisation-based rostering engine, *Proceedings of the 33rd Annual Conference of the Operational Research Society of NZ*, pp10

Ryan, D. M. (1980) ZIP - A Zero-One Integer Programming Package for Scheduling, *Report C.S.S. 85, A.E.R.E.*, Harwell, Oxfordshire.

Ryan, D.M. & Foster, B.A. (1981). An integer programming approach to scheduling. In A. Wren (ed.), *Computer scheduling of public transport urban passenger vehicle and crew scheduling*, North Holland, Amsterdam, 1981, pp. 268-280.

Smith, Mark C., 1995. *Optimal nurse scheduling using column generation*, Masters thesis, University of Auckland

On Asset Reallocation in the New Zealand Electricity Market

Anthony Downward*, David Young, Golbon Zakeri

University of Auckland
New Zealand

Abstract

Recently two high-profile reviews of the New Zealand electricity market (NZEM) have been carried out. The first was an investigation by the Commerce Commission, which had at its heart a report by Professor Frank Wolak which found that there is a lack of competition in the NZEM, particularly during dry years (Wolak 2009). Secondly, a Ministerial Review into the electricity sector resulted in another report, which supported the Wolak Report's suggestions of improving competition by reallocating assets among the market participants and presented three possible reallocation scenarios (ETAG 2009).

In this work, we consider two types of asset reallocation: *swaps* and *divestiture*. In an asset swap, the ownership of a given pair of generation assets is swapped; whereas with asset divestment the ownership of a generation asset may be transferred to another firm in the market, or to a new entrant.

In our model, each firm may own *hydro* and/or *thermal* generation plants, with the objective being to maximize the total profit. Strategic behaviour of the firms is captured using a Cournot model with linear demand curves (see e.g. (Borenstein, Bushnell, and Stoft 2000)).

We first consider some simple examples of asset reallocation in a one-node Cournot setting. Here we find that the welfare benefits of any asset reallocation is very sensitive to the relative costs faced by the firms. This result is particularly relevant to markets with unpredictable hydro inflow patterns such as New Zealand.

We analyze an asset swap in the New Zealand context, whereby a firm in the North Island exchanges one of its thermal units for a hydro plant in the South Island. To model this, we construct a stylized two-node Cournot model loosely based on New Zealand. We find that this asset swap may not have the desired effect; specifically, the presence of transmission constraints on the line linking the islands (HVDC) may mean that the suggested swap will lead to a decrease in competition in some situations. In particular, a suggestion that works well in a dry year may not work well in a wet year.

Our results indicate that the capacity of the HVDC line is important in determining whether the asset swap is effective. If the HVDC is frequently congested then our model suggests the swap will improve welfare. If the HVDC is uncongested, we find that at best the swap makes little or no difference, and at worse can decrease welfare.

Key words: Cournot, electricity markets, asset reallocation, transmission.

References

- Borenstein, S., J. Bushnell, and S. Stoft. 2000. "The competitive effects of transmission capacity in a deregulated electricity industry." *RAND Journal of Economics* 31 (2): 294–325.
- ETAG, Electricity Technical Advisory Group. 2009. "Improving Electricity Market Performance." A preliminary report to the Ministerial Review of Electricity Market Performance.
- Wolak, F. 2009. "An Assessment of the Performance of the New Zealand Wholesale Electricity Market." Report for the New Zealand Commerce Commission.

Can Markets in Agricultural Discharge Permits be Competitive?

R. A. Ranga Prabodanie and John F. Raffensperger

Department of Management

University of Canterbury

New Zealand

r.ranathunga@mang.canterbury.ac.nz

Abstract

Agriculture is a major contributor to the problem of nitrate pollution in groundwater and surface water bodies. Tradable discharge permit programs have been proposed as a means of balancing the demand for nitrate intensive farming and the capacity of ecosystems to dilute the nitrates. These markets are usually designed at the catchment scale.

With limited geographic scope and several environmental constraints, can these markets be competitive enough to justify the cost of implementation? Similar nutrient trading programs in the United States record only a few trades and market failures.

We study the competitiveness of a market in nitrate discharge permits, assuming a centrally controlled market where users bid to buy discharge permits, and a regulator finds the equilibrium prices relative to the bids and the environmental constraints. The market competition is determined by buyer power, which is usually measured by buyer concentration. Using typical indices of buyer concentration, we investigate the competitiveness of these markets. The major purpose is to identify the essential and desirable conditions for proper functioning of a market in nitrate discharge permits. We also study what market designers could do to increase competitiveness.

Key words: tradable discharge permits, nitrate, market competitiveness.

1 Introduction

Agriculture is a major contributor to the problem of nitrate pollution in groundwater and surface water bodies. Intensive application of nitrogen fertilizer and livestock effluents on farm land creates a surplus of nitrogen in the soil which eventually leaches into the underlying groundwater aquifers as nitrate. Once mixed in groundwater, nitrate keeps on flowing with groundwater for decades, until being discharged into a surface water body.

Tradable discharge permit programs have been proposed as a solution to the problem of nitrate pollution. The United States has nutrient credit trading programs which allow trade in nitrate discharge permits (King and Kuch 2003; US EPA 2007). New Zealand's environmental institutions have proposed tradable nitrogen discharge permit programs as a means of controlling the intensive nitrate loading in agricultural catchments (Environment Waikato 2007; Lock 2008). These trading programs are usually designed at the catchment level.

Nitrate discharges from diffuse agricultural sources have spatial and temporal effects. Nitrate leaching from farms in a single time period may cause increases in nitrate levels at different locations in several different time periods. Therefore, trade in nitrate discharge permits is restricted by a number of hydro-geologically complex spatial and temporal constraints. A well-defined trading framework is required to facilitate trading. Under these circumstances, both market design and implementation would cost significantly. With a limited geographic scope and a number of resource constraints, can these markets be competitive enough to justify the cost of design and implementation?

Market competitiveness usually describes the rivalry among the firms. Research into market power and strategic behaviours in general environmental permit markets suggests that those issues are mainly related to the initial distribution of pollution rights (Egtern and Weber 1996). However, the competitiveness in agricultural discharge permit markets may be limited by factors other than the general measures of competitiveness such as market power. Even if the players were ideal competitors and did not behave strategically or exercise market power, the opportunities for trade in agricultural permit markets (mainly nutrient markets) would be mostly limited by the catchment hydro-geology. Discharge permits allocated to two distinct locations in the catchment are not usually comparable due to spatially varying hydro-geological properties. Non-comparable permits provide fewer opportunities for trade. Therefore, the competitiveness in these permit markets is determined by two types of factors: (1) participant characteristics (capabilities, strategic intents, current permit holdings) and (2) catchment hydro-geology.

In this paper, we look at the extent of competition determined by the catchment hydro-geology which is beyond the control of market designers and the players. In agricultural permit markets, the players (both buyers and sellers) are the farms (given that the discharge rights are fully distributed among farms) and the degree of competition is determined by the comparability of the permits they trade. Hence, for the purpose of this work, market competitiveness is loosely defined in terms of tradability as the extent to which the farms can participate in the market. This paper assumes that the only factor which restricts participation is the catchment hydro-geology, not the individual economic characteristics and capabilities.

We study the competitiveness of a market in nitrate discharge permits, assuming a centrally controlled trading system as proposed in Prabodanie et al (2009a). In the proposed trading system, farms submit bids and offers to trade nitrate permits and a market manager finds the equilibrium prices relative to the bids and the environmental constraints using a linear program. We propose two measures to describe the market competitiveness.

In the next section, we present an overview of the trading model. A comprehensive literature survey on pollution permit markets can be found in Prabodanie (2009a). In the third section, we show how to adapt and use a popular measure of buyer concentration to study the competitiveness in the agricultural discharge permit markets. We also present a new index to measure the overall competitiveness in a catchment scale market in nitrate discharge permits.

2 Trading model

To study a market, we first have to be clear about the commodities being traded. The maximum acceptable nitrate level (the nitrate dilution capacity) at each receptor in each

time period is a commodity (resource) being traded in the market. The tradable quantity is calculated after providing allowances for all non-tradable sources such as rain water and nitrates already in the aquifer. We define a “resource permit” for each single resource being traded. Hence, a resource permit is a right to increase the nitrate level at a specified receptor in a specified time period. In an ideal market, anyone should be able to trade resource permits, but to avoid confusion, the proposed trading system allows the farms to trade only “loading permits”, and a single entity (possibly a regional environmental authority or a farm organization) is allowed to buy and sell resource permits to reserve some resources for the future by actively participating in the market. We call this entity as the “resource bank.”

2.1 Loading permits

A farm would like to buy a single discharge permit rather than to buy each resource separately. Discharge permits are bundles of resources in different compositions. The trading system is designed so that the farms can buy and sell discharge permits in the same exchange market, even though the permits are not directly comparable among farms. We define nitrate discharge permits in terms of the maximum allowed nitrate leaching or loading into the groundwater aquifer underlying the farms. We call these as loading permits. Farms have to use some soil nitrogen model to determine the size of the permit required to cover their operations in each year. Then, they should submit bids and offers to the market manager who constructs a profit function for each farm from the bids/offers and the initial permit holdings. The profit functions are used in the objective function of the LP, which determines the optimal allocations and prices.

2.2 Market clearing LP

The market manager uses an LP to determine the optimal distribution of loading permits among the farms and a set of prices which result in the optimal distribution of permits. The LP maximises the social welfare relative to submitted bids and offers subject to the set of resource constraints, which require the farm to collectively meet the nitrate dilution capacity at each receptor in each time period over a long planning horizon.

The LP uses two sets of important environmental parameters: (1) a set of transport coefficients which relates the farm loading permits to the nitrate level at each receptor in each time period and (2) the maximum acceptable nitrate level (dilution capacities) at each receptor in each time period. In this paper, we do not explicitly present the LP used to clear the market. Instead, we present only a model of the capacity constraints which provides an overview of the hydro-geological impacts on market competitiveness. The complete LP model is available in Prabodanie et al (2009b).

2.3 Resource constraints

Indices

i = farm: $1, \dots, N$.

j = receptor: $1, \dots, M$.

t = monitoring year: $1, \dots, T$.

s = loading (permit) year: $1, \dots, S < T$.

d = delay in years: $0, \dots, T-1$.

Parameters

S_{jt} = maximum acceptable nitrate level at receptor j in time period t (after providing allowances for all non-tradable sources).

H_{ijd} = increase in nitrate level that occurs at receptor j after d year of unit (1 kg) nitrate loading in farm i during a single year, due to farms i 's action.

Decision variables

q_{is} = optimal nitrate loading allowed to farm i in year- s . This is the optimal allocation of the years- s loading permits to farm i .

q_{jt}^* = optimal resource position of the resource bank.

Model OptimalLoadingModel

Maximise [social welfare] subject to,

$$\sum_{i=1}^N \sum_{s=1}^{\min(S,t)} H_{ij(t-s+1)} q_{is} + q_{jt}^* \leq S_{jt} \quad \text{for } j = 1, \dots, M \text{ and } t = 1, \dots, T.$$

and other constraints.

Explanation

The constraints imply that the cumulative resource consumption should not exceed the available capacity. The first term in the left-hand-side indicates how much the farm loading permits consume each resource. Define resource $_{jt}$ as the nitrate dilution ability at receptor j in year t . A year- s loading permit of size q_{is} given to farm i is a bundle of resources consisting of $H_{ij(t-s+1)} q_{is}$ of each resource $_{jt}$.

3 Market competitiveness

The market competitiveness describes the extent to which the farms can trade with each other. If all farms are hydro-geologically identical, i.e., if H_{ijd} do not vary over i , farm loading permits are perfectly comparable and the competition is determined by individual farm characteristics such as current permit holdings and strategic intents. However, the situation in agricultural catchments is usually different; even if all the farms are in the same industry (for example, dairy) and have similar characteristics (for example, all are running at lowest cost), they are not hydro-geologically identical. Farm loading permits are not comparable, and therefore they have limited opportunities to trade. We look at the latter case and study the degree of competition determined by catchment hydro-geology. We ignore any effect on competition that may result from economic properties and behaviours of the individual players, including the resource bank and the availability of resources indicated by S_{jt} .

In the following discussion, we assume a single receptor J ($M=1$) case, but the models and measures are applicable to cases where $M>1$ with minor changes. We consider two cases: (1) the farms can trade only year-1 permits ($S=1$) and (2) the farms can simultaneously trade permits for several years ($S>1$). In the former case, if a farm \tilde{i} has $H_{\tilde{i},jd}>0$ for all $d\leq 4$, and $H_{\tilde{i},jd}=0$ for all $d>4$, while all other farms have $H_{i,jd}=0$ for all $d\leq 4$, then farm \tilde{i} is isolated and cannot participate in the market because it does not compete for the same resources as others. However, if farms can simultaneously trade year-1 to year-3 permits, \tilde{i} will compete with others for some resources (resource $_{j4}$ and resource $_{j5}$). Following this general understanding, we derive numerical indicators of competitiveness in the next sections.

3.1 Herfindahl-Hirschman Index (HHI)

The Herfindahl-Hirschman Index (HHI) is a measure of buyer/seller concentration and an indicator of competition among the buyers/sellers (Liston-Heyes, and Pilkington 2004). HHI is calculated from the sum of squared market shares of firms purchasing (selling) a particular product. Clearly, HHI is expressed relative to a single commodity.

To assemble a unit year-1 loading permit, farm \tilde{i} has to buy $H_{\tilde{i}J(t-1)}$ of resource $_{Jt}$. Farm \tilde{i} 's consumption rate of resource $_{Jt}$ is given by the transport coefficient $H_{\tilde{i}J(t-1)}$. Hence, the farm competition for resource $_{Jt}$ is determined by the distribution of $H_{iJ(t-1)}$ over i ($[t-1]^{\text{th}}$ row of the transport matrix for receptor J as shown in Figure 1). We may define farm \tilde{i} 's relative consumption rate of resource $_{Jt}$, $R_{\tilde{i}Jt}$ as a share of the cumulative consumption rate of all the farms.

$$R_{\tilde{i}Jt} = H_{\tilde{i}J(t-1)} / \sum_{i=1}^N H_{iJ(t-1)}.$$

Considering the relative consumption rate of each farm as a market share, we may now define a modified HHI, HHI^M for a market in resource $_{Jt}$, as the sum of squared relative consumption rates of the farms.

$$HHI^M_{Jt} = \sum_{i=1}^N R_{iJt}^2$$

For demonstration, the transport coefficient matrix for a hypothetical catchment with a single receptor (J), 10 farms ($i=1, \dots, 10$) and a 30 year planning horizon ($d=1, \dots, 29$) is given in Figure 1. The estimated competitiveness index HHI^M_{Jt} for each resource being traded (for each $t=d+1$) is given in the last column.

The values of HHI^M_{Jt} for $t=10, \dots, 13$ are around 0.1. With 10 firms, an HHI of 0.1 or below is usually considered as a competitive market (Ruster and Neumann 2006). Therefore, the results indicate that the market is competitive for some resources, but for some resources (for example $t=29$ and $t=30$), the market is dominated by a single player. However, the farms cannot buy resource $_{J29}$ or resource $_{J30}$ without buying resources from the other competitive markets because they can only trade loading permits. Hence, we suggest that even if the competition for few resources are significant relative to the number of farms in the catchment, the combined market would function actively.

$d \setminus i$	1	2	3	4	5	6	7	8	9	10	HHI ^M _{J(d+1)}
0	0.000	0.000	0.000	0.000	0.000	0.000	0.024	0.038	0.000	0.000	0.523
1	0.000	0.000	0.000	0.000	0.000	0.065	0.101	0.113	0.000	0.000	0.350
2	0.000	0.030	0.000	0.029	0.019	0.103	0.130	0.129	0.000	0.000	0.239
3	0.000	0.051	0.000	0.053	0.024	0.114	0.124	0.121	0.000	0.020	0.192
4	0.000	0.066	0.000	0.071	0.027	0.110	0.107	0.106	0.000	0.027	0.173
5	0.000	0.075	0.024	0.084	0.030	0.099	0.090	0.090	0.000	0.033	0.150
6	0.000	0.078	0.033	0.089	0.031	0.086	0.074	0.075	0.020	0.038	0.132
7	0.000	0.077	0.041	0.088	0.032	0.072	0.060	0.061	0.024	0.042	0.127
8	0.024	0.073	0.047	0.083	0.032	0.060	0.048	0.049	0.027	0.044	0.114
9	0.032	0.068	0.053	0.075	0.033	0.049	0.039	0.039	0.029	0.046	0.110
10	0.039	0.061	0.056	0.066	0.033	0.039	0.031	0.031	0.031	0.046	0.109
11	0.045	0.054	0.059	0.057	0.033	0.031	0.024	0.024	0.033	0.046	0.110
12	0.050	0.048	0.059	0.047	0.033	0.025	0.000	0.000	0.034	0.044	0.132
13	0.053	0.041	0.058	0.039	0.033	0.000	0.000	0.000	0.035	0.043	0.148
14	0.055	0.035	0.056	0.032	0.032	0.000	0.000	0.000	0.035	0.041	0.151
15	0.055	0.030	0.052	0.026	0.032	0.000	0.000	0.000	0.035	0.038	0.154
16	0.054	0.025	0.048	0.020	0.031	0.000	0.000	0.000	0.035	0.036	0.157
17	0.052	0.020	0.044	0.000	0.030	0.000	0.000	0.000	0.034	0.033	0.180
18	0.049	0.000	0.040	0.000	0.029	0.000	0.000	0.000	0.034	0.031	0.208
19	0.045	0.000	0.035	0.000	0.028	0.000	0.000	0.000	0.032	0.028	0.207
20	0.041	0.000	0.031	0.000	0.026	0.000	0.000	0.000	0.031	0.025	0.207
21	0.037	0.000	0.026	0.000	0.025	0.000	0.000	0.000	0.030	0.023	0.206
22	0.033	0.000	0.023	0.000	0.024	0.000	0.000	0.000	0.028	0.021	0.206
23	0.029	0.000	0.000	0.000	0.022	0.000	0.000	0.000	0.027	0.019	0.257
24	0.026	0.000	0.000	0.000	0.021	0.000	0.000	0.000	0.025	0.000	0.336
25	0.022	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.023	0.000	0.335
26	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.022	0.000	0.503
27	0.000	0.000	0.000	0.000	0.017	0.000	0.000	0.000	0.020	0.000	0.503
28	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.000	1.000
29	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000	1.000

Figure 1: Transport coefficient matrix (H_{ijd}) for receptor J .

The index derived above gives an approximate idea about the buyer concentration for each of the resource being traded in the combined market. However it does not give a clear picture of the overall market competitiveness. Therefore, we derive a new measure called the catchment tradability index, to give an idea about the overall market competitiveness.

3.2 Catchment tradability index (CTI)

Assuming $S=1$, we first define a tradability index for a pair of farms \tilde{i} and \tilde{j} to measure the extent to which they can interact in the market.

$$\text{Define } CTI_{(\tilde{i}, \tilde{j})} = \frac{1}{T} \sum_{t=1}^T (H_{\tilde{i}j_t} \times H_{\tilde{j}i_t})^{|H_{\tilde{i}j_t} - H_{\tilde{j}i_t}|},$$

where T is the number of years for which $H_{\tilde{i}j_t} > 0$ or $H_{\tilde{j}i_t} > 0$.

If either $H_{\tilde{i}j_t} = 0$ or $H_{\tilde{j}i_t} = 0$ for all t (if the columns corresponding to \tilde{i} and \tilde{j} in the transport coefficient matrix are not overlapping) $CTI_{(\tilde{i}, \tilde{j})} = 0$, and the pair of farms cannot trade at all. If $H_{\tilde{i}j_t} = H_{\tilde{j}i_t}$ for all t , (if the columns corresponding to \tilde{i} and \tilde{j} in the transport coefficient matrix coincide perfectly), $CTI_{(\tilde{i}, \tilde{j})} = 1$, the pair of farms can trade perfectly. Since $H_{ij_t} \leq 1$ for any i, j , and t , when the difference between $H_{\tilde{i}j_t}$ and $H_{\tilde{j}i_t}$ ($|H_{\tilde{i}j_t} - H_{\tilde{j}i_t}|$) increases, $CTI_{(\tilde{i}, \tilde{j})}$ tends to zero. When the difference between $H_{\tilde{i}j_t}$ and $H_{\tilde{j}i_t}$ decreases $CTI_{(\tilde{i}, \tilde{j})}$ tends to one. $CTI_{(\tilde{i}, \tilde{j})}$ can take a value within the range $[0, 1]$, it indicates the extent to which the pair of farms can interact in the market, and a greater the value of $CTI_{(\tilde{i}, \tilde{j})}$ is always preferable for a market.

From N farms in the market, we can extract ${}^N C_2 = (N-1)N/2$ distinct pairs. Therefore, the overall catchment tradability can be given by,

$$CTI = \frac{2}{(N-1)N} \sum_{(\tilde{i}, \tilde{j})} \frac{1}{T} \sum_{t=1}^T (H_{\tilde{i}j_t} \times H_{\tilde{j}i_t})^{|H_{\tilde{i}j_t} - H_{\tilde{j}i_t}|}$$

The table in Figure 2 lists the calculated values of pair-wise tradability index for the hypothetical catchment example in Figure 1. We see that the farms 7 and 8 which nearly coincide have the highest value of 0.984. The calculated overall catchment tradability index is 0.479. The value is in the mid of the two extremes of zero and one. Therefore, the market should be moderately competitive. The pair-wise $CTI_{(\tilde{i}, \tilde{j})}$ values in Table 2

may also be useful to understand the degree of market segmentation. Therefore, further study is required to better interpret the proposed index.

$i \backslash j$	2	3	4	5	6	7	8	9	10
1	0.359	0.668	0.318	0.633	0.177	0.139	0.139	0.701	0.637
2		0.539	0.912	0.532	0.555	0.459	0.460	0.368	0.604
3			0.484	0.633	0.302	0.257	0.256	0.619	0.811
4				0.485	0.585	0.477	0.479	0.327	0.5459
5					0.319	0.278	0.278	0.763	0.775
6						0.790	0.789	0.205	0.352
7							0.984	0.174	0.304
8								0.174	0.303
9									0.628

Figure 2: Tradability between pairs of farms ($CTI_{(i,j)}$).

3.3 Multiple year permit markets

When $S > 1$, farms can trade permits for several years simultaneously. Prabodanie et al (2009b) shows that multiple year permit markets provide more opportunities for trade compared to single year permit markets. Assuming that the farms would trade the same quantities of year-1 to year- S permits, we may define a cumulative transport coefficient, $H_{i,j,d}^S$ to measure the cumulative affect of a unit (1 kg) permit continuously valid from year-1 to year- S .

$$H_{i,j,d}^S = \sum_{s=1}^{\min(S,d+1)} H_{i,j,(d-s+1)}.$$

We can calculate HHI_{jt}^M for the case of $S > 1$, by replacing $H_{i,j,d}$ with $H_{i,j,d}^S$ in the formulation given above. To calculate CTI for this case, we have to replace $H_{i,j,d}$ with $H_{i,j,d}^S / \min(S,d+1)$ in the formulation of CTI . For the example given in Figure 1, the calculated catchment tradability index for $S=5$ is 0.580. This is a significant increase in the index compared to the value for the year-1 permit market. The competitiveness increases with S . This result is consistent with the results generated from a simulation of the market in Prabodanie et al (2009b).

4 Discussion and conclusions

In agricultural discharge permit markets, the competitiveness driven by individual farm characteristics and the structure of the market is manageable. Therefore, the competitiveness driven by hydro-geology worth to be studied before a market is implemented.

In this paper, we proposed some measures to evaluate the competitiveness of a market in nitrate discharge permits. We studied the competitiveness driven by hydro-geology, which is beyond the control of both the market participants and the designers. Our results suggest that the extent to which the farms can participate in the market depends on the extent to which their transport coefficient matrices overlap. If the farms are hydro-geologically isolated, tradable permits do little favour to the farms. However, if they plan far into the future and like to buy future permits, and the trading system allows them to buy permits for several future years, the farms would be able to trade with each other in a competitive market.

Even though the models were specifically designed for markets in nitrate discharge permits, the concepts are generally applicable to markets in diffuse ecological effect permits. Further research on applying the proposed measures would help the market designers to understand the nature of competition better (for example, presence of market segments).

5 References

- Egteren, H. V. and M. Weber. 1999. "Marketable Permits, Market Power, and Cheating." *Journal of Environmental Economics and Management* **30**:161-173.
- Environment Waikato. 2007. "Proposed Waikato Regional Plan Variation 5 - Lake Taupo Catchment." Hamilton.
- King, D. M. and P.J. Kuch. (2003). "Will Nutrient Credit Trading Ever Work? An Assessment of Supply and Demand Problems and Institutional Obstacles." *Environmental Law Reporter: News and Analysis*, 5-2003.
- Liston-Heyes, C. and A. Pilkington. 2004. "Incentive concentration in the production of green technology." *Science and Public Policy* **31**:15–25.
- Lock, K. and S. Kerr (2008). "Nutrient Trading in Lake Rotorua: Overview of a Prototype System." Motu Economic and Public Policy research.
- Prabodanie, R.A.R, J.F. Raffensperger, and M.W. Milke. 2009a. "A pollution offset system for trading non-point source water pollution permits." *Journal of Environmental and Resource Economics*. DOI:10.1007/s10640-009-9325-1.
- Prabodanie, R.A.R., J.F. Raffensperger, and M.W. Milke. 2009b. "Simulation-optimization approach for trading point and non-point source nutrient permits." 18th World IMACS / MODSIM Congress, Cairns, Australia.
- Ruster, S. and A. Neumann. 2006. "Economics of the LNG Value Chain and Corporate Strategies An Empirical Analysis of the Determinants of Vertical Integration." 26th USAEE International Conference, Michigan, U.S.
- US EPA. 2007. "Water Quality Trading toolkit for Permit Writers." Office of Wastewater Management. Us Environmental Protection agency,

A proposed smart market for impervious cover runoff under rainfall uncertainty

Antonio Pinto¹, John F. Raffensperger¹, Thomas Cochrane² and Shane Dye¹

(1) Department of Management

(2) Department of Civil and Natural Resources

University of Canterbury

New Zealand

Abstract

Damage caused from stormwater runoff is becoming more frequent and occurring in places that were previously free of these problems. In order to reduce this problem, authorities have looked at different mechanisms for controlling and minimising the costs of floods. This paper proposes a smart market (SM) for runoff from impervious cover to reduce the mitigation and prevention costs for damages in the catchment. As rainfall events are uncertain, the SM must incorporate this stochasticity into the model formulation. Two stage stochastic programming (TSSP) with recourse is proposed to deal with uncertainty of rainfall events. The recourse actions or penalties would be priced according to the cost of damage or mitigation at different places in the catchment. The market would allow hedging against a range of rainfall events until an established maximum rainfall event, and above it, penalties for more extreme events.

We expect the proposed SM would encourage users to internalize the expected costs of their runoff and cost of flooding. In addition, as the SM reduces transaction costs, there will be efficient allocation outcomes at minimum cost for society. The clearing prices and allocations would be based on auction bids, desirable environmental standards and the expected cost of flooding.

Key words: Smart market, impervious cover, runoff, flooding, stochastic programming.

1 Introduction

Society faces high costs due to environmental degradation from stormwater runoff (RO). In recent years, the occurrence of problems due to excess of stormwater and, therefore, flooding is getting worse. The frequency of these disasters has been a consequence of human activities, changes in land management and development in hazard areas. Rogers and Defee II (2005) observed that impacts on catchment outflow arose when development, impervious cover and edge density of roads increased. Those entailed more frequent flooding and threatened natural habitat in rural and urbanized areas (E.P.A. 1993; Strappazon et al. 2003; Walls and McConnell 2004; Eigenraam et al. 2005; Tang et al. 2005; Westra, Zimmerman and Vondracek 2005; Bradshaw et al. 2007; Hill, Pugh and Mullen 2007; Pappas et al. 2008).

Environmental problems have motivated governments and policymakers to create mechanisms and policies to achieve economic development with minimum environmental impact in order to avoid tragedy of the commons (Hardin 1968). Only a few studies have been set up to solve the problems related to storm-water runoff and floods through market-based instruments. One system with market instruments is transferable development rights (TDR), which has been tried to indirectly control flooding problems (McConnell, Walls and Kopits 2006; Walls and McConnell 2007). Other systems use fees and rebates, tradable runoff allowances (Thurston et al. 2003), flood management and risk analysis (Harman, Bramley and Funnell 2002; Purnell 2002; Sayers, Hall and Meadowcroft 2002; Ermolieva and Ermoliev 2005; Liu and Huang 2009), or merely command and control (Parker 1995).

Despite the theoretical plausibility of these methods, empirical evidence has shown problems with efficiencies, prices, allocations and, especially, transaction costs.

An alternative proposal to solve stormwater runoff problems, while reducing transaction costs, would be a smart market. A smart market (SM) is an auction system assisted by mathematical and computing tools (McCabe, Rassenti and Smith 1989; McCabe, Rassenti and Smith 1991), to manage complexities and third-party effects which are impossible to handle with ordinary auctions. The main difference between a smart market and an ordinary auction is that the former allows management of directly doing trading. The SM would reduce the transaction costs because users do not need to search for trading partners, bargaining is simpler, price information can be made available, and the manager ensures market discipline. Thus, theoretically, a SM would enable the attainment of efficient allocations and prices, as well as a greater surplus for society (McCabe, Rassenti and Smith 1989; McCabe, Rassenti and Smith 1991; Murphy et al. 2000; Gallien and Wein 2005; Raffensperger, Milke and Read 2008; Murphy et al. 2009; Raffensperger and Cochrane 2009).

Concerning the SM operation, Sayers et al., (2002) noted that dealing with flood management is quite complex and that any integrated management would necessarily need to be supported by a computer-based system. This statement reinforces the need for a smart market which is able to incorporate the consequences of land changes and trade in the catchment.

A deterministic smart market for impervious cover to control problem from excess of runoff was proposed by Raffensperger and Cochrane (2009). In this market, users trade consent to change impervious cover as measured by the Curve Number of the land while the authority limits a maximum capacity of storm water runoff at channel control points. Although this study introduces the idea of the SM, it does not consider the stochastic nature of rainfall. The current proposal tries to extend the research of Raffensperger and Cochrane to a SM that incorporates the stochasticity of the rainfall events.

Hydrological phenomena and, especially, rainfall events are complex and hard to predict. Calculations about rainfall are often based on simple averages, which creates problems when designing infrastructure and developing policy instruments. Ignoring the stochastic nature of rainfall events may lead to poor decisions and inefficient outcomes with significant social cost. Malcolm and Zenios (1994) suggest incorporating uncertainty into planning and design of infrastructure, to allow obtaining robust solutions and outcomes. That also is applicable to SM design.

The next section describes the smart market. The third section presents the model, and the fourth section shows how the SM would operate.

2 Smart market under rainfall uncertainty

Under uncertainty of rainfall events, the authority should make decisions about design, planning and environmental thresholds to minimise risk and keep society safe from flooding. The decision is influenced by rainfall distribution and the cost of extreme events. The regulator must deal with these concerns and should consider them in the design of the SM, to encourage hedging against a range of events.

A smart market under uncertainty works based on mathematical programming. This paper will use a two stage stochastic program (TSSP) with recourse. The main source of uncertainty will be the rainfall events.

The TSSP with recourse model is particularly useful as it allows working with infeasibilities due to extreme events (Li, Huang and Nie 2009). Tilmant et al. (2008) noted a stochastic model would be useful to help to hedge against extreme events such as drought and floods. Thus, the TSSP with recourse approach would model the randomness from different rainfall events and would consider violation of the constraints.

Stochastic formulations with recourse have been applied to different economic problems. For example, in electricity markets Carrion et al. (2007) proposed a stochastic program with recourse for solving an electricity supply problem of a large consumer. Calatrava and Garrido (2005) analysed different water market systems under uncertainty for water supply in Spain. Tilmant et al. (2008) presented a stochastic program for valuing the marginal water value in an integrated economic hydrologic model of a multipurpose multireservoir system for which the main activities requiring water were agriculture and hydroelectric production. Hollinshead and Lund (2006) minimized the expected cost of long-term, spot and option water purchases to meet environmental demands with a three stage stochastic formulation with recourse. The authors did not penalize their formulation against extreme drought corrections in dry years, which could result in infeasibilities.

The stochastic nature of the rainfall is incorporated in the SM model by relating the distribution of probabilities of rainfall events with the impact coefficients due to land uses, best management practices (BMPs) and technologies. Those impact coefficients correspond to the flow of runoff per unit of time (runoff hydrographs) at different points across the channel, and they vary according to the rainfall scenarios.

The SM for impervious cover maximizes the expected economic surplus for trading the level of impervious cover for land and accounting for recourse actions or penalties. The market considers the users' willingness to pay for impacting the channel's system and the corresponding environmental thresholds established by the authority.

Recourse actions and penalties are associated with flooding costs for extreme storms. Thus, the market would present incentives to reduce runoff and consequently to reduce damage due to flooding. Additionally, the health of a catchment would be hedged against a range of storms. Figure 1 illustrates a possible scenario for the rainfall distribution and flooding cost. The dashed line indicates the cost increase due to storms greater in intensity than the hedged range, E_d . Those costs of flooding would be considered in the model as penalties.

The penalties related to flooding cost are incorporated into the SM model. The mathematical formulation is presented in the next section.

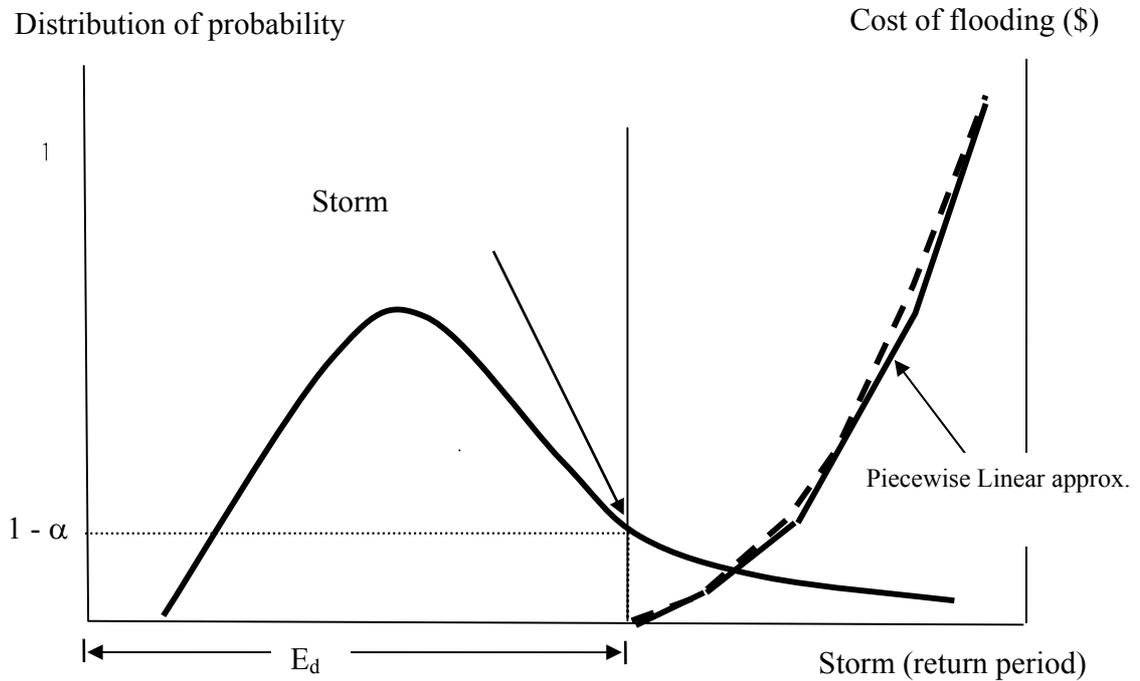


Figure 1 Rainfall event of probabilities distribution and flooding costs for extreme events. E_d represents the range of storms that the market would be hedged against and α represents the level of certainty.

3 Math models

Indices

- $i = 1, \dots, N$ user.
- $j = 1, \dots, J$ land type.
- $b = 1, \dots, B$ bid step.
- $k = 1, \dots, K$ control point.
- $r = 1, \dots, R$ range (interval) in the piecewise linear approximation.
- $s = 1, \dots, S$ scenarios of rainfall (storms).
- $t, u, r = 1, \dots, T$ time.

Parameters

- $C_{i,j}$ = Total initial land of type j owned by user i (ha).
- $D_{i,j,b}^{max}$ = Maximum amount of land type j (ha) that user i in the bid step b is willing to buy at price $P_{i,j,b}^D$.
- $S_{i,j,b}^{max}$ = Maximum amount of land type j (ha) that user i in bid step b is willing to sell at price $P_{i,j,b}^S$.

$P_{i,j,b}^D$ = Demand price (\$/ha) for land type j from user i and bid step b .

$P_{i,j,b}^S$ = Bid price (\$/ha) for land type j from user i and bid step b .

L_k^t = Maximum allowable runoff (m³/time) at channel the control point k , time t . These maximum capacities depend on the channel sectional shape at the control point k .

$P_{k,r}^f$ = Piecewise linear cost of flooding at control point k . This represents a linear cost under marginal changes in the total quantities in m³ at a place in range r . This m³/time cost will be incorporated as a penalty.

ϕ^s = Probability of storm scenario s . This parameter satisfies the following properties:
 $0 \leq \phi^s \leq 1$, $\sum_s \phi^s = 1$, $\phi^{s \cup s'} = \phi^s + \phi^{s'}$.

$H_{i,j,k}^{t-u+1,s}$ = Marginal impact of land type j from user i at control point k and scenario s , at the end of period $t - 1$ after one m³ of runoff was discharged from user i . This coefficient relates marginal impact at control points according to land type j and users' properties conditions and scenarios, e.g., m³/(time ha). u is the lag time between the discharged flow from a property and the flow that reaches the control point.

Decision variables

$qsell_{i,j,b}$ = Amount in hectares (ha) type j of runoff and bid steps $b = 1, \dots, B$ sold by user i .

$qbuy_{i,j,b}$ = Amount in hectares (ha) type j of runoff and bid steps $b = 1, \dots, B$ bought by user i .

$g_{i,j}$ = Total hectares type j for user i (ha).

$\mu_{i,j}$ = Expected land use price for firm i and land type j (\$/ha).

$\phi^s \lambda_{t,k}^s$ = Expected price to discharge at the control point k , time t and scenario s (\$ m³/time).

$f_{k,m}^s$ = Recourse action which accounts for the maximum level of flooding in scenario s at control point k in range r (m³/time).

Model SmartMarketIC

$$\text{Maximize } \sum_i^N \sum_j^J \sum_b^B P_{i,j,b}^D qbuy_{i,j,b} - \sum_i^N \sum_j^J \sum_b^B P_{i,j,b}^S qsell_{i,j,b} - \sum_s^S \phi^s \sum_k^K \sum_r^R P_{k,r}^f z_{k,r}^s$$

Subject to

- 1) $qbuy_{i,j,b} \leq D_{i,j,b}^{max}$: $\beta_{i,j,b}$. for all i,j,b
- 2) $qsell_{i,j,b} \leq S_{i,j,b}^{max}$: $\gamma_{i,j,b}$. for all i,j,b
- 3) $g_{i,j} = \sum_b^B qbuy_{i,j,b} - \sum_b^B qsell_{i,j,b} + C_{i,j}$: $\mu_{i,j}$ (free) for all i,j
- 4) $\sum_i^N \sum_j^J H_{i,j,k}^{t-u+1,s} g_{i,j} \leq L_k^t + \sum_r^R f_{k,r}^s$: $\phi^s \lambda_{t,k}^s$ for all t,k,s
- 5) $\sum_j^J \sum_b^B qbuy_{i,j,b} - \sum_j^J \sum_b^B qsell_{i,j,b} = 0$: v_i (free) for all i
- 6) $qbuy_{i,j,b}, qsell_{i,j,b}, f_{k,r}^s \geq 0$, and $g_{i,j}$ (free) : $\zeta_{i,j,b}, \delta_{i,j,b}, \theta_k^s$ for all i,j,b,s,k

Explanation

The objective function maximizes the expected total economic surplus from trading impervious cover-runoff less the penalties for flood damage under different rainfall events. This formulation considers a discrete distribution of probabilities for rainfall events and recourse actions for flooding damage at different control points. The model assumes that users bid truthfully.

- 1) Total expected area bought in each tranche or step is bounded by demand quantities.
- 2) Total expected area sold in each tranche is bounded by bid quantities.
- 3) The final amount of area of land type j of user i equals the net traded plus i 's initial allocation. The dual variable $\mu_{i,j}$ of this constraint is the expected marginal value for another unit of land type j for user i across the different rainfall scenarios according to the impact at each control points.
- 4) For each scenario s , the total runoff at control point k in period t should be less than the maximum capacity in the channel. The dual price $\phi^s \lambda_{t,k}^s$ represents the expected marginal social welfare if the authority allows another unit, e.g., m³/hr, at the control point k .
- 5) The total land sold for each user must be equal to the total land bought. This constraint ensures that all users will retain their initial land holding. For instance, if a user sold 0.25 hectares with 40% impervious cover, then she/he must buy 0.25 hectares with other imperviousness.

The model allows efficient allocation and provides prices for each participant, $\mu_{i,j}$. This price is a weighted value of impacts at different control points across all scenarios. It measures the expected improvement in social welfare due to the authority increasing the channel capacity by one unit, resulting in a corresponding reduction in the expected cost of flooding at the control point. The dual price decomposition of $\mu_{i,j}$ is as follows:

$$7) \mu_{i,j} = -\sum_s \phi^s \sum_k \sum_{t=u}^T H_{i,j,k}^{t-u+1,s} \lambda_{t,k}^s \quad \text{for all } i,j \quad : g_{i,j}$$

$$8) \phi^s \lambda_{t,k}^s + \theta_k^s = \phi^s P_k^f \quad \text{for all } k,s \quad : f_{k,r}^s$$

The price depends on the probable flooding costs $\phi^s P_k^f$ at the control point (equations 7 and 8). This expected cost will increase for each land use with greater impervious cover. Participants who want to increase their imperviousness would face a higher price due to the expected cost from extreme events. In other respects, if $\lambda_{t,k}^s > 0$, then by complementary slackness ($\theta_k^s = 0$) the new condition will be $\lambda_{t,k}^s = P_k^f$. Thus, if the constraint were bounded in only one time-period, the new condition would be $\lambda_{t,k}^s = P_k^f$, and so $\mu_{i,j} = -\sum_s \phi^s \sum_k \sum_{t=u}^T H_{i,j,k}^{t-u+1,s} P_k^f$.

The dual price $\mu_{i,j}$ will be used to charge or pay each user i in the catchment $\sum_j \mu_{i,j} \left(\sum_b^B q_{sell}{}_{i,j,b} - \sum_b^B q_{buy}{}_{i,j,b} \right)$ and, finally, after clearing the market, the regulator would receive a net payment (NP) for all transactions $NP = \sum_i \sum_j \mu_{i,j} \left(\sum_b^B q_{sell}{}_{i,j,b} - \sum_b^B q_{buy}{}_{i,j,b} \right)$. The auction is not necessarily revenue neutral.

4 How the market will work

The proposed SM would work as an auction system where the demand is represented by users that want to increase their impervious cover and the supply is defined by those persons that want to reduce their impervious cover. The SM considers tranches of sell offers and buy bids, to buy or sell impervious cover. As mentioned in Section 2, imperviousness is measured by an empirical parameter called the curve number (CN) which measures runoff from a rainfall event based on the land owner's hydrologic soil group, land use, management practice, and hydrologic conditions. Given the suitability of the curve number to represent the impervious cover, the SM would trade CN rights for a piece of land.

Concerning administration and control of the SM, a regional environmental authority would facilitate the CN trade. Trading will be done in a centrally-controlled auction where users will trade impervious cover quota (rights). These quotas may be for long- or short-term, but they cannot be issued for an indefinite time period.

Users may bid in advance for the impervious cover quota according to a timeframe which would depend on the hydrological seasons and the main economic activities. For instance, a catchment comprised mainly of farms would run with regard to the agricultural season, i.e., twice or four times per year; but in an urban catchment, the market may run monthly. In any case, the optimal timeframe is depends on catchment land uses and consequently, the timeframe should be evaluated constantly by the authority.

Because users may bids in advance, the impervious covers can be planned across a year or longer time frames. Those actions can also be linked to a contract which stipulates future physical actions on the land use regarding impervious cover or equivalent imperviousness at a specified time. Users who do not inherit impervious cover rights or those who want to develop new project can also participate in the spot market.

Outside the market, users may evaluate the cost of different options to control their own runoff, whilst satisfying the regulator that they would not change their initial runoff hydrograph. This could be quite expensive. The change of imperviousness and technology might increase the runoff from the property, raising penalties if a user does not comply with obligations. To simplify the calculation of impervious cover, the regulator could develop a web site where participants could estimate their land covers.

Concerning the role of the SM operator, the authority will first need to validate the property condition (impervious surface) for each participant in the catchment, and different methods could be used for this purpose. For instance, the manager could use a satellite-derived impervious map area to estimate the imperviousness of the area (Dougherty et al. 2004).

Secondly, individual impacts and runoff hydrographs need to be estimated in the catchment at control points by scenarios. The estimation can be done by hydrological and hydraulic models based on geographical information systems (GIS) such as the simulators HEC-HMS, HEC-RAS and SWMM. These models calculate the components of a hydrological cycle and are able to simulate runoff hydrographs in a routed channel or pipe system. In addition, the authority will estimate the distribution of rainfall events in the catchment. Thus, users' impacts can be measured as impact coefficients by time and by scenario.

Thirdly, the regulator must define the threshold at different control points. The threshold can be defined as maximum capacity of channels, pipes, and streams. With those thresholds, the authority can estimate the probable cost of flooding for extreme events. The costs would enable a more accurate assessment of the risk that the authority faces, while leaving the health of the catchment to be hedged against a range of events.

Fourthly, to accurately measure the impact coefficients, the regulator needs to monitor the individual impact coefficients along channels, especially in environmentally sensitive areas. These coefficients should be evaluated and controlled periodically to update the SM model.

Finally, the authority, with all previous issues, will clear the market and obtain prices and allocations.

5 Conclusion

In this paper, we have presented a smart market to manage problem of flooding. The market model considers a TSSP with recourse to incorporate the uncertainty of rainfall events. This market allows obtaining efficient allocation and prices while transaction cost are reduced. In addition, the authority could keep the health of the catchment within a range of storms.

This market does not account the risk of greater events, nor incorporate those in the market model. This topic will be further studied along our research, in particular the risks involved in the market design and how they would effect allocations and prices. In addition, the research will incorporate water quality into the market, where this quality will be measured as sediment discharge.

6 References

- Bradshaw, C. J., Sodhi, N. S., Peh, K. S., and Brook, B. W. (2007). "Global evidence that deforestation amplifies flood risk and severity in the developing world." *Global Change Biology*, 13(11), 2379-2395.
- Calatrava, J., and Garido, A. (2005). "Modelling water market under uncertainty water supply." *European Review of Agricultural Economics*, 32(2), 119-142.
- Carrion, M., Philpott, A. B., Conejo, A. J., and Arroyo, J. M. (2007). "A Stochastic Programming Approach to Electric Energy Procurement for Large Consumers." *Power Systems, IEEE Transactions on*, 22(2), 744-754.
- Dougherty, M., Dymond, R. L., Goetz, S. J., Jantz, C. A., and Goulet, N. (2004). "Evaluation of impervious surface estimates in a rapidly urbanizing watershed." *Photogrammetric Engineering and Remote Sensing*, 70(11), 1275-1284.
- E.P.A., E. P. A. (1993). *Handbook: Urban Runoff Pollution Prevention and Control Planning*, Diane Pub Co (March 1993), Cincinnati, OH. United States.
- Eigenraam, M., Beverly, C., Stoneham, G., and Todd, J. (2005). "Auctions for multiple environmental outcomes, from desk to field in Victoria, Australia." *Paper presented to the Annual Conference of the Western Economic Association, 4-8 July 2005, San Francisco, California.*, 30.
- Ermolieva, T., and Ermoliev, Y. (2005). "Catastrophic risk management: flood and seismic risk case study." Application of stochastic programming, S. W. Wallace and W. T. Ziemba, eds., Society for industrial and applied mathematics, SIAM, Philadelphia, USA, 709.

- Gallien, J., and Wein, L. M. (2005). "A smart market for industrial procurement with capacity constraints." *Management Science*, 51(1), 79-91.
- Hardin, G. (1968). "The Tragedy of the Commons." *Science*, 162(3859), 1243-1248.
- Harman, J., Bramley, M. E., and Funnell, M. (2002). "Sustainable flood defence in England and Wales." *Civil Engineering*, 150(5), 3-9.
- Hill, E., Pugh, S., and Mullen, J. (2007). "Use of the Hedonic Method to Estimate Lake Sedimentation Impacts on Property Values in Mountain Park and Roswell, GA." *Annual Meeting, July 29-August 1, 2007, Portland, Oregon*, 21.
- Hollinshead, S., and Lund, J. R. (2006). "Optimization of environmental water purchases with uncertainty." *Water Resources Research*, 42(W08403, doi:10.1029/2005WR004228), 10.
- Li, Y., Huang, G., and Nie, S. (2009). "Water Resources Management and Planning under Uncertainty: an Inexact Multistage Joint-Probabilistic Programming Method." *Water Resources Management*, 23(12), 2515-2538.
- Liu, Z., and Huang, G. (2009). "Dual-Interval Two-Stage Optimization for Flood Management and Risk Analyses." *Water Resources Management*, 23(11), 2141-2162.
- Malcolm, S. A., and Zenios, S. A. (1994). "Robust Optimization for Power Systems Capacity Expansion under Uncertainty." *The Journal of the Operational Research Society*, 45(9), 1040-1049.
- McCabe, K. A., Rassenti, S. J., and Smith, V. L. (1989). "Designing Smart computer-assisted markets." *Journal of political economy*, 5, 259-283.
- McCabe, K. A., Rassenti, S. J., and Smith, V. L. (1991). "Smart Computer-Assisted Markets." *Science*, 254(5031), 534-538.
- McConnell, V., Walls, M., and Kopits, E. (2006). "Zoning, TDRs and the density of development." *Journal of Urban Economics*, 59(3), 440-457.
- Murphy, J. J., Dinar, A., Howitt, R. E., Rassenti, S. J., and Smith, V. L. (2000). "The Design of "Smart" Water Market Institutions Using Laboratory Experiments." *Environmental and Resource Economics*, 17(4), 375.
- Murphy, J. J., Dinar, A., Howitt, R. E., Rassenti, S. J., Smith, V. L., and Weinberg, M. (2009). "The design of water markets when instream flows have value." *Journal of Environmental Management*, 90(2), 1089-1096.
- Pappas, E. A., Smith, D. R., Huang, C., Shuster, W. D., and Bonta, J. V. (2008). "Impervious surface impacts to runoff and sediment discharge under laboratory rainfall simulation." *CATENA*, 72(1), 146-152.
- Parker, D. J. (1995). "Floodplain development policy in England and Wales." *Applied Geography*, 15(4), 341-363.
- Purnell, R. (2002). "Flood risk: A government perspective." *Civil Engineering*, 150(5), 10-14.
- Raffensperger, J., Milke, M., and Read, E. G. (2008). "A deterministic smart market model for ground water." *Forthcoming in Operation Research*.
- Raffensperger, J. F., and Cochrane, T. (2009). "A smart market for impervious cover." *Working Document. University of Canterbury*.
- Rogers, G. O., and DeFee II, B. B. (2005). "Long-term impact of development on a watershed: early indicators of future problems." *Landscape and Urban Planning*, 73, 215-233.
- Sayers, P. B., Hall, J. W., and Meadowcroft, I. C. (2002). "Towards risk-based flood hazard management in the UK." *Civil Engineering*, 150(5), 36-42.

- Strappazon, L., Ha, A., Eigenraam, M., Duke, C., and Stoneham, G. (2003). "Efficiency of alternative property right allocations when farmers produce multiple environmental goods under the condition of economies of scope." *Australian Journal of Agricultural & Resource Economics*, 47(1), 1-27.
- Tang, Z., Engel, B. A., Pijanowski, B. C., and Lim, K. J. (2005). "Forecasting land use change and its environmental impact at a watershed scale." *Journal of Environmental Management*, 76(1), 35-45.
- Thurston, H. W., Goddard, H. C., Szlag, D., and Lemberg, B. (2003). "Controlling Storm-Water Runoff with Tradable Allowances for Impervious Surfaces." *Journal of Water Resources Planning and Management*, 129(5), 409-418.
- Tilmant, A., Pinte, D., and Goor, Q. (2008). "Assessing marginal water values in multipurpose multireservoir system via stochastic programming." *Water Resources Research*, 44(W12341, doi:10.1029/2008WR007024), 17.
- Walls, M., and McConnell, V. (2004). "Incentive-Based Land Use Policies and Water Quality in the Chesapeake Bay." Resources For the Future.
- Walls, M., and McConnell, V. (2007). *Transfer of Development: Rights in U.S. Communities*, Resources for the Future, Washington D.C.
- Westra, J. V., Zimmerman, J. K. H., and Vondracek, B. (2005). "Bioeconomic analysis of selected conservation practices on soil erosion and freshwater fisheries." *Journal of the American Water Resources Association*, 41(2), 309.

Shaping More Sustainable Communities: a Case Study in Urban Water Management

Robyn M. Moore
Victoria Management School
Victoria University of Wellington
New Zealand
robyn@j.co.nz

Extended Abstract

Purpose

Urban water systems, in particular, are under increasing pressure to meet the expectations of communities, with water managers required to articulate sensible and sustainable management initiatives that will secure water supplies and protect water for its intended use, now and in the future. Despite policy and regulation intended to advance outcomes and integrate efforts within the complex area of urban water management, fragmented approaches persist, while a pattern of decline in the quality of New Zealand's urban water resources remains a cause for concern. Nearly half of urban rates in New Zealand apply to water and wastewater management. Thus, this study is concerned with increasing awareness of the critical constraints to achieving healthier, more sustainable systems that are affordable for New Zealand communities. The specific challenges facing a community pursuing sustainable urban water management objectives are examined and solutions sought and tested.

Design/methodology/approach

Subsequent to a piloted investigation, a methodological framework was proposed, based on integrating three complementary perspectives. The Theory of Constraints (TOC) was used with a Stakeholder Typology to identify 'typical' and 'atypical' system stakeholders and examine their perspectives, while Causal Loop Diagrams (CLDs) from Systems Dynamics were constructed with participants to explore and circumvent potential negative outcomes. Thus, a case study in a community resource management setting is described that tests the value of the combined framework.

Findings

The combined framework provided a source of deep insights into the challenges, dilemmas, potential solutions and side effects facing resource managers and other stakeholders in an urban water system under pressure from population growth and climatic/topographical conditions. It is possible that the combined theoretical framework can be applied to other resource management cases. The use of the Stakeholder Typology to complement TOC provided a tactical element not routinely evident in systems studies, valuing the experiential and historical perspectives of those who might otherwise be treated as being outside the system, their perspectives marginalised or ignored. Solutions that were sought and tested using TOC and CLDs have been put into practice and are driving actions and dialogue that to date, appear to be delivering positive change for the community and other stakeholders (see Moore 2009: Appendix 10).

Research limitations/implications

The present study provides a starting-point for further research combining TOC with a stakeholder engagement methodology in the resource management sector. One perceived limitation is that once the TOC practitioner disengages from the research, this leaves stakeholder insights to be shared with other stakeholders in a potentially ad hoc manner; if indeed they are shared at all, limiting ongoing improvement. Training an in-house TOC practitioner would help to resolve this. To a limited extent, this has occurred in this instance, with a Kapiti Coast District Council (KCDC) Water Project Manager receiving guidance in IO mapping from the researcher and having access to the full thesis. Following the Kapiti case as it progresses, will reveal further study limitations.

Originality/value

The combined TOC, CLD, and Stakeholder Typology framework has proven of value in seeking and testing a number of solutions to the long standing problem of water insecurity on the Kapiti Coast. In particular, the Kapiti Coast District Council has adopted a Water Communications Strategy and a stakeholder engagement process. These are necessary conditions for a more sustainable urban water system, according to the IO maps and CLDs prepared with Councillors and other participant stakeholders. That the thesis played some part in informing actions – with the researcher consulted to review KCDC’s Water Communications Strategy (in September 2009) – is a notable and promising outcome of the study, from a resource management – and also a personal – perspective.

Key words: Sustainable urban water systems, Theory of constraints, urban water management, Stakeholder typology, Decision making, Case study.

1 Introduction

The motivation for this study was to consider how communities might take a more integrated and systematic approach to meeting the challenges of water management in New Zealand, and achieve more sustainable systems. Urban water systems, in particular, are under increasing pressure to meet the expectations of communities, with water managers required to articulate sensible and sustainable management initiatives that will secure water supplies and protect water for its intended use, now and in the future. Despite policy and regulation intended to advance outcomes and integrate efforts within the complex area of urban water management, fragmented approaches persist, while a pattern of decline in the quality of New Zealand’s urban water resources remains a cause for concern. Nearly half of urban rates in New Zealand apply to water and wastewater management. Thus, this study is concerned with increasing awareness of the critical constraints to achieving healthier, more sustainable systems that are affordable for New Zealand communities. It tests the use of the Theory of Constraints (TOC) systems framework and a Stakeholder Typology to examine ways that communities might gain better outcomes from their investment in urban water management initiatives. The thesis demonstrates the methodology by focusing on Kapiti, a settlement north of Wellington, which has been debating and responding to water quality and security issues for more than a decade.

2 Methodological Approach

2.1 Ethical Considerations

Early in the investigation, the researcher applied to the University Human Ethics Committee for leave to request participants to be named. All participants subsequently agreed to their comments being attributed to their name, though not all wished to attribute their comments to an organisation they were affiliated with. Identifying participants is a departure from the generally accepted procedure in qualitative studies. However, the sharing of knowledge and perspectives in a systematic and transparent manner (the researcher adopted the term ‘thinking out loud’) is in keeping with the intention of the research design: to foster a supportive environment for stakeholder engagement, revealing deep insights and critical understandings by encouraging stakeholder participants to share their ‘thinking out loud’.

2.2 Combining The Theory of Constraints and Stakeholder Typology

A brief Pilot Study conducted by the researcher revealed that the selection of participants with a stake in the system under investigation might be assisted by applying stakeholder analysis. Stakeholder mapping (Elias, Cavana and Jackson, 2002; Freeman, 1984), and Mitchell’s (1997) Stakeholder typology, informed the initial participant selection process, while an award-winning paper on stakeholder analysis in Public Relations by Rawlins (2006) was discovered during a later literature search. This led to a further stakeholder group (the Starorough Flaxbourne Conservation Project from Marlborough) being identified and included in the study. Specific stakeholders linked with the strategic issue were identified according to the ten categories that appear in the figure below. Note the two directional arrows, illustrating the nature of the relationship between the stakeholder and the system issue.

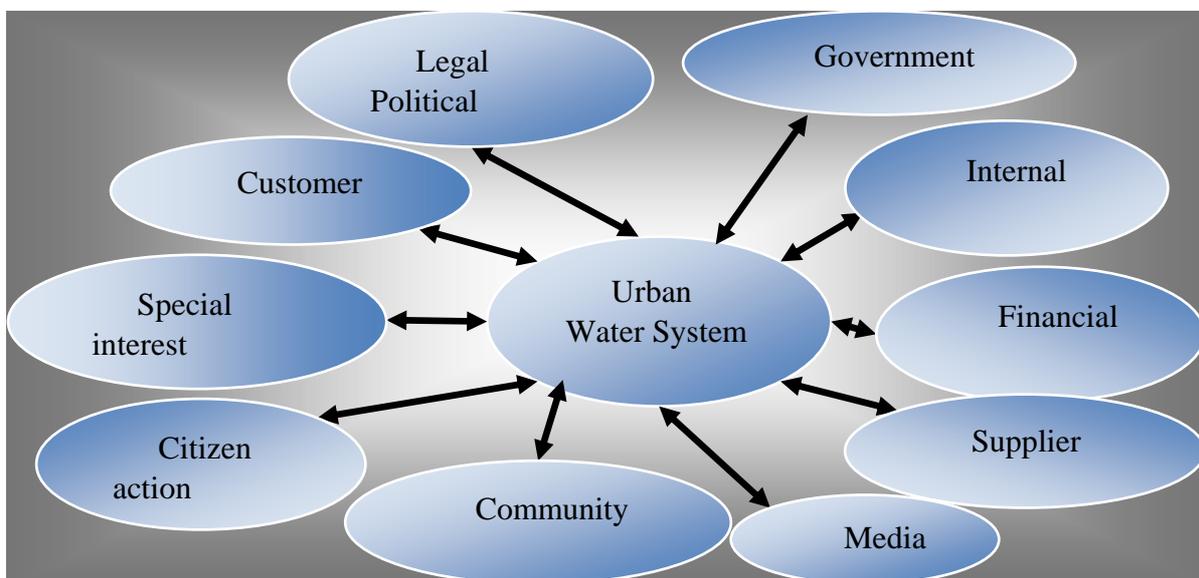


Figure 1. The Stakeholder map

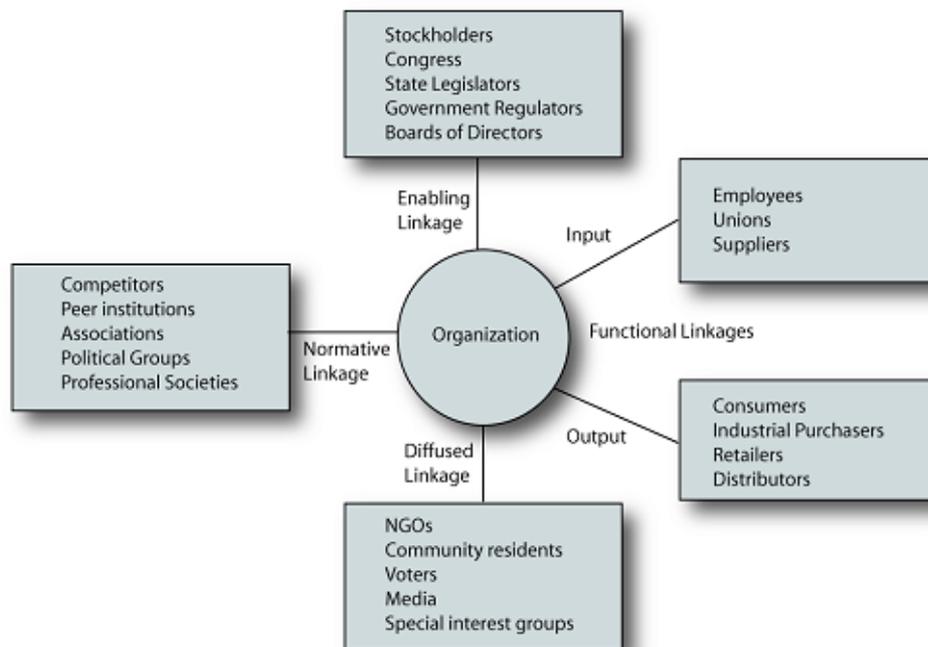


Figure 2. Stakeholder linkage model (Rawlins, 2006:4)

Following the mapping of participants using the generic stakeholder map (Freeman, 1984; Elias et al., 2002), emphasis turned to the linkages (Rawlins, 2006), or the connections between the system and the participant, with care taken to ensure the participant stakeholders were represented in each of the linkages.

The justification for the selection of the Starborough Flaxbourne participants is their situational linkage to the system problem, described in Rawlins (2006) as a ‘normative’ linkage. The ‘community’ and ‘consumers’, comprising around a third of the participants on the stakeholder grid (see Elias et al., 2002), have either a diffused or a functional (output) linkage. However, a normative linkage is also possible, given that a high proportion of the community members/consumers interviewed appeared to share common concerns and goals for their urban water system. KCDCs role in the thesis stems from a functional (input) linkage, while the Ministry of Health (MOH), Greater Wellington Regional Council (GWRC) and to an extent, the Department of Building and Housing (DBH), demonstrate enabling linkages.

By engaging with the participant stakeholders identified through the Stakeholder Typology, and using the Thinking Processes from the TOC methodology, the problems with the most undesirable effects on the system were identified. Solutions were sought and tested using Intermediate Objective (IO), Current Reality Tree (CRT/B) and Prerequisite Tree (PRT) mapping procedures, coupled with Evaporating Cloud conflict resolution diagrams (ECs) from TOC, together with Causal Loop Diagrams (CLDs) from Systems Dynamics. The last in a series of IO maps and the CRB are shown.

2.3 Reaching the Destination - with TOC IO Maps

The study began with an idea to agree a ‘clear, unequivocal goal statement’ (Dettmer, 2007) among participants. The vehicle for this is the IO or Destination map, which Dettmer argues is critical to the success of the Thinking Processes. The IO map ‘fixes a firm baseline in space and time’ (Dettmer, 2007: 68), with the researcher finding it necessary to change/refine the IO at various stages of the research, as certain

dynamics (plan changes for example) suggested that some IOs be reconsidered. IOs are connected in a logical hierarchy leading to the system goal. Applying knowledge of what *is* happening and what *should* be happening identifies gaps and determines the actions needed as part of systemic change. Read the IO map top down: *In order to... we must (ensure)...*

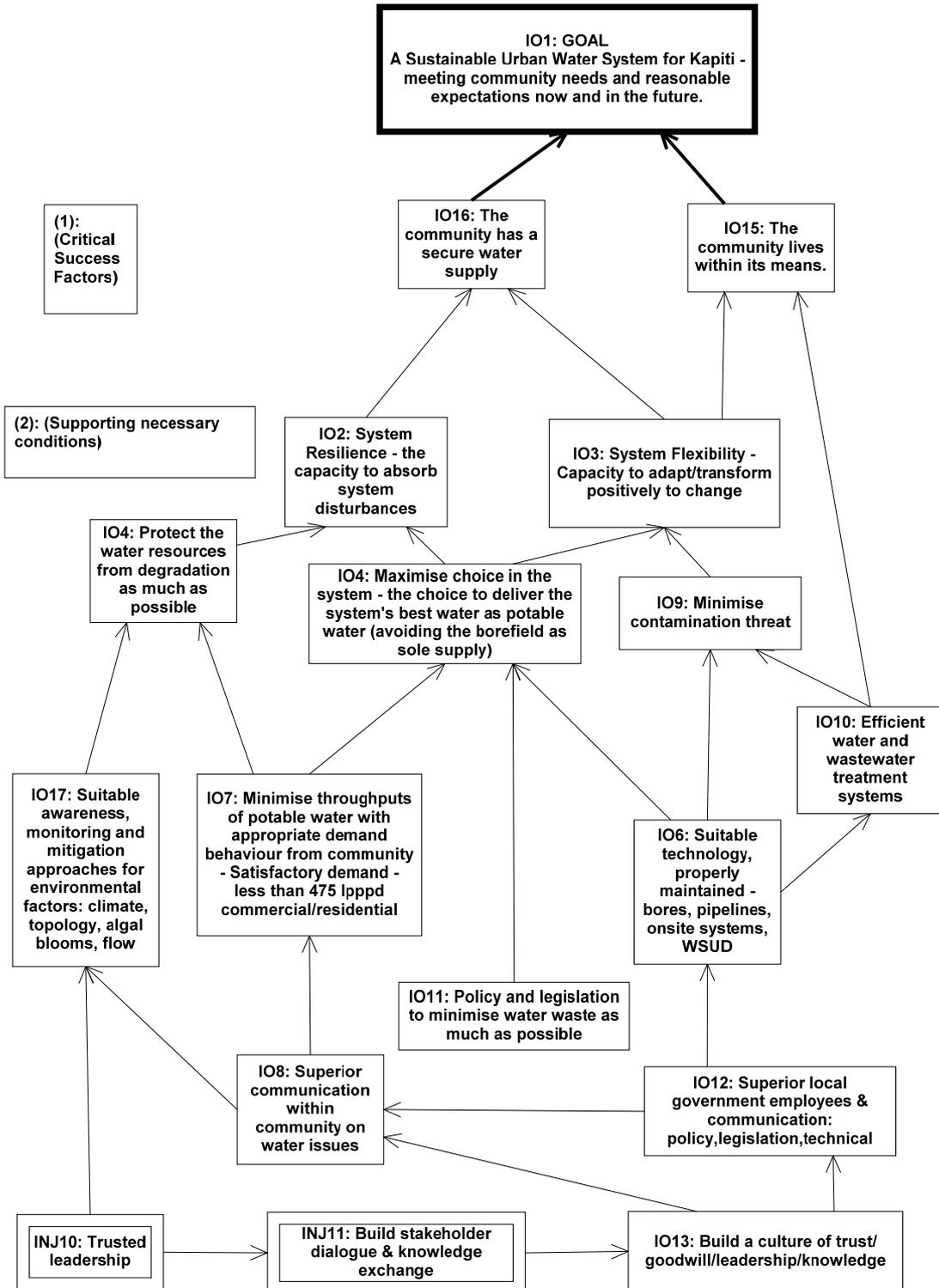


Figure 3. The final IO map prepared with input from all participants.

2.4 The Current Reality – What really is happening?

CRBs and CRTs are sufficiency-based (if...then) logic trees used to compare reality with system benchmarks in order to isolate what needs changing in a system. As such they only need to reflect the part of the system that is unfavourable (Dettmer, 2007:92).

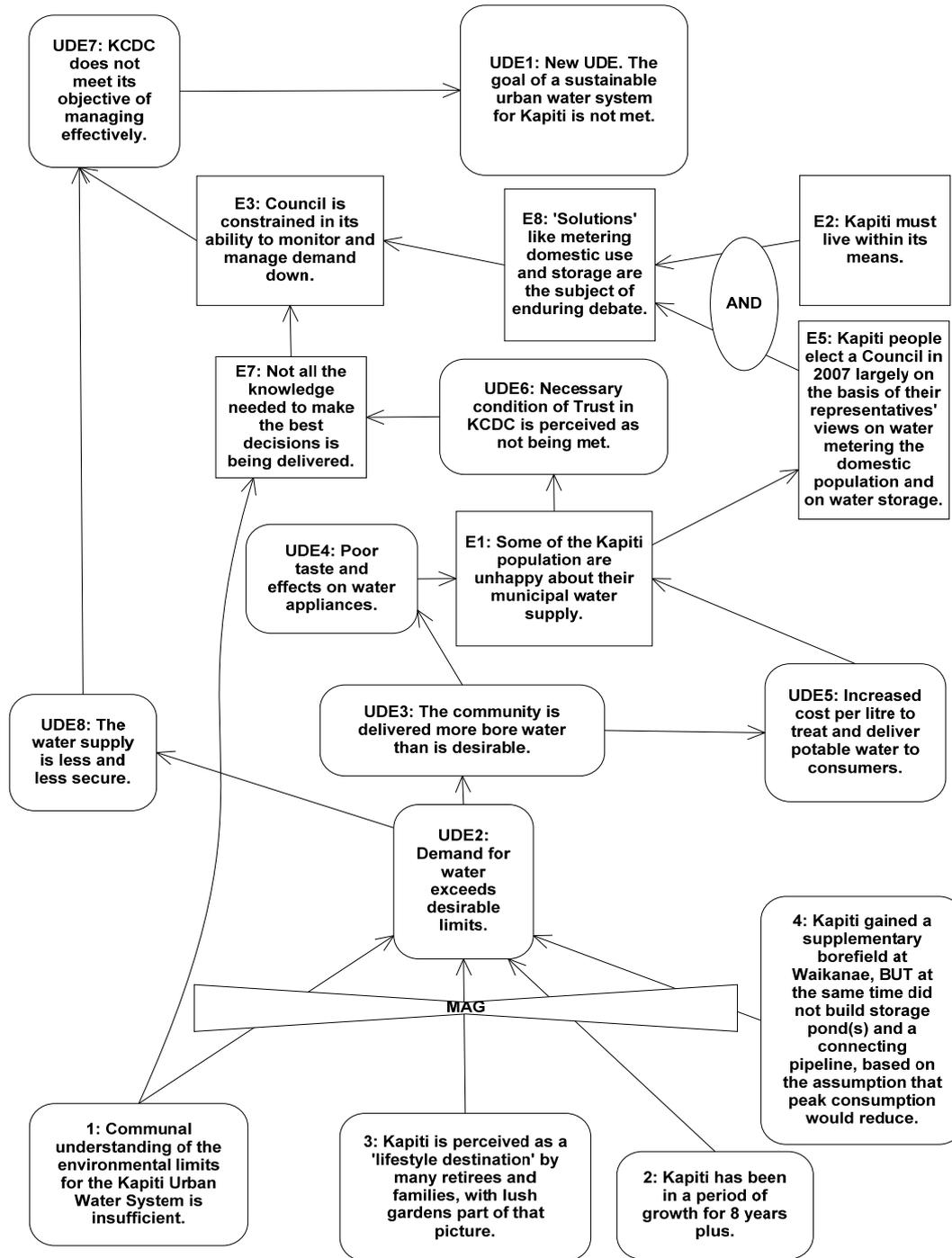


Figure 4. The CRB depicting what is happening in one part of the system.

The CRB reflects what is happening in the system now. Figure 3 maps the destination (the Intermediate Objectives), while Figure 4 shows the Current Reality Branch (CRB), focusing on the dilemma facing Waikanae/Paraparaumu/Raumati, and reflecting the part of the system most likely to impede the attainment of the system goal.

The CRB (*what to change*) was prepared in conjunction with a series of Evaporating (or conflict resolution) Clouds (or EC – see Dettmer, 2007). A UDE signifies an undesirable effect. The MAG shows four entities combining in a magnitudinal way to influence UDE2. One or more of these acts as a critical constraint to achieving the goal. Having focused on one part of the system using the participants' viewpoints, a broader CRT was constructed that could be compared with the final, most objective destination (IO) map of the system. The CRB and CRT are essentially gap-analysis tools (Dettmer, 2007). By comparing the CRB in Figure 4 with the IO map in Figure 3, the reasons behind the current reality differing from the preferred system were able to be determined.

3 Findings

The methodologies worked synergistically to give structure and clarity to situations, problems and perspectives, evoking a rich and valid picture of the system issues and potential solutions. The author contends that the TOC mapping procedures are akin to 'thinking out loud'. Participant stakeholders' contributions were captured and represented in the TOC trees and maps, that were analysed and refined with the involvement of willing participants. The participatory way that TOC was operationalised with the stakeholder typology and CLDs ensured that a variety of stakeholder perspectives and ideas came to light. Constructing the participants' 'conversations' into a systematic series of TOC trees and maps evoked the sense of an 'ongoing conversation' between system participants, even though most of them never met one another. This 'conversation' was the source of a 'roadmap' for change that participants could take ownership in.

Using a stakeholder typology with TOC to identify stakeholders in an analytical way ensured that stakeholders with important historical perspectives such as Greypower, were not unintentionally marginalised. The Starborough Flaxbourne Conservation Group, with their long experience of farming under drought conditions, was another stakeholder group identified using the stakeholder typology. This group would seem to have a tenuous link to the Kapiti urban water system at best, under a narrow systems definition. In reality, group members were able to provide valuable historical perspectives and critical insights into understanding Kapiti's problems and suggested innovative ways of dealing with them.

The methodologies of TOC, stakeholder mapping, and CLD's, combined to provide a means for the researcher to systematically work with a variety of key stakeholders, including Councillors, mana whenua (this and associated definitions in Environment Waikato, 2009), and central government agencies, without the need to have them all in the same room, or even in the same town. The value of this approach is not only to gain some deeper insights into how to protect urban investment in water assets, but also to conserve other scarce resources, notable among them, time. One practitioner can gather and articulate the viewpoints of all the stakeholder participants, and from them, gain agreement and ultimately ownership, of the desirable 'future reality' and the 'roadmap' needed to reach it.

The combination of problem structuring methodologies served to increase the participants' sense of connection and engagement with both problem and solution, with everyone's viewpoints validated by their part in the 'ongoing conversation'.

4 Conclusion

This thesis was essentially a conversation on the subject of how we might manage our fresh water for better outcomes. The study engaged with a number of willing individuals, all stakeholders in a New Zealand urban water system, and asked them for their perspectives about *the system destination, what to change, what to change to, and how to make the changes* necessary for more sustainable urban water systems.

The overarching message from the participants in this study is that for the urban water system to serve our communities and businesses better, the range of stakeholders must develop a deeper understanding of the system's limits and opportunities. If limits are not understood and agreed, it is difficult to live within them. If opportunities are not revealed, it is hard to grasp them (planting drought-resistant lucerne instead of rye-grass and clover for example, with its obvious parallel in the Kapiti context of planting gardens that do not need watering, and other less obvious implications). The thesis findings suggest that better decision-making is required to develop more sustainable environmental, and sound business agendas, that address the widest possible range of stakeholder interests. This is no small task. The participants were almost unanimous in the view that resource managers must take responsibility for raising the depth of understanding and gaining agreement towards a defined goal for the system and its range of stakeholders – and they need the resources and mechanisms to do it.

TOC provides a constructive and non-threatening way to encourage a kind of deep level reflection. However, the Pilot Study suggested that the TOC tools alone could not achieve the research objectives. There had to be a way of opening up the field of enquiry, and to facilitate ongoing reflexivity (Basset, 1995) and system improvement. The stakeholder engagement process developed in this thesis proved invaluable in this regard. Dettmer (2007) urges that simplifying approaches is the key to finding out what we know and that it is better to be approximately right than precisely wrong. This premise of simplicity was fundamental to the design of this research and to the process of ongoing engagement upon which this enquiry was constructed.

The Parliamentary Commissioner for the Environment (2000, 2001) had predicted nearly a decade ago that reaching consensus between stakeholders on environmental, social and economic goals for urban water systems would become one of the greatest challenges facing communities New Zealand-wide. The veracity of this appears indisputable in 2009 and is reason to value a methodological approach that might recognise the perspectives of diverse or divergent stakeholders, and at the same time provide the means to logically evaluate system issues and opportunities, and reveal suitable leverage points for motivating change.

5 Contribution

This paper has sought to make two contributions: the first is to test a methodology that might facilitate more integrated and better approaches to meeting the challenge of achieving more sustainable urban water systems in New Zealand; the second is to present the insights of participants to reveal assumptions underlying the not uncommon dilemmas faced by urban communities regarding water, and present a TOC 'roadmap' of the minimum changes required to resolve these dilemmas.

6 References

- Bassett, K. (1995). On reflexivity: further comments on Barnes and the sociology of science, *Environment and Planning*, 27, 1527-38.
- Dettmer, H. W. (2007). *The logical thinking process: a systems approach to complex problem solving*. Milwaukee WI: Quality Press.
- Elias, A. A., Cavana, R. Y. and Jackson, L. S. (2002). Stakeholder analysis for R&D project management. *R&D Management* 32 (4), 301-310. Malden, Mass.: Blackwell Publishers Ltd.
- Environment Waikato, (2009). *Appendix VI: Glossary*. Retrieved May 2009 from <http://www.ew.govt.nz/policy-and-plans/Regional-Coastal-Plan/Regional-Coastal-Plan/APPENDIX-VI-Glossary/>
- Freeman, R. E. (1984). *Strategic management: a stakeholder approach*. Boston MA: Pitman.
- Mitchell, R. K., Agle, B. R. and Wood, D. J. (1997). Towards a theory of stakeholder identification and salience: defining the principle of who and what really counts. *Academy of Management Review*, 22 (4), 853-886.
- Moore, R. M. (2009). *Shaping more sustainable communities: a case study in urban water management*. A thesis submitted to Victoria University of Wellington in partial fulfilment of the requirements for the degree of Master of Management Studies. For the thesis, please contact Victoria University Library or the author (robyn@j.co.nz).
- Parliamentary Commissioner for the Environment. (2000). *Aging pipes and murky waters. Urban water system issues for the 21st century*. Wellington: Parliamentary Commissioner for the Environment. June 2000.
- Parliamentary Commissioner for the Environment. (2001). *Whose water is it? The sustainability of urban water systems on the Kapiti Coast*. Wellington: Parliamentary Commissioner for the Environment. May 2001.
- Rawlins, Brad L. (2006). *Prioritizing Stakeholders for Public Relations*. Sourced from http://www.instituteforpr.org/research_single/prioritizing_stakeholders. Published by the Institute for Public Relations, www.instituteforpr.org.

Supply chain based agent simulation: Towards a normative approach

Luciano Ferreira
Computer Science Department
University of Cruz Alta
Brazil
lferreira@unicruz.edu.br

Denis Borenstein
Management School
Federal University of Rio Grande do Sul
Brazil
denisb@ea.ufrgs.br

Abstract

This paper presents an agent-based framework to the simulation of supply chain following the idea of normative agents. Normative agent is being developed to model elements in a system that are able to define norms and to control its use by other agents. Although some methodologies and more generic solutions have been proposed in SC modelling, they are not able to cope with supply chains in which regulation plays an important role. Several supply chains such as in the energy, food, chemical, and forestry areas are highly regulated. The main objective of this paper is to fulfill this gap in the literature, developing a modelling tool able to cope with regulation in supply chain. The modeling of a Brazilian biodiesel supply chain is presented as a case study.

Key words: Simulation, normative multi-agent system, supply chain modelling

1. Introduction

Manufacturers normally buy components from other organizations and sell to distributors, who then sell to retailers. Maximizing the efficiency of the supply chain (SC) linking business is increasingly important and difficult as products involve more parts, drawn more widely from around the world, and as managers attempt to reduce inventory and increase the availability of goods (Simchi-Levi et al, 2002).

The SC is a network of suppliers, factors, warehouses, distribution centres and retailers through which raw materials are acquired, transformed and delivered to customers. SC management (SCM) is the strategic, tactical and operational decision making that optimizes supply-chain performance. The strategic level defines the supply chain network; that is, the selection of suppliers, transportation routes, manufacturing facilities, production levels, and warehouse number and locations. The tactical level plans and schedules the supply chain to meet day-by-day demand. The operational level

executes the plans. Tactical - and operational - level decision making functions are distributed across the supply chain (Fox et al., 2000).

Modelling supply chain is a good way of studying order fulfilment processes and investigating the effectiveness of management policies. Multi-agents models are increasingly being used for this purpose. According to Gilbert (2008, p.11) a multi-agent model fits well with the task of simulation supply chain since the business elements involved in a SC can be modelled as agents, each with its own rules. It is also easy to model the flow of products down the chain and the flow of information, such as order, volumes and lead times, from one organization to another, using the agent approach.

This work aims to contribute with the state of the art in the SCM area as follows: (1) building a framework for SC modelling and simulation, providing generic agents which may be easily extended and used in other application contexts, and (2) exploiting normative agents in the context of SCM. Normative multi-agent systems may be approached as the intersection among normative systems and multi-agent systems (Boella and van der Torre, 2006). The integration of these areas (SCM and normative multi-agent systems) increases the possibilities of supply chain modelling, allowing the inclusion of external entities which normally influence the SCM management, such as governmental organizations and regulation agencies.

This work is organized as follows: section 2 presents a brief literature review on the state of the art of the area and points out the potential contributions of this work; section 3 presents some definitions related to agent-based modelling and the basic premises to elaborate an agent based modelling using normative agents; section 4 presents the developed framework; section 5 shows the applicability of the model using the Biodiesel supply chain as a case study. The last section presents the final considerations and conclusions.

2. Previous Research

Swaminathan et al. (1998) was one of the first papers to deal with the application of agents to SCM. The authors presented a generic agent that is specialized to execute different activities in a supply chain. The study divided the agents in two main types: structural agents (retailer, distributor, producer, supplier and transporter) and control agents (demand, stock and information). Similar study was presented by Julka et al. (2002), where a unified framework for SC modelling and monitoring is discussed.

Fox et al. (2000) presented an architecture based on agents for the SCM that include the tactical and operational levels. The authors presented the development of an agent that offers re-usable and generic components, besides offering support for the development of cooperative agent models. Sadeh et al. (1999) presented a review of the MASCOT architecture (Multi-Agent Supply Chain Coordination Tool).

Nissan (2001) presented a model based on agents for the General Motors supply chain, where the capacity of the agents to perform business in the name of users, buyers and sellers is evaluated. Davidsson and Wernstedt (2002) utilized multi-agent systems to coordinate the production and distribution in a just-in-time system.

Scheritz and Gröbler (2002) presented a modeling strategy which combines system dynamics with the agent-based simulation approach. Cavalieri et al. (2003) utilized multi-agent systems to evaluate lateral and vertical coordination mechanisms among the supply chain components. Janssen (2005) developed a simulation model based on

agents which utilizes a combination of structural and functional agents proposed by Cavalieri et al. (2003). The simulation showed how to reduce the lead time and the stock out in those peak movement periods of a case study.

Chiu and Lin (2004) discussed the utilization of software agents and artificial neural network (ANN) for the context of a collaborative SC that works under the idea of assembler-to-order. In their work, agents were modeled like nodes of the ANN and agents with similar functions are grouped in the same layer; the model is composed by a structure of four layers, each one representing an actor of the supply chain (supplier, producer, distributor, and consumer).

Melle et al. (2007) proposed an optimization strategy which uses genetic algorithms and agent-based simulation models to determine the parameters of the stock control policy of the distribution center (R, s, S). Zarandi et al. (2008) also developed a simulation model composing the idea of software agents with genetic algorithms, but aimed at reducing the bullwhip effect and at minimizing the total cost in a four-stage supply chain. Chatfield et al. (2007) presented a generic architecture for supply chain modeling. The authors divide the supply chain into five types of constructs, or fundamental classes: node, arc, component, action and policy. The physical aspects of the supply chain were described as nodes, arcs and components, while the functional and logical aspects were described as actions and policies.

Although the cited previous researches present some interesting ideas, there are several gaps that are neglected by these studies. The most important ones are as follows: (1) the actors of the supply chain and their relationships are only defined during the supply chain project; (2) most of the research works are concentrated in finding the solution of a particular problem, specifically carried out for an enterprise or specific supply chain.; (3) the studies are focused on determining behavioral differences caused by changes of parameters or agents that are components of the supply chain. In addition, there is no presence of a regulator agent, monitoring, structuring, and resolving conflicts.

The introduction of a new agent with the purpose to regulate the behavior of the supply chain and of its actors, proposing indicators, defining taxes, authorizing or not performance of a given agent which may be proven inefficient or evil intentioned, among other tasks, is an important modelling aspect to be achieved. Once many supply chains are inserted in government controlled environments or regulation agencies, which dictate policies that must be followed by the actors of a SC, the development of such agents become a very important feature in SCM modelling. Moreover, generic solutions are needed, which may serve as basis to solve new problems that still need to be better exploited.

3. Normative Multi-Agent Systems

Conceptually, we can define an agent as a computational system which is situated in an environment and that is able to execute actions in an autonomous way in this environment aiming to reach its goals (Wooldridge, 2002). A complementary definition is given by Weiss (2001) in which an agent is defined as an application that may operate with robustness in environments that are quickly modified and that require precise answers even to unexpected events. Agents must react quickly and show characteristics able to allow them to operate even under non-programmed situations. Furthermore, agents must demonstrate capabilities of interacting with other agents. The flexible and

rational behavior is reached through problem resolutions, planning, decision-making and learning (Weiss, 2001).

A multi-agent system, or society, is defined as a set of agents where each one can work in the name of different types of entities whose goal must be satisfied. In societies like this, it is not uncommon to appear conflicts of interests among agents. To avoid these situations or, at least, minimize them, the introduction of rules to command a behavior is very important. Rules may be defined as a special constraint, defining a behavioral pattern for agents (Boella and van der Torre, 2006). However, as the priority of every agent is to satisfy its goals, before being in agreement with rules accomplishment, each agent must evaluate the positive or negative impact of such rules on its goals. This way, the normative behavior of an agent can not be assured and thus the society must have mechanisms to reward agents that accomplish the proposed rules (López and Márquez, 2004). A normative agent, thus, may be defined as an autonomous agent able to adopt, deliberate and comply with norms, respecting its goals and preferences (López *et al.*, 2005). Thus, its behavior is partially determined by tasks that must be accomplished, prohibitions that limit its goals, social commitments that must be created during its life cycle and social codes that represent social satisfaction for the agent.

Figure 1 presents the basic architecture of a normative agent. According to this scheme, the adoption process of a rule is based on beliefs, on motivations and on goals of an agent. For an agent to adopt a rule it is necessary to recognize it, through its perception, that this rule has not yet been adopted, and that it had been sent by a recognized authority. The deliberation process of a rule involves the evaluation of two aspects: goals that the agent will not perform (if the rule is adopted) and the associated benefits brought by the adoption of the rule. The deliberation process divides the rules in two groups: intended and rejected. Once the agent has deliberated on which rules it is going to execute, a new process is initiated (norm compliance) in order to update the goals of each agent as well as adapt its behavior.

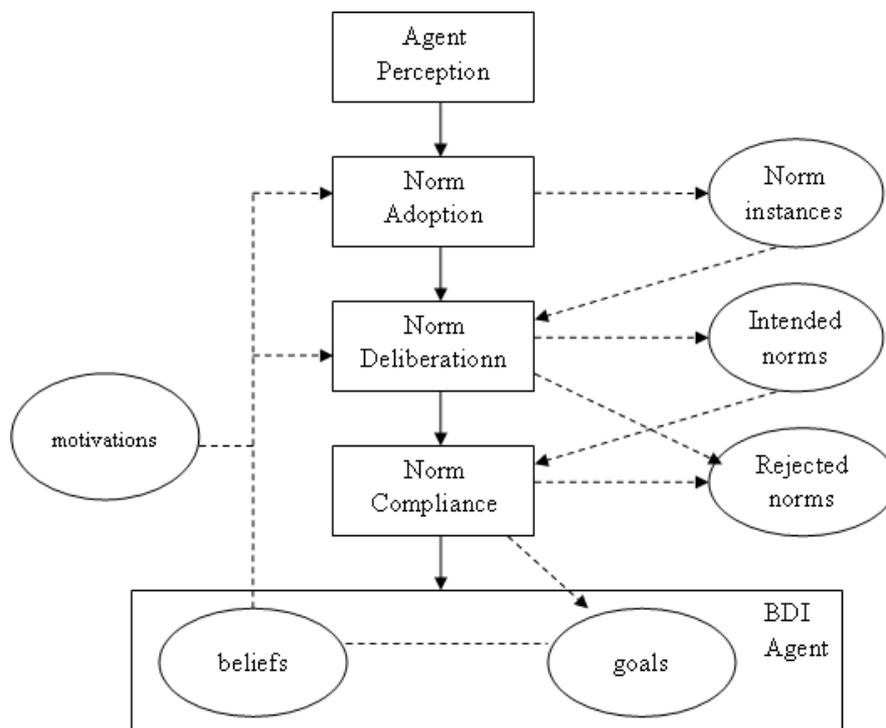


Figure 1. Normative agent architecture (López *et al.*, 2005)

4. Normative Agent Framework

In this section, a framework for supply chain modelling and simulation using normative agents is presented. A supply chain may be defined as “a network of autonomous business entities or semi-autonomous, collectively responsible by the acquisition, production and distribution associated with one or more family of products” (Swaminathan et al., 1998). It is possible to represent a SC as a set of agents acting in a collaborative way to produce a given product; interacting with other agents to acquire products; consumer agents performing requests; and regulating agents acting in this context. Thus, the configuration possibilities of SC environments are immense. Taking this fact into account, it is important to define a framework from which different simulation scenarios may be built and/or quickly expanded, as well as different dominion problems may be studied and evaluated, without the need of a previous construction of all particularities of the model.

Initially, to construct this framework, a generic agent (SCMAgent) derived from a normative agent was defined. The normative agent model presented by López et al. (2005) was used as a main reference. In this work the authors propose a framework for normative multi-agent systems whose main components are as follows (see Figure 2):

- A set of normative agents (*NormativeAgent*): these agents might adopt, deliberate and comply with norms;
- A set of legislator agents (*LegislatorAgent*): these agents might create, change and abolish norms;
- A set of norm defender agents (*DefenderAgent*): these agents might give rewards or to apply punishments to other agents according to the compliance to norms;
- A set of norms (*Norm*) directed to regulate the behavior of the agent;
- A set of norms whose purpose is to enforce and to determine the fulfillment of the most recent set of norms;
- A set of norms directed to promote the fulfillment of norms through rewards;
- A set of emitted norms to allow the creation and the abolition of norms.

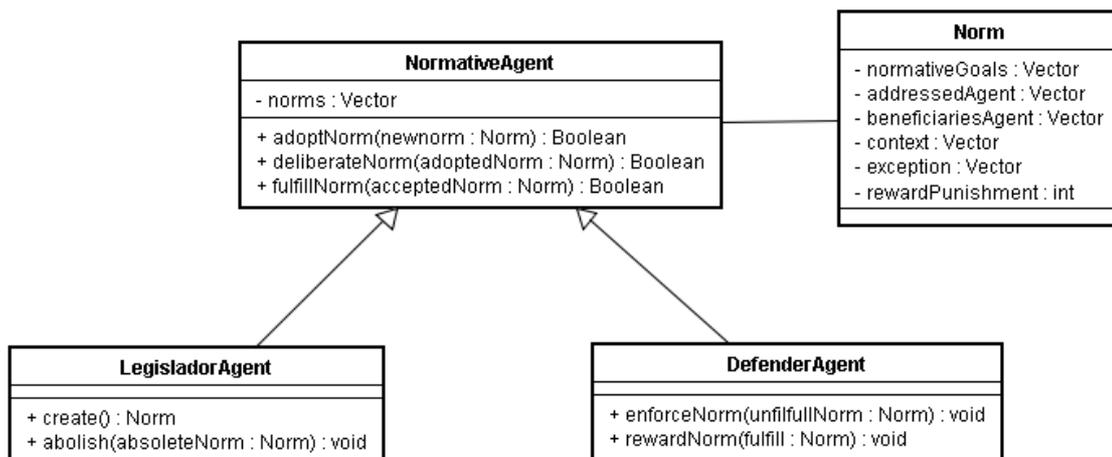


Figure 2 – Types of agents in NMAS (López et al., 2005)

After the design of normative agents, a framework for supply chains was built. Figure 3 presents the main components of the proposed framework. The model presents a generic agent (*SCMAgent*) that inherits all the functionalities of a normative agent.

From this agent, two main classes of agents are defined, following the framework proposed by Swaminathan *et al.* (1998): (1) structural agents: to represent the actors of the supply chain and (2) control agents: to manage processes and to make decisions. A brief description of these agents is presented below:

- *Retailer*: supply the consumers demand, each agent of this type may have different necessities in terms of products, time and level of service;
- *Distribution Center*: is an intermediate agent, receives products or parts from suppliers and delivers to the consumer;
- *Manufacturer*: assemble or manufacture a product or a set of product. Each product has an associated bill of material;
- *Supplier*: supply parts to the manufacturer or other agents. It has a level of production, production costs and a lead time defined per unit of time and per product;
- *Inventory*: control the inventory levels of a particular product or raw-material;
- *Demand*: implements a mechanism to calculate how much inventory is needed and when it will be needed. Its goal is to have the right product in the right spot at the right time.

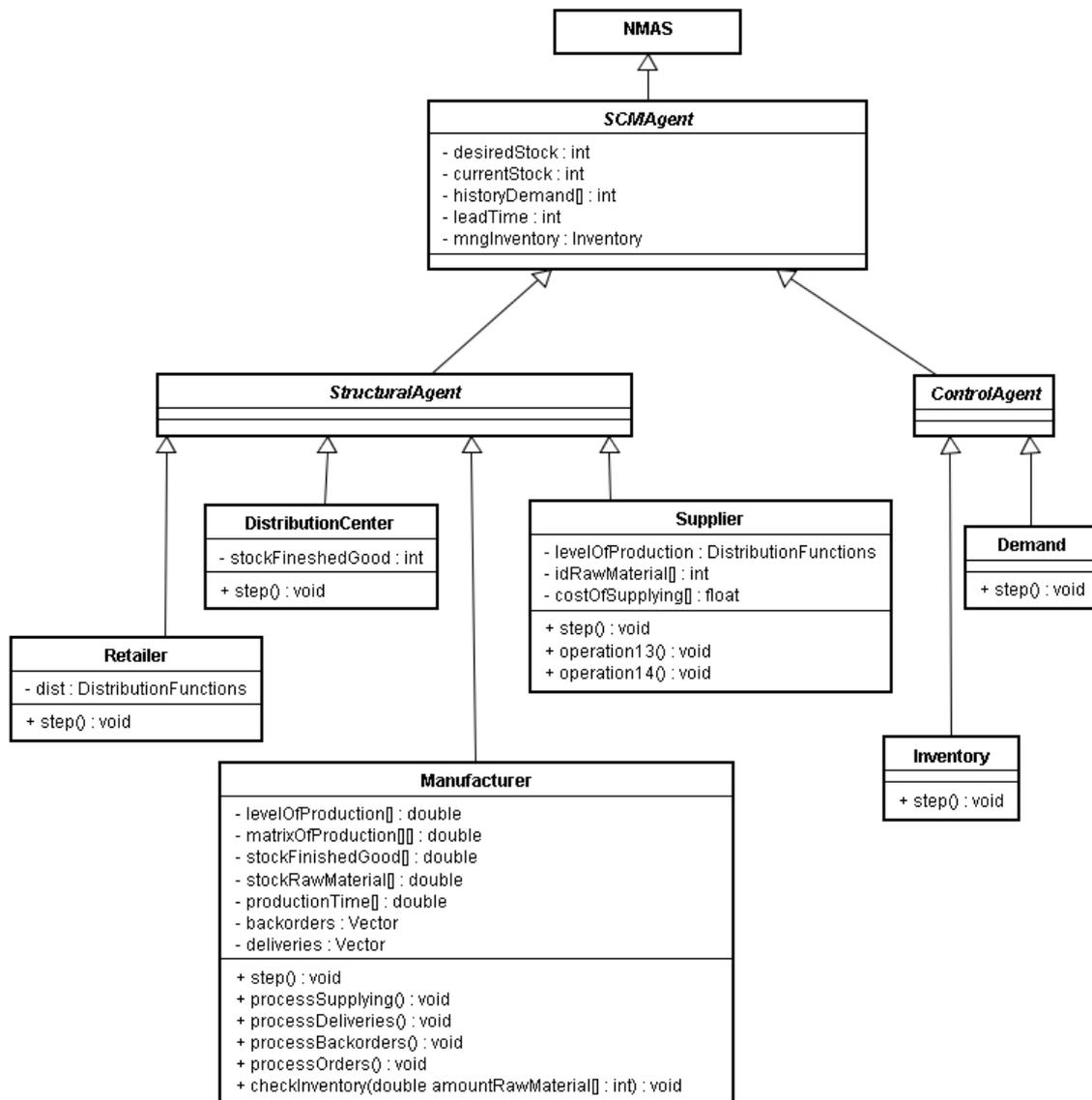


Fig. 3 – Normative supply chain framework

5. Case Study

This section demonstrates the applicability of the model proposed in this work. The Biodiesel supply chain modelling in Brazil is presented as a case study. This is an appropriate case to demonstrate the necessity of modeling supply chains using normative agents, since this is the typical regulated SCM by energy regulating agencies.

Biodiesel is a biodegradable fuel derived from renewable sources that may be obtained through different processes, such as cracking, esterification or transesterification (Hass *et al.*, 2006). In Brazil, there are hundreds of vegetable species from which it is possible to produce Biodiesel, such as castor-oil plant, African oil-palm, sunflower, peanut, babassu, pine seed and soybean, among others. The federal government launched the National Biodiesel Production and Use Program in December 2004 to stimulate the introduction of the biofuel in the national energetic matrix. The refineries and distributors are authorized to add 2% bio-fuel to the mineral fuel (B2), requiring a production superior to 800 million liters of bio-fuel per year. In 2013 the tax will increase to 5%, equivalent to 2.5 billion annual liters.

Figure 4 presents a general view on the actors of the biodiesel supply chain. Initially, it is considered a set of suppliers of raw-material for the biodiesel production plant. The mass balance for the production using the vegetable oil is presented. From this process glycerin and a residue are also extracted.

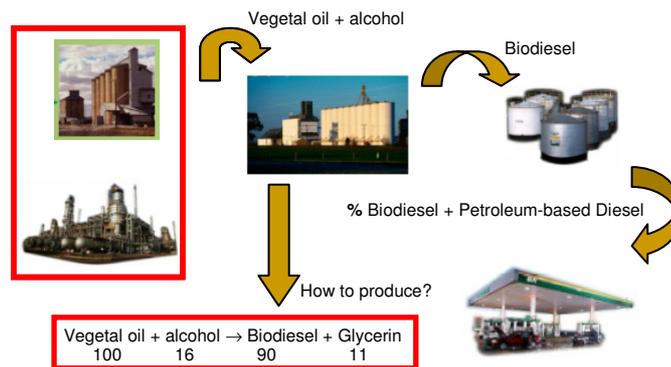


Figure 4 – The Biodiesel supply chain

After the production, Biodiesel is sold for fuel distributors and is then mixed to mineral diesel (according to the governmental stipulated percents) before being distributed for consumption. According to the Brazilian Law, the distribution can only be carried out by companies licensed by the National Petroleum, Natural Gas and Biofuels Agency. Thus, biodiesel can not be directly sold to retailers, but rather to regional distribution bases controlled by this agency.

The model presented in Figure 5 was elaborated taking into consideration these definitions. The supply chain was structured containing raw-material suppliers, production plant, distribution center and retailer. Besides, normative agents are included in order to guarantee some norms imposed by the federal government regarding the production of this biofuel, such as the percentage of Biodiesel mixture to the mineral diesel (*Legislator-Distribution and Defender-Distribution*). Another one being considered is related to the compliance with the “social stamp” plan. Some biodiesel producing plant receive the so called “social stamp” if the raw-material is derived from

small producers (*Legislator-Biodiesel and Defender-Biodiesel*). As counterpart they receive fiscal incentive from the government. The proposed framework was developed using the Repast tool (North and Macal, 2005) for agent modelling.

Different experiments may be carried out by the model, serving as subsidy for the establishment of public policies, as well as dynamic analysis of the Biodiesel supply chain. Figure 6 shows the results of one of the performed experiments, where the production cost for a given production mix and the consequent demand for raw-material is presented. In this experiment we are not searching for the raw-material mix in order to minimize production cost, but rather the one that results in the fulfillment of the obligations of the production plant with the “social stamp”. The demand generated by the distribution bases and the resulting demand for raw-material in the supplier stage try to satisfy the mixture percents authorized by the government. Among the results obtained by the model it is possible to mention the equilibrium point between soybean offer and Biodiesel production. It is estimated that if the same intern consumer patterns and soybean exportation of the last few years are maintained, a 4% increase in the average soybean production will be necessary to supply the demand for B2 and 13% to supply the demand for B5.

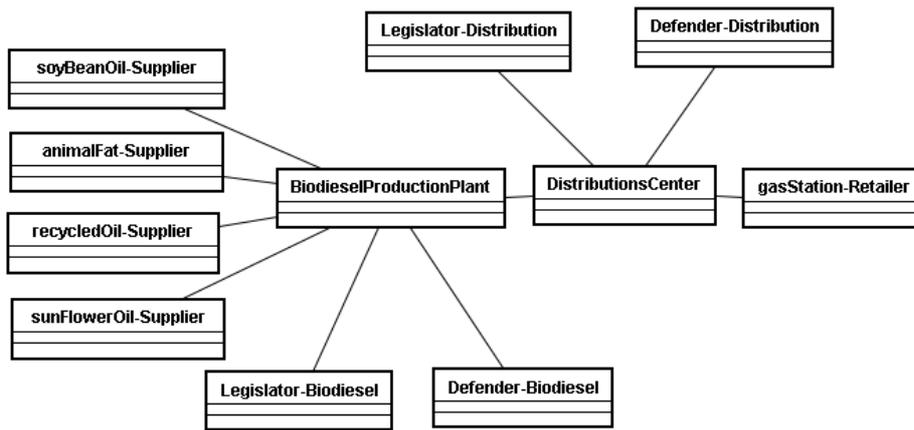


Figure 5 – Biodiesel supply chain modelling

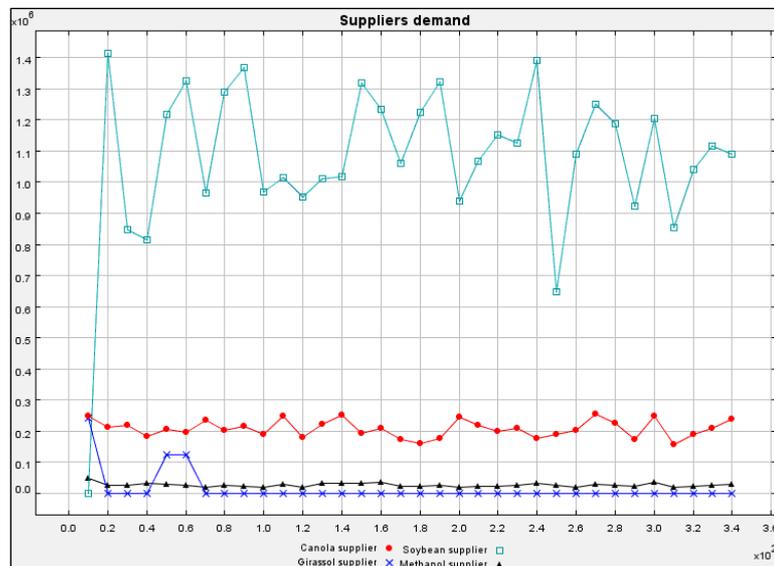


Figure 6 – Results obtained

6. Conclusions

This work presented the design of a framework for supply chain modeling proposing the idea of utilizing normative agents to study supply chains. The case of the Biodiesel supply chain in Brazil was presented. The developed model can also be utilized for other purposes, such as inventory control, estimate production price, determinate the raw-material mix for production and analyze the bullwhip effect.

Many areas of our research can be further extended as follows: (1) developing a graphic interface to assist the users in the design of the supply chain. Nowadays the construction of the model and the agents parameters configuration is made in the program code; (2) expanding the developed model in order to consider logistics issue such as location, distribution and routing, for example; (3) adding new agents to the framework; and (4) adding and removing norms during the simulation running. Norms are defined in the project phase in the current version.

7. References

- Cavaliere, S., V. Cesarotti, V. Introna. 2003. A Multiagent model for coordinated distribution chain planning. *Journal of Organizational Computing and Electronic Commerce* **13**:267–287.
- Chatfield, D. C., J. C. Hayya, T. P. Harrison. 2007. A multi-formalism architecture for agent-based, order-centric supply chain simulation. *Simulation Modelling Practice and Theory* **15**:153–174.
- Chiu, M., G. Lin. 2004. Collaborative supply chain planning using the artificial neural network approach. *Journal of Manufacturing Technology Management* **15**: 787-796.
- Davidsson, P., F. Kwerntedt. 2002. A multi-agent system architecture for coordination of just-in-time production and distribution. *The Knowledge Engineering Review* **17**: 317–329
- Fox, M. S., M. Barbuceanu, R. Teigen. 2000. Agent-oriented supply chain management. *The International Journal of Flexible Manufacturing Systems* **12**:165–188.
- Gilbert, N. 2008. *Agent-Based Models*. Sage Publications, Thousand Oaks, California.
- Janssen, M. 2005. The architecture and business value of a semi-cooperative, agent-based supply chain management system. *Electronic Commerce Research and Applications* **4**:315–328.
- Julka, N., R. Srinivasan, I. Karimi. 2002. Agent-based supply chain management – 1: Framework. *Computers and Chemical Engineering* **26**:1755–1769.
- López, F., A.A. Márquez. 2004. An architecture for autonomous normative agents. In: *Proceedings of the Fifth Mexican International Conference in Computer Science*.
- López, F., M. Luck, M. d’Inverno 2005. A normative framework for agent-based systems. In: *Proceedings of the Symposium on Normative Multiagent Systems*, 24–35.
- Mele, F. D., G. Guillén, A. Espuña, L. Puigjaner. 2007. An agent-based approach for supply chain retrofitting under uncertainty. *Computers and Chemical Engineering* **31**: 722–735.
- Nissan, M. E. 2001. Agent-based supply chain integration. *Information Technology and Management* **2**: 289–312.
- Simchi-Levi, D., P. Kaminsky, E. Simchi-Levi. 2003. *Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies*. McGraw-Hill, New York.

- Swaminathan, J. M., S.F. Smith, N.M. Sadeh. 1998. Modeling supply chain dynamics: A multiagent approach. *Decision Sciences* **29**: 607–632.
- Zarandi, M. H. F., M. Pourakbar, I.B. Turksen. 2008. A Fuzzy agent-based model for reduction of bullwhip effect in supply chain systems. *Expert Systems with Applications* **34**:1680–1691.
- Weiss, G. 2001. Agent orientation in software engineering. *The Knowledge Engineering Review*, **16**: 349–373
- Wooldridge, M. 2002. *An Introduction to Multiagent Systems*. John Wiley & Sons, New York.

Some thoughts on model use in OR/MS

Michael Pidd

Department of Management Science
Lancaster University
Lancaster LA1 4YX
UK
m.pidd@lancaster.ac.uk

Visiting Professor
Victoria Management School
PO Box 600
Victoria University of Wellington
Wellington 6140

Extended abstract

1.0 Introduction

OR/MS models are used in many different ways. Sometimes the models are used to automate a decision-making process: for example, credit scoring models are used routinely to determine whether or not to grant loans to applicants. However this is not the only way that OR/MS models are used. Some models are built to facilitate human decision-making in complex problems and may only be used once in support of a specific decision. Such models are not intended, in any sense, to automate decision making, but rather to support human decision makers as they puzzle over difficult issues.

Pidd (2010, p10) suggests that an OR/MS model is an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality. This definition is far from perfect and avoids the question of what may constitute reality, which can loom large. It is not concerned whether a model is based on a sophisticated mathematical formulation or whether it is just a simple flow diagram showing how entities are believed to relate to one another. It stresses that models are approximations, built with some intended use(s) in mind and that they are the product of human thought and ingenuity.

2.0 A spectrum of model use

There are many ways in which OR/MS models can be used. Figure 1 shows a spectrum of model use, with two opposing axes that are used to position different types of model use. The two axes might be regarded, by some, as orthogonal, but this does not matter for present purposes. The axes of figure 1 are:

- The degree to which use is intended to be frequent and routine, rather than occasional or one-off.
- The amount of human interaction involved in making the model ready for each use.

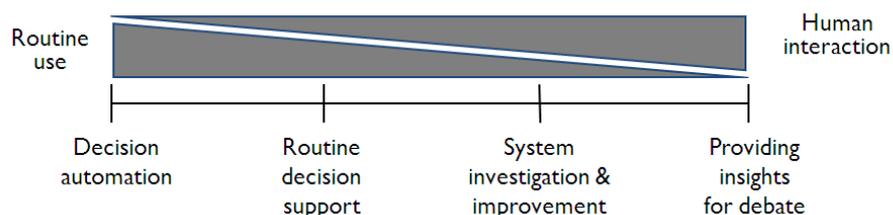


Figure 1: A spectrum of model use

To illustrate the point of the spectrum, figure 1 has four archetypes marked on it, though there are obviously others that could be identified.

2.1 Decision automation

This refers to model use that is frequent and routine, with, in general, no need to prepare the model for each use. As mentioned earlier, credit scoring models commonly replace human decision making on a frequent and routine basis, as they do in dynamic pricing. A user operating in this mode supplies data to the models. Most decision automation is implemented so that a single wrong decision will not bring a business to its knees. Typically, the value of such individual decisions is rather small, though their cumulative effect can be considerable. Because the use of these models often forms a core business process, it is important that this decision automation is done well. Though such models are usually employed for decisions of individually low value and replace human decision making, it would clearly be unwise to allow them to operate unmonitored. Hence, in the background, will be humans who track the performance of the models (probably using other models to aid them).

2.2 Routine decision support:

This second form of model use occurs when they are used to assist, but not replace, people making routine, repeated decisions. This might be a preliminary to attempts at decision automation, hoping that successive models will eventually become good enough for this. However, real-life is usually more complex or rather less certain than can be wholly encoded in an OR/MS model. Hence models may be just one component in a decision-making process; that is, used for decision support and the human decision makers and planners may sensibly over-ride them when appropriate. This is common in the scheduling applications such as operating theatres or airlines and trains. A variation on the same theme occurs in forecasting, when statistical forecasts are combined with judgemental forecasts.

2.3 Modelling for investigation and improvement

This is close to the view of model use usually presented in OR/MS textbooks. It occurs when models are used to support investigations that are relatively unique, possibly in system design, system improvement or an attempt to gain understanding of a very complex situation. There are many such examples, such as locating a distribution depot, the investigation of improved performance in an accident and emergency department, the design of business processes to support automated decision-making or the study of diseases to improve intervention and treatment. Models used this way are often purpose-built for the particular study, though it is sometimes possible to re-use one that already exists or to parameterise a generic model.

2.4 Modelling to provide insights

This applies to situations close to what Rittel and Webber (1973) term wicked problems. This mode of model use is most strongly associated with what have come to be known as problem structuring methods or 'soft OR', such as SSM or cognitive and causal mapping, but it is important to realise that quantitative models may also be used in this way. Chelst and Bodily (2000) discusses how decision analysis and decision trees can be used to help explore options and reduce uncertainty rather than as a way to take a decision. Robinson (2001) argues that many simulation models are actually used in this

mode by helping people to gain useful insights so they can move to a form of accommodation that leads to action.

3.0 So what?

Identifying the four archetypes of model use in figure 1 enables a discussion of aspects of model use that may be useful in developing a theory of model use. In the presentation at the conference, three aspects of model use are discussed to illustrate important features of each of the four modes.

- The importance of model validation.
- Data requirements for this type of model use.
- The value added by such model use.

In addition, the paper reports on a small-scale, empirical pilot investigation of the model use spectrum and considers the possible value of a theory of model use in OR/MS.

Acknowledgements

This extended abstract is based on a paper, Why Modelling and Model Use Matter, due to appear in the Journal of the Operational Research Society in January 2010.

4.0 References

- Chelst K and Bodily S.E. (2000) Structured risk management: filling a gap in decision analysis education. *Journal of the Operational Research Society*, 51, 12: 1420-1432.
- Pidd M. (2010) *Tools for thinking: modelling in management science*, 3rd edition. John Wiley & Sons, Chichester.
- Rittel H.W.J. and Webber M.M. (1973) Dilemmas in a general theory of planning. *Policy Sciences*, 4: 155–69.
- Robinson S (2001). Soft with a hard centre: discrete-event simulation in facilitation. *Journal of the Operational Research Society*, 52, 8: 905–915.

Commuter Cyclist Route Choice and the Bi-Objective Shortest Path Problem

A. Raith, C. Van Houtte, J.Y.T. Wang and M.Ehrgott
The University of Auckland, New Zealand
a.raith@auckland.ac.nz

Abstract

Commuter cyclists choose their route differently to drivers of private vehicles. It is commonly assumed that commuter drivers have only one route-choice objective, to reduce their generalised travel cost (a monetary value representing a combination of travel time and vehicle operating cost). However, commuter cyclists may have multiple objectives when choosing their route: the travel time and the suitability of the route for cycling. Some of the factors that characterise the suitability or attractiveness of a route include safety (influenced by traffic volumes, traffic speeds, the presence of bicycle lanes, etc.) and whether the terrain is flat or hilly. We model the potential routes considered by a commuter cyclists as solutions to the bi-objective shortest path problem with objectives travel time and attractiveness. Rather than determining a single route for a cyclist, we determine a choice set of optimal alternative routes (efficient routes) from which a cyclist selects one according to their personal preference. We show how this method is applied in a case study in Auckland. Future research will try to answer the question what portion of cyclists chooses each of the routes in the choice set. Subsequently, we aim to derive a traffic assignment algorithm for cyclists.

Key words: Bi-objective Shortest Path, Cycling, Route Choice

1 Motivation

Modelling the route choice of motorised vehicles is an important part of strategic transport planning called *traffic assignment*. One major application is the evaluation of the benefit of transport infrastructure developments and how they affect traffic congestion and travel patterns. While route choice of motorised vehicles is well researched, only little is known about how commuter cyclists choose their routes. In particular, the equivalent of the traffic assignment problem has not been formulated for cyclists. With increasing demand on transport networks, the bicycle becomes more important as alternative to personal motorised vehicles. In order to encourage cycling, the cycling infrastructure must be developed to make cycling safer and more convenient. Cyclist traffic assignment may assist in identifying parts of the network in need of improvement, and can also be used as a tool evaluate those infrastructure

developments. A first step towards a cyclist traffic assignment algorithm is understanding cyclist route choice and determining a *choice set* of routes cyclists may travel on. Then, the percentage of cyclists that choose to travel on each of the paths in the choice set needs to be determined and incorporated into a traffic assignment algorithm.

2 Route Choice Factors

Modelling route choice of motorised vehicles and commuter cyclists are two significantly different problems. Car travel is subject to network congestion implying that travel time is not proportional to distance travelled. It is assumed that motorists choose their route with the aim of minimising their individual travel time or some generalised cost function combining time and other route choice factors.

The identification of main factors that influence cyclist route choice is the subject of numerous studies. Travel time (or distance) is identified as the most important objective according to which cyclists choose their route, e.g. Aultman-Hall, Hall, and Baetz 1997. It is found that 58% of all commuters actually follow their shortest-distance path, which clearly indicates that travel distance (or travel time, which should be highly correlated with distance) is one of the most important factors of route choice. However, there are other influencing factors that lead cyclists to divert from their shortest-distance route. Stinson and Bhat 2003; Aultman-Hall, Hall, and Baetz 1997 study other factors affecting cyclist route choice. Main factors are identified as traffic volume, road gradient, presence of dedicated cycling facilities, presence of on-street parking, etc.

Therefore, we formulate cyclist route choice as a bi-objective problem with *travel distance* as one objective, whereas all other route choice factors are combined into a second objective that we call *attractiveness*.

3 Commuter Cyclist Route Choice Model

$G = (\mathcal{V}, \mathcal{A})$ is a *directed graph* with a set of nodes $\mathcal{V} = \{1, \dots, n\}$ and a set of arcs $\mathcal{A} = \mathcal{V} \times \mathcal{V}$. $\mathcal{P}_{s,t}$ is the set of all paths from origin $s \in \mathcal{V}$ to destination $t \in \mathcal{V}$. The distance along path $p \in \mathcal{P}_{s,t}$ is $\sum_{a \in p} d_a$ with d_a being the length of arc a . The attractiveness along p is the distance-weighted average attractiveness score α_a , $\frac{\sum_{a \in p} d_a \alpha_a}{\sum_{a \in p} d_a}$. The choice set of cyclists commuting from s to t is obtained as the set of efficient solutions of a bi-objective shortest (BSP) path problem with a minimisation and a maximisation objective:

$$\begin{aligned} \min \quad & d(p) = \sum_{a \in p} d_a \\ \max \quad & \alpha(p) = \frac{\sum_{a \in p} d_a \alpha_a}{\sum_{a \in p} d_a} \\ \text{s.t.} \quad & p \in \mathcal{P}_{s,t}. \end{aligned} \tag{1}$$

Unfortunately, this problem cannot be solved with a standard BSP algorithm (Raith and Ehrgott 2009) as adding new arcs to a path may decrease the value of the attractiveness objective $\alpha(p)$. As we are unable to solve problem (1) as it is, we define an auxiliary problem by removing the denominator of the attractiveness objective, we replace $\alpha(p)$ by $\hat{\alpha}(p) = \sum_{a \in p} d_a \alpha_a$. It can be shown that efficient solutions of the original problem (1) are always efficient solutions of the auxiliary problem. Hence, a

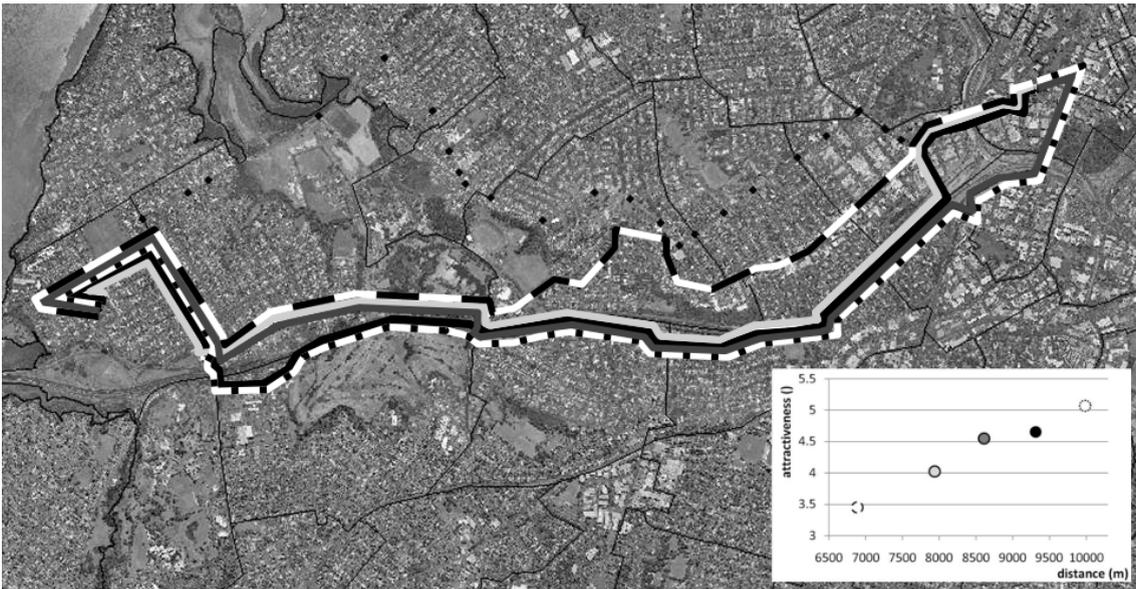


Figure 1: Efficient paths for cyclist trips between Pt. Chevalier and the CBD, and their distance / attractiveness scores. Source of aerial map: Auckland City Council.

standard BSP algorithm is used to solve the auxiliary problem and from its efficient solutions those that are also efficient for the original problem (1) can be selected. The efficient paths of problem (1) represent the choice set of commuter cyclists.

More details can be found in Raith et al. 2009.

4 Example

We demonstrate the derived cyclist route choice model for a selected origin and destination in Auckland, from Pt. Chevalier to Auckland CBD (from west to east) as indicated in Figure 1. The obtained efficient paths are highlighted in the figure. The shortest path is the top-most dashed path in the figure, whereas all other paths are longer but safer as they all follow an off-road cycleway. The corresponding path length and attractiveness values are also shown in the lower-right corner of the figure.

References

- Aultman-Hall, L., F.L. Hall, and B.B. Baetz. 1997. "Analysis of Bicycle Commuter Routes Using Geographic Information Systems – Implications for Bicycle Planning." *Transportation Research Record* 1578:102–110.
- Raith, A., and M. Ehrgott. 2009. "A comparison of solution strategies for biobjective shortest path problems." *Computers & Operations Research* 36:1299–1331.
- Raith, A., C. Van Houtte, J.Y.T. Wang, and M. Ehrgott. 2009. "Applying Biobjective Shortest Path Methods to Model Cycle Route-choice." *ATRF 2009 Proceedings*.
- Stinson, M.A., and C.R. Bhat. 2003. "Commuter Bicyclist Route Choice – Analysis Using a Stated Preference Survey." *Transportation Research Record* 1828:107–115.

Optimization of a Single Ambulance Move up

Lei Zhang

Andrew Mason

Andy Philpott

Department of Engineering Science

University of Auckland

New Zealand

Emergency Medical Services (EMS) are the organizations which provide on-scene medical care and transport to healthcare facilities. A common performance measure is the coverage achieved by the ambulances. A zone is considered covered if an ambulance can reach the zone centroid within a target response time (e.g., 9 minutes). With the rapid increase in call volume, the rising costs of medical equipment and the worsen traffic conditions, it is a great challenge for the EMS to achieve satisfactory performance goals and therefore efficiently utilizing the limited resources is of great importance to the EMS.

There are two main streams of research on improving the performance of EMS. The first stream is the static siting problem. Early models such as (Church and ReVelle 1974) are deterministic models in which ambulances are assumed to be always available. The hypercube (Larson 1974) and approximate hypercube models (Larson 1975) are developed to calculate the probability of unavailability of each mobile server. Subsequently, a few models such as (Susan, Armann, and Erhan 2008) seek to embed the (approximate) hypercube model into their optimization framework in order to give more realistic solutions.

The second stream of research is the dynamic relocation problem (i.e., move up). Idle ambulances are relocated dynamically in real time in order to compensate for the loss of service due to vehicles becoming unavailable in busy areas. Integer Programming models (Richards 2007) and (Gendreau, Laporte, and Semet 2001) use a score function based on the number of vehicles covering each zone to give a move up strategy. An approximate dynamic programming approach (Henderson, Topaloglu, and Restrepo 2006) uses an approximate value function which is trained using simulations to move up idle ambulances.

Our research falls into the second stream—move up problems. We first formulate a dynamic programming model to move up an single ambulance in order to maximize the probability of reaching the next call on time. Examples of moving up an ambulance on a line and on a small network are used to give insights. We also present a node label setting algorithm to reduce the solution time for large problems. We then formulate a dynamic programming model to move up an single ambulance in order to maximize the average number of calls reached on time in an infinite horizon. Examples of moving up an ambulance on a line are used to give insights. We also show a variable substitution method to reduce the state space in order to solve large problems.

Finally we show the impact of a time-dependant spatial distribution of call arrivals on the move up policy for the infinite horizon problem.

References

- Church, R.I., and C.S. ReVelle. 1974. "The maximal covering location problem." *Papers of the Regional Science Association* 32:101–118.
- Gendreau, M., G. Laporte, and S. Semet. 2001. "A dynamic model and parallel tabu search heuristic for real time ambulance relocation." *Paralell Computing* 27:1641–1653.
- Henderson, S.G., H. Topaloglu, and M. Restrepo. 2006. "Approximate Dynamic Programming for Ambulance Redeployment." *Technical Report*.
- Larson, R.C. 1974. "A hypercube queueing model for facility location and re-districting in urban emergency services." *Computers and Operations Research* 1:67–75.
- . 1975. "Approximating the performance of urban emergency service systems." *Operations Research* 23:845–868.
- Richards, D.P. 2007. "Optimised ambulance redeployment strategies." *Department of Engineering Science, The University of Auckland*.
- Susan, B., I. Armann, and E. Erhan. 2008. "Optimal ambulance location with random delays and travel times." *Health Care Management Science* 11:262–274.

Trip Assignment under Energy and Environmental Constraints

Kenneth D. Kuhn

Department of Civil and Natural Resources Engineering
University of Canterbury
Christchurch, New Zealand
kenneth.kuhn@canterbury.ac.nz

Abstract

Limits on available energy and allowable environmental impacts may soon restrict transportation systems. Relatively little work has been done to investigate how this will impact the techniques used in the field of transportation engineering. This research considers one such technique, all-or-nothing trip assignment. A process that today involves solving thousands of shortest path problems may involve solving thousands of constrained shortest path problems in the future. This research examines the impact such a shift will have on computational burdens.

The results of computational studies involving energy-constrained trip assignment are presented here. It is found that solving constrained shortest path problems can take several orders of magnitude more time than solving traditional shortest path problems. This is worrying given that such problems will have to be solved large numbers of times in order to use one of the simplest techniques of transportation engineering. A specialized algorithm (Carlyle and Wood 2003) typically outperforms a generic solver, but occasionally takes an excessively long amount of time to select a path. The results indicate that it may become increasingly important for transportation engineers to be well versed in optimization.

Keywords: energy-constrained transportation, constrained shortest path problem, trip assignment, transportation planning.

1 Introduction

Research suggests that constraints on energy use and environmental impacts may soon restrict transportation. This relates to both individual vehicles, which may have reduced range, and whole systems, which may face new local, regional, or global restrictions. Transportation engineering as a discipline is, by and large, unprepared for a future where energy and environmental constraints take on increased prominence (Dantas et al. 2005).

This research investigates how energy and environmental constraints could be incorporated in a relatively simple fashion into one technique in transportation engineering, all-or-nothing trip assignment. Essentially, a process that involves solving thousands of shortest path problems would be replaced by a process that involves solving thousands of constrained shortest path problems. The implications of such a change, in terms of computational burden, are analyzed here.

The next section describes why energy and environmental concerns may seriously constrain transportation in the future. The following section defines trip assignment in the presence of energy and environmental constraints. Algorithms for the constrained shortest path problem are described next, including an efficient algorithm presented at a previous ORSNZ conference (Carlyle and Wood 2003). Computational studies are introduced and conclusions drawn in subsequent sections of this paper.

2 Transportation in a constrained future

The overwhelming majority of vehicles in use today are powered by burning petroleum-based fuels. A number of researchers are predicting serious declines in the availability of petroleum and/or a transition to alternate fuels in the next ten to twenty years (Farrell and Brandt 2006). At the same time, significant environmental concerns are being raised. For example, one study found that child asthma prevalence rates were associated with local rates of carbon monoxide and nitrogen oxides production from automobile traffic (Guo et al. 1999). Other studies have indicated that the burning of fossil fuels is causing dangerous and irreversible climate change (see www.ipcc.ch).

A significant and growing number of vehicles are being powered by fuels extracted from biological materials. The production and consumption of biofuels releases pollutants at a lower rate than that of petroleum, and carbon is stored in materials as they are grown to make biofuels. However, the widespread use of biofuels globally would require dramatic land use changes and new distributional systems that may not be feasible. Many biofuels are produced on land that could otherwise be used for food production, setting up a *food vs. fuel* dynamic that could constrain production of both. Finally, evidence indicates that the net result of a large-scale switch from petroleum to biofuels, when incorporating land use changes, may be a dramatic *increase* in pollutant emissions (Searchinger et al. 2008).

It now looks likely that many transportation systems of the not-too-distant future will be largely electric. When power is produced in wind, solar, hydroelectric, tidal, or geothermal power plants, zero harmful pollutants are emitted. Unfortunately, current global electricity production is insufficient to meet current demands and power transportation. Electricity produced by renewable resources is particularly limited. Many sustainable modes of electric energy production depend on factors beyond our control. One effort to design a bus route in Christchurch, New Zealand using sustainable resources concluded that “no amount of investment in wind and solar energy capacity can provide the same service as the fossil fuel system” (Dantas et al. 2005). Pumped-storage hydroelectricity could provide zero-emissions energy when wind and sun are not strong enough. However, the land area and funds required to establish a wind, solar and pumped-storage hydroelectricity system capable of

providing the continuous power required by just one bus route (operating in a manner consistent with current practice) are impractical (Ibid.). On a smaller scale, personal electric vehicles will likely be powered by batteries and have a range substantially lower than current vehicles.

Summarizing the points listed above, research suggests that transportation systems and individual vehicles may be significantly more constrained in terms of energy use and environmental impacts in the future. This would clearly have substantial implications for the practice of transportation engineering. Yet little has been written on this point. In the words of one of the articles cited above, “a survey of transportation engineering texts illustrates that current modelling and planning techniques do not include any method to consider constraints in natural resources, emissions, or, most importantly, energy” (Dantas et al. 2005)

3 Trip assignment

Transportation engineering often requires forecasting vehicle loads on different sections of a transportation network. One example would be estimating the traffic impacts of the opening of a new shopping centre. Another example would be planning the future maintenance requirements of an airport taxiway. The most commonly used method for forecasting vehicle loads is known as the *four-step method*, the roots of which can be traced back to the seminal Chicago Area Transportation Study (CATS 1959). First, *trip generation* estimates the numbers of trips that will depart from different origins as well as the numbers of trips that will arrive at different destinations. Next, *trip distribution* links origins and destinations, forecasting the numbers of trips between location pairs. Where relevant, *mode choice* breaks down trip counts by mode of transportation. Finally, *trip assignment* forecasts the numbers of vehicles that will travel on individual sections of a given transportation network, based on the preceding analyses. Trip assignment is one of the most researched areas of transportation engineering, and one of the most important to practitioners.

The simplest form of trip assignment follows what is known as the *all-or-nothing approach*. This technique assigns fixed travel times to different sections of the transportation network. All trips between given origins and destinations are then assigned to the sections of the network that allow the trips to be completed in the shortest possible time. Travel times can be replaced with generalized cost estimates without substantively altering the process. The all-or-nothing approach does not consider congestion, the idea that travel times (or costs) increase as the number of vehicles on the relevant section of the network increase. However, the all-or-nothing approach remains useful when congestion is relatively unimportant or when estimating the costs of congestion.

The research presented here considers trip assignment using the all-or-nothing approach in scenarios involving energy and environmental constraints. The assumption is made that vehicles using individual sections of a transport network consume fixed amounts of various *resources*. Resource consumption is additive across the sections of a transport network. There are budgets of the various resources available, and a vehicle may not be assigned to a route that would require it to consume

more of any resource than the available budget. This is a relatively simple way to imagine integrating energy and environmental constraints into trip assignment. Essentially, we will be solving large numbers of constrained shortest path problems. Mathematical details are presented below.

Let a transportation network consisting of a set V of vertices and another set E of edges be given. Each element in E will consist of a start and an end vertex. Let o and d be the origin and destination, respectively, of the trip currently being assigned. The parameter c_{ij} is the generalized cost associated with the edge (i, j) . Let R be the set of resources constraining us. The parameters b^r and q_{ij}^r are the budget (available amount) of resource r and the quantity of resource r used when travelling on the edge (i, j) respectively. The decision variable α_{ij} is binary and is to be set to 1 if we travel on edge (i, j) and 0 otherwise. The nominal formulation of the problem of interest, for one trip, follows.

Model Nominal formulation

$$\min \sum_{(i,j) \in E} c_{ij} \alpha_{ij} \tag{1}$$

s.t.

$$\sum_{i:(i,j) \in E} \alpha_{ij} - \sum_{k:(j,k) \in E} \alpha_{jk} = \begin{cases} -1 & j = o \\ 0 & \forall j \in V \setminus \{o, d\} \\ 1 & j = d \end{cases} \tag{2}$$

$$\sum_{(i,j) \in E} q_{ij}^r \alpha_{ij} \leq b^r \quad \forall r \in R \tag{3}$$

$$\alpha_{ij} \in \{0, 1\} \quad \forall (i, j) \in E \tag{4}$$

The objective function, (1), minimizes the total generalized cost. Constraint set (2) ensures that the edges chosen represent a logical path, departing only from the origin and arriving only at the destination. Constraint set (3) captures energy and environmental constraints. Finally, constraint set (4) ensures decision variables take on values consistent with the desired interpretation.

4 Solution techniques

The formulation presented above is well known in operational research and has been labelled the *constrained shortest path problem* (Irnich and Desaulniers 2004). This problem can be computationally challenging, and is NP hard (Garey and Johnson 1979). The constrained shortest path problem is a special case of the *shortest path problem with resource constraints*, which is often solved via labelling algorithms (Irnich and Desaulniers 2004). Indeed, a specialized labelling algorithm has been proposed for the problem studied here (Dumitrescu and Boland 2003). Efficient techniques based on Lagrangian relaxation have also been proposed (Handler and Zang 1980; Carlyle and Wood 2003). A number of researchers have proposed special techniques applicable when there is only one resource constraint (Handler and Zang 1980; Santos, Coutinho-Rodrigues, and Current 2007). This work considers two approaches for solving the identified problem. One approach encodes the nom-

inal formulation shown above and asks the generic *glpk* solver to find the optimal solution. An alternate approach uses a Lagrangian relaxation and enumeration algorithm presented at an earlier ORSNZ conference (Carlyle and Wood 2003) to find a solution. The two approaches were used to see the ranges of solution times we might expect depending upon our choice of solution technique.

The Carlyle and Wood approach begins by finding optimal, or near-optimal values for Lagrange multipliers associated with resource constraints. A formulation for the Lagrangian relaxation is as follows.

Model Lagrangian relaxation

$$\max_{\lambda \geq 0} \min_{\alpha} \sum_{(i,j) \in E} (c_{ij} + \sum_{r \in R} \lambda_r q_{ij}^r) \alpha_{ij} - \sum_{r \in R} \lambda_r b^r \quad (5)$$

s.t.

$$\sum_{i:(i,j) \in E} \alpha_{ij} - \sum_{k:(j,k) \in E} \alpha_{jk} = \begin{cases} -1 & j = o \\ 0 & \forall j \in V \setminus \{o, d\} \\ 1 & j = d \end{cases} \quad (6)$$

$$\alpha_{ij} \in \{0, 1\} \quad \forall (i, j) \in E \quad (7)$$

Note that the inner problem of finding values for α is a traditional shortest path problem where the cost of taking edge (i, j) is $c_{ij} + \sum_{r \in R} \lambda_r q_{ij}^r$. The outer problem of finding values for λ can be solved using relatively simple techniques making use of solutions of the inner problem. Such techniques include subgradient optimization, or bisection search when there is only a single resource constraint. Note that it is possible to begin such techniques by setting all λ terms equal to zero and solving the resulting shortest path problem for α . In other words, the first step for solving the constrained shortest path problem is to solve the (unconstrained) shortest path problem. This is helpful in comparing the two problems.

Once optimal, or near-optimal, values for the Lagrange multipliers have been found, the Carlyle and Wood approach enumerates paths that are feasible in terms of resource budgets, as well as being near-optimal in terms of the nominal and Lagrangian objective functions. The algorithm can be set up to finish when a solution is found within a set distance of a lower bound previously identified. A complete description of the algorithm including pseudo-code has been presented at this conference previously (Carlyle and Wood 2003). The referenced text claims that the proposed approach was able to solve constrained shortest path problems “an order of magnitude faster” than arguably the most promising alternate approach (Ibid.).

5 Computational studies

The results of computational studies are presented here. In these computational studies we considered only a single resource constraint on energy use. Considering a single constraint keeps analysis simple, although the techniques used in this study could easily be applied in situations involving multiple constraints. 1,000 vehicles were assigned paths from an origin to a destination. In each case, 1,000 vertices were

set up in a 4 by 250 grid. Edges linked vertices adjacent to one another vertically or horizontally. Edges went up and down vertically, but only to the right horizontally. In total there were 2,496 edges. The vertex at the top left was chosen to be the origin and the vertex at the bottom right was chosen to be the destination. The generalized costs and energy use associated with individual edges were independent identically distributed random variables uniformly distributed between 0 and 1. It is unlikely that costs and energy use would be independent in real life, but the assumption of independence simplifies problem parameterization. The cost and energy use data were randomized each time a new vehicle was to be assigned a path. Other researchers have used vertex grids similar to the one used here to test constrained shortest path algorithms (Carlyle and Wood 2003; Dumitrescu and Boland 2003). The grid used here is somewhat unique in that one dimension is significantly longer than the other. It is believed that this reflects many transportation path planning problems, where there are a handful of parallel routes to take and the decision boils down to when to switch between routes.

Vehicles were only allowed to use 155 units of energy to get from the origin to the destination. Vehicles had to transverse at least 252 edges to get from the origin to the destination, and a shortest cost path could include significantly more edges. The constraint on energy use was binding around half the time. Paths proposed by a simple shortest path algorithm (ignoring the energy constraint), the Carlyle-Wood algorithm, and the glpk solver were saved. The Carlyle-Wood algorithm was set up to quit and return the best feasible path found if the optimality gap was reduced to 0.1 units of cost or if 600 seconds had elapsed. There were cases where 600 seconds elapsed and a solution not proven optimal or near-optimal was returned (more on this later). Figure 1 shows histograms of energy use and generalized costs on paths found by the various algorithms.

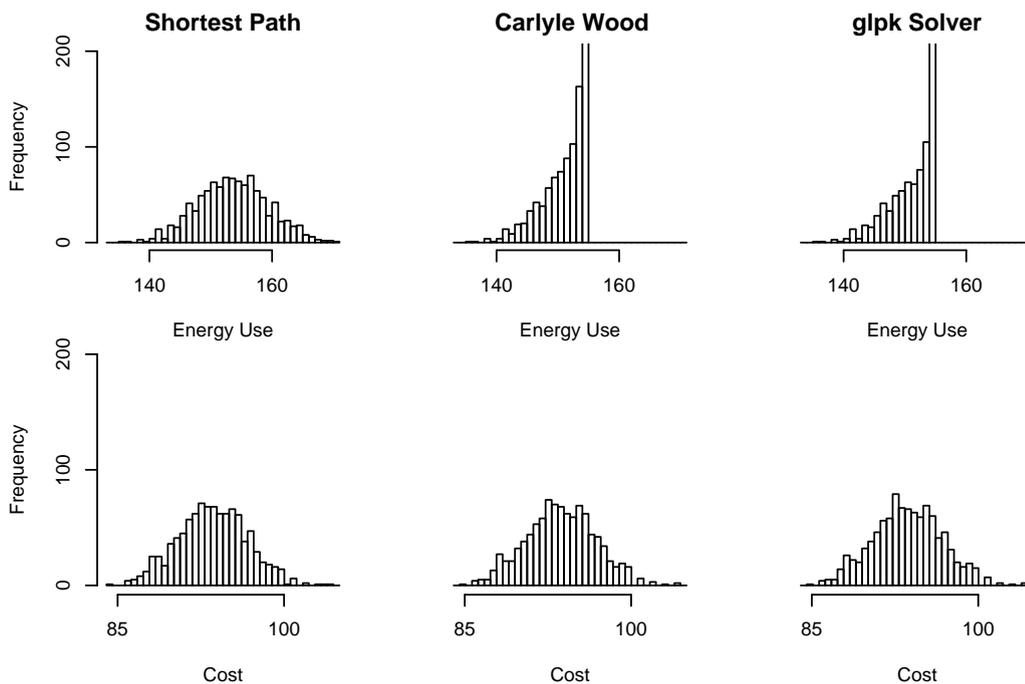


Figure 1: Energy use and generalized costs associated with chosen paths.

The Carlyle-Wood algorithm and glpk solver produced close to identical results. Each was able to limit energy use to below 155 units for each vehicle without substantially altering the costs of travel. The results reflect the assumption that energy use and cost are independent. Given a good deal of independence, large enough choice sets and good decision making, it is possible to constrain resource use without dramatically increasing costs. This is encouraging, although it seems likely that resource use and costs would be highly correlated in reality. Further research could investigate the link between resource use and generalized travel cost estimates in reality, or look at the sensitivity of simulation results to resource-cost correlation.

Special attention was paid to the solution times of the various algorithms. On the 1,000 test problems, the shortest path algorithm took between 0.01 and 0.02 seconds while the glpk solver took between 27 and 35 seconds. These results highlight how much more complicated the constrained shortest path problem is than the shortest path problem. The Carlyle-Wood algorithm was able to finish after the shortest path problem had been solved in cases where the budget for energy use was not fully used. This meant more than half the time, the Carlyle-Wood algorithm took only around 0.01 seconds to find a vehicle path. Over ninety percent of the time, the Carlyle-Wood algorithm took less than 0.25 seconds to find a solution, significantly outperforming the glpk solver. However, on 28 out of the 1,000 trials, the Carlyle-Wood algorithm had not found a solution after ten minutes and had to be stopped. Some summary statistics regarding the distribution of computation times are presented in Table 1.

	minimum	median	75th perc.	90th	95th	99th
Shortest Path	0.0112	0.0113	0.0113	0.0113	0.0113	0.0193
Carlyle Wood	0.0112	0.0113	0.242	0.243	11.1	600*
glpk Solver	27.2	27.6	27.9	28.3	28.7	35.0

* - Algorithms were forcibly stopped after 600 seconds.

Table 1: Distributions of computation times, in seconds, by algorithm.

It is not clear why the Carlyle-Wood algorithm performed so well most of the time, but so poorly occasionally. The results mirror those of another study investigating the biobjective shortest path problem (Raith and Ehrgott 2009). Raith and Ehrgott conclude that a specialized enumerative algorithm “is a very successful approach to solve some problem instances, but the run-time on others is very long” (Ibid.). The data presented in Figure 1 and an informal analysis show that the Carlyle-Wood algorithm was always able to identify a high quality feasible solution, but not to verify that the solution was optimal or near-optimal.

One of the difficulties associated with the constrained shortest path problem is that it is difficult to obtain good lower bounds on the optimal objective function value. Actually, one of the strengths of the Lagrangian approach is that the relaxed problem provides a reasonable lower bound. Further research investigating ways to dynamically determine stronger lower bounds and incorporate their use into an enumerative algorithm may be worthwhile. It’s worth mentioning that the author of this paper is a casual programmer and there may have been inefficiencies, unknown to the author, in the code used here to test the various algorithms.

Extra long computation times are likely associated with problems where there are large optimality gaps. In such situations, there would also likely be a large separation between the energy used on the shortest cost path and the available energy budget. Figure 2 displays the computation times of the Carlyle-Wood algorithm as a function of the energy used on the shortest cost path. There is some, but not a dramatic amount of, evidence for the hypothesis that computation times increase with energy use on the shortest cost path.

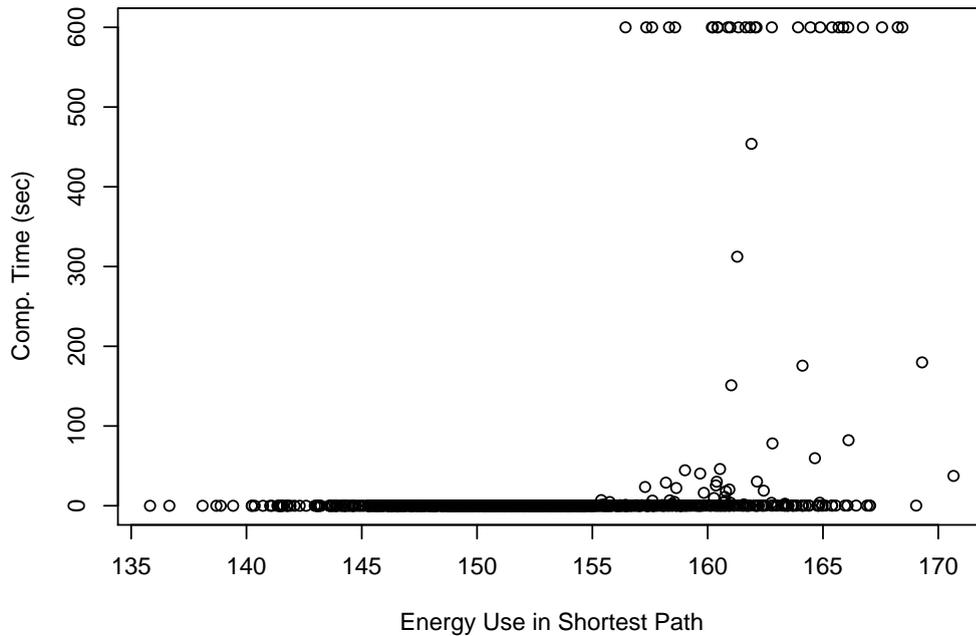


Figure 2: Exploring the performance of the Carlyle-Wood algorithm.

6 Conclusion

This article discussed the possibility of there being significant new energy use or environmental impact constraints on transportation in the next ten to twenty years. Consideration was given to how such constraints would impact techniques used in the field of transportation engineering. One of the simplest of these techniques, all-or-nothing trip assignment, was used as an example. It was hypothesized that what today involves solving thousands of shortest path problems may, in the future, involve solving thousands of constrained shortest path problems. It was found that such a change would increase computation times by more than three orders of magnitude if a generic solver was used to solve constrained shortest path problems. Using a specialized algorithm resulted in computation times that were typically only 20 to 30 times those associated with the nominal shortest path problem, in cases involving binding resource constraints. However, the specialized algorithm chosen occasionally took an exceptionally long amount of time to select an optimal path. Further research is needed to identify what caused these outlying results. Overall, the results indicate that it may become increasingly important for transportation engineers to be well versed in optimization.

Acknowledgments

The author would like to thank Andrea Raith for providing technical guidance regarding shortest path problems. Andrea, Stephan Hassold, and Matthias Ehrgott convinced the author to attend this conference. Susan Krumdieck, Shannon Page, André Dantas, and Judith Wang interested the author in transportation planning for an energy-constrained future. The author wishes to thank those mentioned for several enjoyable discussions.

References

- Carlyle, W.M., and R.K. Wood. 2003. "Lagrangian relaxation and enumeration for solving constrained shortest-path problems." *Proceedings of the 38th Annual ORSNZ Conference*.
- CATS. 1959. *Chicago Area Transportation Study*. Volume 1. Chicago.
- Dantas, A., S. Krumdieck, A. Hamn, M. Saunders, and S. Minges. 2005. "Performance-objective design for energy constrained transportation system." *Journal of the Eastern Asia Society for Transportation Studies* 6:3276–3292.
- Dumitrescu, I., and N. Boland. 2003. "Improved preprocessing, labeling and scaling algorithm for the weight-constrained shortest path problem." *Networks* 42:135–153.
- Farrell, A.E., and A.R. Brandt. 2006. "Risks of the oil transition." *Environmental Research Letters* 1:1–8.
- Garey, M.R., and D.S. Johnson. 1979. *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco: Freeman.
- Guo, Y.L., Y.C. Lin, F.C. Sung, S.L. Huang, Y.C. Ko, J.S. Lai, H.J. Su, C.K. Shaw, R.S. Lin, and D.W. Dockery. 1999. "Climate, traffic-related air pollutants, and asthma prevalence in middle school children in Taiwan." *Environmental Health Perspectives* 107:1001–1006.
- Handler, G.Y., and I. Zang. 1980. "A dual algorithm for the constrained shortest path problem." *Networks* 10:293–310.
- Irnich, S., and G. Desaulniers. 2004. "Shortest Path Problems with Resource Constraints." *Les Cahiers du GERAD*.
- Raith, A., and M. Ehrgott. 2009. "A comparison of solution strategies for biobjective shortest path problems." *Computers & Operations Research* 36:1299–1331.
- Santos, L., J. Coutinho-Rodrigues, and J.R. Current. 2007. "An improved solution algorithm for the constrained shortest path problem." *Transportation Research Part B* 41:756–771.
- Searchinger, T., R. Heimlich, R.A. Houghton, F. Dong, A. Elobeid, J. Fabiosa, S. Tokgoz, D. Hayes, and T.H. Yu. 2008. "Use of U.S. croplands for biofuels increases greenhouse gases through emissions from land-use changes." *Science* 319:1238–1240.

The Performance Evaluation of Turkey's Export to Ireland

A.Ulgen Ozgul, S.Sebnem Ahiska
Industrial Engineering Department, Galatasaray University
Ciragan Cad. No:36 34357 Ortakoy, Istanbul, Turkey

Abstract

This study aims to evaluate the performance of Turkey's export to Ireland in 2008 by focusing on the manufacturing sectors of the top 100 products exported. A robust evaluation of the export performance requires the consideration of several quantitative criteria as well as qualitative criteria, and therefore, as the solution methodology, we employ an imprecise data envelopment analysis (IDEA) model, which is a mathematical programming based multi-criteria decision making (MCDM) model that can successfully deal with imprecise input and output criteria. In this study, through IDEA, the manufacturing sectors considered (namely, Manufactured Food Products and Beverages, Textile Industry, Wearing Apparel, Manufacture of Paper and Paper Products, Plastic and Rubber Products, Manufacture of Non Metallic, Basic Metals Industry, Machinery and Equipment, Electric Machinery and Apparatus, Manufacture of Motor, Vehicles, Trailers, and Manufacture of Furniture) will be ranked according to their export performance, and managerial insights will be provided regarding which sectors should receive more attention in order to increase the overall export performance of Turkey.

Keywords: export performance evaluation, data envelopment analysis, multi-criteria, decision making, imprecise data

1. Introduction

As the current global market trend points out the role of exporting in a nation's economy, the interests in export performance and the actions needed in order to improve the export efficiency are getting more and more important each day.

The studies concentrated on Turkish Economy show that the export performance is a very important topic for Turkey and according to the strategic plans made for the upcoming years, its importance will continue to grow in the future. According to TURKSTAT (Turkish Statistical Institute), the value of Total Export of Turkey in 2008 is 132 billion dollars, which reflects 23.1% growth compared to 2007. When the export origins investigated, it is worth noting that 94.8% of Turkish exports are made by the manufacturing industry, which motivated us to focus specifically on the export performance evaluation of the manufacturing sectors in this study.

Ireland, despite being a small country with the estimated population of 4,422,100, was listed among the top 50 countries that Turkey made export for the first 9 months of 2008[Turkstat, 2008]. When the top 100 products exported from Turkey to Ireland is analyzed, we see that while the total import of Ireland for these products from different countries decreased by 20% overall in 2008 because of the global crisis, the value of the import made from Turkey had a lower decrease, which is 17%. In the light of this

healthy commercial relationship between Turkey and Ireland, we decided to provide a quantitative analysis that evaluates the export performance of Turkey with Ireland.

This paper will focus on the performance of Turkey for the top 100 manufacturing products (in terms of value) exported to Ireland. Since this analysis requires multiple performance evaluation criteria that can be quantitative as well as qualitative, we employ an imprecise data envelopment analysis (IDEA) model, which is a mathematical programming based multi-criteria decision making model that can deal with imprecise as well as exact data simultaneously.

There exist some studies in the literature regarding a firm's export performance evaluation by using an MCDM methodology. However, to our knowledge, there is no study that evaluates the performance of Turkey's export at country level using real data and employing an MCDM method. Hence, we believe that this study which evaluates the performance of Turkey's export to Ireland (a target market) using an MCDM technique, namely IDEA, applied to real data gathered from reliable sources, is an important contribution to literature.

This paper is organized as follows. Section 2 provides a brief literature survey related to the export performance evaluation problem as well as the data envelopment analysis (DEA) models. In Section 3, the problem of the export performance evaluation of Turkey to Ireland is presented. Section 4 describes the IDEA methodology employed. In section 5, the results of our study are given, and managerial insights are provided. Finally, concluding remarks are made in section 6.

2. Literature review

2.1 Literature Review on Export Performance Evaluation Criteria

As the importance of the export increases at the national economy, export performance evaluation becomes important especially at the international marketing-business areas. For many years the researchers from economics or marketing disciplines have tried to identify the best criteria to take into consideration. However, a fully consensus hasn't been settled yet (Sousa, 2004).

Researchers emphasize the usage of multiple criteria for the evaluation of export performance. Al-Khalifa & Morgan (1995) and Zou & Stan (1998) try to obtain a consensus related with the export performance input and output measurements. However, even though they both propose some financial outputs as the main export performance criteria, such as; percentage of total sales, export growth level and export profitability, they both propose different additional measurement criteria for their analysis. Al-Khalifa et al. focuses on competitor centered and custom focused output measurement criteria, whereas Zou et al. deals with a wide range of objective and subjective input and output measures.

Even though most of the studies provide the relevant criteria for a firm's export performance evaluation, the criteria related with the products, markets or countries are not explicitly provided. (Zou & Stan, 1998) Only a few studies related with this issue can be found in the literature (Thirkell & Dau,1998). Sousa (2004) provides a more detailed list compared to Zou and some of these additional criteria were also used at the study of Katsikeas, Leonidou & Morgan (2000).

When Sousa's and Katsikeas et al.'s studies are examined, it is seen that they have different criteria classification proposals. Sousa is classifying criteria into two groups: measures, which are considered as the dependent variables (outputs), and determinants,

which are considered as the independent variables (inputs). However Katsikeas et al. forms three groups of criteria which are defined as; measures, which are considered as the dependent variables (outputs), intervening variables, which are independent variables that affect directly the dependent variables (inputs) such as; targeting factors and marketing strategy factors and background variables, which are independent variables that have indirect affects on dependent variables (inputs) such as; managerial factors, organizational factors, environmental factors.

2.2 Literature Review on Data Envelopment Analysis (DEA)

DEA is a linear programming based methodology formally developed by Charnes, Coopers and Rhodes in 1978 building on the ideas of Farrell (1957). The DEA model is a multi-factor productivity analysis model for measuring the relative efficiencies of a homogeneous set of decision making units (DMUs). The efficiency score in the presence of multiple input and output factors is defined as: $\text{Efficiency} = (\text{Weighted sum of exact outputs} / \text{Weighted sum of exact inputs})$ (Talluri, 2000). The proposed measure of the efficiency of any DMU is obtained as the maximum of a ratio weighted outputs to weighted inputs subject to the condition that the similar ratios for every DMU be less than or equal to unity (Charnes, Coopers and Rhodes, 1978). Since each DMU is trying to maximize its own efficiency score in CCR DEA model, because of having weight flexibility and due to the structure of the method which makes us to solve the problem for each DMU separately, sometimes the optimal weight outcomes can be out of logic and that's why sometimes the inefficient ones can be better in overall performance in the real life (Talluri, 2000).

Several approaches have been proposed to overcome this problem: one approach is to detect the efficient DMUs using the DEA model and then, apply other MCDM techniques such as TOPSIS or AHP to provide ranking between these efficient units. Other approaches include the use of objective functions that have a better discriminating power among DMUs such as the minimax efficiency objective (i.e. the minimization of the maximum deviation from the ideal efficiency score 1), the addition of constraints into the model such as weight restriction constraints, and the cross efficiency analysis.

Cross efficiency analysis is developed by Sexton in 1986. In this analysis, the performance of each DMU is evaluated with respect to the optimal input and output weights of other DMUs, and the efficiency scores obtained this way are presented in a matrix named "Cross Efficiency Matrix" (CEM). A weakness of this method can outcome when weights aren't unique. This makes it difficult to distinguish the good and bad performers for the problem set. In order to overcome this problem, Doyle and Green developed a technique in 1994 by introducing an aggressive formulation. This technique tries to find the optimal weights that maximize the efficiency of the DMU under consideration while minimizing the average efficiency of other DMUs. Moreover, this technique provides a full ranking of the DMUs by indicating as well the false positive ones.

In addition to these developments, with the help of the procedure developed by Cook et al. (1993, 1996), DEA is extended to deal with imprecise data in addition to exact data.

There are also some other techniques developed for specific types of problems. Ahiska and Karsak (2005) propose a model for the single input, multi output models. By this model, they are suggesting to use efficiency measures that aren't specific to a particular DMU but common to all. To do that first the problem is rewritten in terms of

deviation between the ideal efficiency ($=1$) and efficiency obtained. Then in order to obtain a solution wrt common weights, minimization of max deviation will be used as objective function.

3. Problem Description

We consider here the problem of the performance evaluation of Turkey's export to Ireland in year 2008. We employ an imprecise DEA methodology to rank the manufacturing sectors considering both exact and ordinal real data sets.

For this analysis, the top 100 products exported to Ireland are categorized into 11 sectors by the help of an expert. The 11 Sectors that will be considered as DMUs in our IDEA analysis are reported below at Table 1:

Table 1-DMU List of the Problem

DMUs	Manufacturing Sectors
DMU 1	Manufactured Food Products and Beverages
DMU 2	Textile Industry
DMU 3	Wearing Apparel
DMU 4	Manufacture of Paper and Paper Products
DMU 5	Plastic and Rubber Products
DMU 6	Manufacture of Non Metallic
DMU 7	Basic Metals Industry
DMU 8	Machinery and Equipment
DMU 9	Electric Machinery and Apparatus
DMU10	Manufacture of Motor, Vehicles, Trailers
DMU11	Manufacture of Furniture

As a result of our literature survey on export performance evaluation criteria, we had a very long criteria list, but we had to eliminate most among them due to lack of real data, which is a common problem for most MCDM problems.

After we determined the relevant criteria to consider, we classified them into two groups: inputs and outputs. Further, these inputs and outputs are classified according to the type of data collected as objective and subjective. In order to obtain the data related with subjective criteria, we formed a survey by using 5-Likert scale (where score of 5 represents "the best" and score of 1 "the worst"), and received the expert evaluations as ordinal data. In order to collect the data regarding the quantitative criteria, we used TURKSTAT, CSO (Central Statistics Office Ireland), and DTM (Undersecretariat of the Prime Ministry for Foreign Trade) as the data sources.

The input and output criteria employed in our study are the following:

Objective Outputs

*Export Share: This criterion is calculated as the ratio of sectoral export to Ireland in 2008/ total sectoral export of Turkey in 2008 for the top 100 products we consider (Sousa, 2004).

* Export Sales Volume: Exports in terms of monetary value.

* Import Share: This criterion is calculated as sectoral export of Turkey to Ireland in 2008/total sectoral import of Ireland in 2008 for the top 100 products we consider (Al-Khalifa & Morgan, 1995)

Subjective Outputs

* Overall Export Performance: (Sousa, 2004)

* Strategic Export Performance: This criterion is used for measuring compliance of the exports realized with Turkey's foreign trade policy for year 2008 (Sousa, 2004)

* Contribution of exporting to the growth of the country: (Sousa, 2004)

*Goal Achievement: This criterion is used for measuring compliance of exports with the pre-determined export targets to Ireland for year 2008 (Zou & Stan, 1998)

*Contribution of exporting to the country reputation: (Katsikeas, Leonidou & Morgan, 2000).

Objective Inputs

*ULC ("The cost of labor required to produce one unit of output in a particular industry, sector or total economy"): This criterion is accepted to be one of the major indicators of the international competitiveness calculated as: Wage per hour worked (euro/hour)/labor productivity (Aysan & Hacıhasanoglu, 2007). Since Labor Productivity is no longer calculated by Central Bank of Republic Turkey since 2006, we calculated this parameter using the instructions given at their webpage. (Labor Productivity: Indices of Partial Productivity of Production Workers per Capita and per Hours Worked at Production in Manufacturing Industry= Quarterly Industrial Production Index/ index of hours worked at production in manufacturing industry)

* Hours Spent: This data is supplied from our survey in terms of precise data.

* Import Growth of Ireland(%): This criterion is calculated in terms of % change of the import of Ireland between 2007-2008 for the top 100 products exported

Then we formed two different data sets one of which is composed of just quantitative input and outputs, the other is composed of all input and outputs including subjective data in order to see the effect of considering the subjective criteria to the efficiency of the analyzed sectors.

4. Proposed methodology

While there are many variations of the DEA methodology, we use as our starting point the original model for the simple efficiency calculations in the existence of exact data. The basic linear DEA model, which is developed by Charnes et al.(1978), is known as the CCR DEA model, and is formulated as follows:

$$\begin{aligned} \max E_{j_0} &= \sum_r u_r y_{rj_0} \\ \text{st} \quad & \sum_i v_i x_{ij_0} = 1 \\ & \sum_r u_r y_{rj} - \sum_i v_i x_{ij} \leq 0, \forall j \\ & u_r \geq \varepsilon, \forall r \\ & v_i \geq \varepsilon, \forall i \end{aligned}$$

where E_{j_0} is the efficiency value of the evaluated decision making unit, DMU j_0 ; u_r and v_i are the weights assigned to output r and input i , respectively; y_{rj} is the amount of output r produced by DMU j ; x_{ij} is the amount of input i consumed by DMU j .and ε is a very small positive number used to assure the positivity of multipliers μ_r and ω_i in order to avoid neglecting any of inputs or outputs under consideration.

For determining the relative efficiency values of all DMUs, the above model is solved separately for each DMU j_0 , $j_0 = 1, 2, \dots, 11$.

The above model can only deal with exact data. Hence, for our data set which contains ordinal data in addition to exact one, we employed the IDEA model introduced by Cook et al. The IDEA model that fits the data types of our problem is formulated below:

$$\max h_0 = \sum_{r \in EXO} u_r y_{rj_0} + \sum_{r \in ORDO} \mathbf{w}_r^1 \gamma_{rj_0}$$

st

$$\sum_{i \in EXI} v_i x_{ij_0} = 1$$

$$\sum_{r \in EXO} u_r y_{rj} + \sum_{r \in ORDO} \mathbf{w}_r^1 \gamma_{rj} - \sum_{i \in EXI} v_i x_{ij} \leq 0$$

$$u_r \geq \varepsilon \quad \text{for } r \in EXO$$

$$v_i \geq \varepsilon \quad \text{for } i \in EXI$$

$$\mathbf{w}_r^1 \in \Psi = \{w_{r,l+1}^1 - w_{r,l}^1 \geq \varepsilon, w_{r,1}^1 \geq \varepsilon, \quad \text{for } l = 1, 2, \dots, L-1; r \in ORDO\}$$

where u_r and v_i are the importance weight variables assigned to output r and input i , respectively; \mathbf{w}_r^1 is the worth vector including variables $w_{r,l}^1$, which indicate the worth (weighted values) of being rated in the l th place with respect to output r ; L is the size of the likert scale; γ_{rj} is the vector indicating the rating assigned to DMU j with respect to output r ; y_{rj} is the amount of output r produced by DMU j ; x_{ij} is the amount of input i consumed by DMU j ; and ε is a positive constant. Finally, EXO, ORDO is the set of exact and ordinal outputs, respectively. Similarly, EXI is the set of exact input.

Our calculations for both exact and exact + ordinal data sets show us that even though we can eliminate some of the DMUs by simple efficiency calculation models, in order to obtain a ranking and find the most efficient sector, we need to apply models with more discriminating power, such as the minimax efficiency model formulated below:

$$\min M$$

st

$$M \geq d_j$$

$$\sum_{i \in EXI} v_i x_{ij_0} = 1$$

$$\sum_{r \in EXO} u_r y_{rj} + \sum_{r \in ORDO} \mathbf{w}_r^1 \gamma_{rj} - \sum_{i \in EXI} v_i x_{ij} + d_j = 0$$

$$u_r \geq \varepsilon \quad \text{for } r \in EXO$$

$$v_i \geq \varepsilon \quad \text{for } i \in EXI$$

$$\mathbf{w}_r^1 \in \Psi = \{w_{r,l+1}^1 - w_{r,l}^1 \geq \varepsilon, w_{r,1}^1 \geq \varepsilon, \quad \text{for } l = 1, 2, \dots, L-1; r \in ORDO\}$$

$$d_j \geq 0$$

where d_j is considered as the deviation from efficiency for DMU j and M is the maximum deviation and the minimax efficiency score of DMU j_0 is calculated by $1 - d_{j_0}$.

Our results showed that the minimax model was not sufficient to fully rank the DMUs. Hence, we apply aggressive cross efficiency analysis which helped us obtain the

results provided in the next chapter. Even though it requires solving $2n$ formulations as opposed to the minimax efficiency model that only requires solving n formulations, it provides a full ranking of the DMUs, and indicates as well the false efficient ones. The aggressive cross efficiency formulation that deals with both exact and ordinal data is formulated below:

$$\begin{aligned}
 & \min \sum_{r \in EXO} u_{rk} \sum_{j \neq k} y_{rj} + \sum_{r \in ORDO} w_{rk}^1 \sum_{j \neq k} \gamma_{rj} \\
 & \text{st} \\
 & \sum_i v_{ik} \sum_{j \neq k} x_{ij} = 1 \\
 & \sum_{r \in EXO} u_{rk} y_{rk} + \sum_{r \in ORDO} w_{rk}^1 \gamma_{rk} - E_{kk} \sum_{i \in EXI} v_{ik} x_{ik} = 0 \\
 & \sum_{r \in EXO} u_{rk} y_{rj} + \sum_{r \in ORDO} w_{rk}^1 \gamma_{rj} - \sum_{i \in EXI} v_{ik} x_{ij} = 0, \quad \forall j \neq k \\
 & u_{rk} \geq \varepsilon, \quad \forall r \text{ for } r \in EXO \\
 & v_{ik} \geq \varepsilon, \quad \forall i \text{ for } i \in EXI \\
 & w_{rk}^1 \in \Psi = \{w_{rl+1}^1 - w_{rl}^1 \geq \varepsilon, w_{r1}^1 \geq \varepsilon, \text{ for } l = 1, 2, \dots, L-1; r \in ORDO\}
 \end{aligned}$$

where E_{kj} is the cross efficiency value of DMU j with respect to DMU k , u_{rk} is the optimal weight assigned to output r for DMU k and v_{ik} is the optimal assigned to input i for DMU k , w_{rk}^1 is the worth vector including variables w_{rl}^1 , which indicate the worth (weighted values) of being rated in the l th place assigned to output r for DMU k .

The aggressive cross efficiency analysis has two main steps: First step is to obtain the simple efficiency values ($E_{kk}, k = 1, 2, \dots, n$) using the CCR-DEA or Cook et al.'s models. Second step is to use the aggressive cross efficiency model. The objective of this model is to obtain the efficiency scores of each DMU calculated using the optimal weights of other DMUs.

After finding all E_{kj} , the formula given below is used in order to find the mean cross efficiency values of the DMUs. The mean cross efficiency value e_j is a measure of how the DMU j is evaluated by other DMUs.

$$e_j = \frac{\sum_{j \neq k} E_{kj}}{n-1}, \quad \forall j$$

Since the mean cross efficiency value is obtained by taking average, it can't distinguish the good overall performers from the poor performers (false positives). To detect these poor performers and measure the degree of their false positiveness in an effective manner, Doyle and Green suggest Maverick index to be calculated. The index is formulated as follows:

$$M_j = \frac{E_{jj} - e_j}{e_j} \times 100, \quad \forall j$$

where M_j is the Maverick index of DMU j . Higher Maverick index indicate higher degree of false positiveness for the DMU under consideration.

5. Results

After the data of the each criteria normalized using max value normalization, we applied the CCR-DEA to the exact only data set whereas Cook et al.'s model to the exact and ordinal data set in order to obtain the simple efficiency values of the sectors.

Since our computations show that there are more than one efficient DMUs at both of the data set analysis, we needed to do further analysis. First, we employed the minimax efficiency model. Event though this model helped us eliminate some of the DMUs, it did not provide a full ranking of the DMUs.

On the other hand, by applying the aggressive cross efficiency model, we could identify the best sector with respect to both of the data sets. Further, we were able to recognize the false efficient sectors by the help of Maverick Index.

Table 2-Analysis Results for Exact Data

	CCR-DEA	Minimax	Aggressive Cross	
			Efficiency	Maverick Index
DMU10	1,00	1,00	0,72	39,65
DMU11	1,00	1,00	0,65	53,49
DMU9	1,00	0,95	0,65	53,75
DMU1	1,00	0,92	0,61	63,32
DMU3	1,00	0,80	0,60	65,70
DMU4	1,00	0,63	0,59	69,92
DMU2	0,71	0,58	0,44	61,69
DMU7	0,78	0,51	0,42	88,00
DMU5	0,58	0,49	0,33	78,00
DMU8	0,53	0,47	0,29	84,51
DMU6	0,49	0,43	0,25	92,50

Table 3-Analysis Results for Exact and Ordinal Data

	Cook et al.	Minimax	Aggressive Cross	
	Model		Efficiency	Maverick Index
DMU9	1,00	1,00	0,67	49,25
DMU10	1,00	1,00	0,66	50,92
DMU8	1,00	0,91	0,56	77,12
DMU11	1,00	1,00	0,43	133,64
DMU7	1,00	0,92	0,40	147,89
DMU4	1,00	1,00	0,38	163,02
DMU3	1,00	0,91	0,38	163,78
DMU5	1,00	0,90	0,37	168,67
DMU2	1,00	1,00	0,26	288,20
DMU1	1,00	0,93	0,11	777,96
DMU6	1,00	0,89	0,06	1473,23

As it is presented at Table 2 and Table 3 above, results of analysis show that when just the exact data is taken into consideration, the best manufacturing sector is Manufacture of Motor, Vehicles, Trailers(DMU 10) whereas when ordinal data as well is taken into account, the best sector is Electric Machinery and Apparatus(DMU 9). This means that if the time spent for the least efficient DMU (DMU 6) can be

transferred to the best sector, there can be an increase in the overall export value of the country.

In addition to that, when the data sets are formed, it is also seen that total amount of time spent for the support of the export of the top 100 products are 63% of the whole time spent for the export related issues. If the remaining 37% can be used for the export support of the best sectors (DMU9 and DMU 10) it can also contribute to the overall export performance.

When the CCR-DEA model is applied to both of the data sets, it is seen that, the inefficient sectors obtained by the analysis of the exact data become efficient by the usage of the ordinal data. This indicates that even though some sectors could not reach the efficiency in terms of numerical performance, when the side effects and the strategic plans are taken into account, they can be considered as efficient.

6. Conclusion

This study describes the export performance evaluation process of Turkish Exports to Ireland for year 2008, using an imprecise DEA methodology, which gives us the opportunity to consider multiple criteria to assess the best sectors for the used exact and exact + ordinal real data sets.

As further studies, a similar analysis can be done at product level or for other markets.

References

- 1) Al-Khalifa, A. and Morgan, N.A. (1995), "Export performance measurement: a review and suggested directions", *Marketing Theory and Applications*, Vol. 6, American Marketing Association, Chicago, IL, pp. 313-18.
- 2) Aysan, F. A., Hacıhasanoglu, Y. S., "Investigation into the Determinants of Turkish Export-Boom in the 2000s", *Journal of International Trade and Diplomacy* 1 (2), Fall 2007, 159-200
- 3) Charnes, A., Cooper, W.W. and Rhodes, E., Measuring the efficiency of decision making units, *European Journal of Operational Research*, 1978, 2, 429-444.
- 4) Cook, W.D., Kress, M. and Seiford, L.M., "On the use of ordinal data in data envelopment analysis", *Journal of the Operational Research Society*, 1993, 44 (2), 133-140.
- 5) Cook, W.D., Kress, M. and Seiford, L.M., "Data envelopment analysis in the presence of both quantitative and qualitative factors", *Journal of the Operational Research Society*, 1996, 47(7), 945-953.
- 6) Farrell, M.J., "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society*, 1957, 120(3):253-290.
- 7) Karsak, E. E. and Ahiska, S.S., Practical Common Weight Multi-Criteria Decision Making Approach with an Improved Discriminating Power for Technology Selection, *International Journal of Production Research*, 2005, 43, 1537-1554.
- 8) Katsikeas, C.S., Leonidou, L.C. and Morgan, N.A., "Firm-Level Export Performance Assessment: Review, Evaluation, and Development", *Journal of the Academy of Marketing Science*, 2000, Vol.28 No.4 pages:493-511
- 9) Sousa, C. M. P., "Export Performance Measurement: An Evaluation of the Empirical Research in the Literature", *Academy of Marketing Science Review*, volume:2004, No:9, pages:1-22

- 10) Talluri, S., Data Envelopment Analysis: Models and Extensions, Decision Line, 2000 (May), 8-11.
- 11) Thirkell, P., C. and Dau, R., "Export performance: success determinants for New Zealand manufacturing exporters", European Journal of Marketing, Vol. 32 No. 9/10, 1998, pp. 813-829.
- 12) TURKSTAT, Foreign Trade, Export by Selected Fifty Countries Reports, 2008, http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=1930&tb_id=7
- 13) Zou, S. And Stan, S., "The determinants of export performance: a review of the empirical literature between 1987 and 1997", International Marketing Review 15(5):333-356.

Improving research students' performance by two contrasting methodologies: Theory of Constraints (TOC) and Appreciative Inquiry (AI)

Garoon Pongsart

3rd Year PhD student, Victoria Management School, Victoria University of Wellington

Supervised by A/P Dr. Victoria J. Mabin and Dr. Deborah Laurs

Garoon.Pongsart@vuw.ac.nz

Abstract

This research aims to compare how effective the Theory of Constraints (TOC), Appreciative Inquiry (AI), and a hybrid approach (a combination of TOC and AI) are in terms of understanding and improving Masters thesis students' performance. In addition, it will utilize exploratory research to find out the major issues of performance encountered by Masters thesis students at Victoria University of Wellington (VUW). The researcher will apply TOC to address and manage the root causes of problematic issues and will apply AI to exploit and enhance the root cause of success of these students. A hybrid approach, combining aspects of both methods, will be developed and applied separately. A mixed methodology will be employed. Firstly a web-based survey will be conducted to elicit student views on the thesis experience, and secondly a semi-structured individual interview (15 students) to explore in more detail and construct solutions using one of three approaches (TOC/AI/Hybrid) for each student. In the third stage of this research, the researcher will recruit 3 students, one from each individual interview group (TOC/AI/Hybrid) to take part in action research. The main purpose of conducting the action research is to understand the students' problems and success in greater depth and over time, further developing and applying the above-mentioned methods in order to help them improve/enhance their performance.

2. Introduction & Background of Study

The foundations of this research derive from two separate directions: firstly methodological and secondly a problematic situation. These two directions are strongly driven by the "highs" and "lows" of the researcher's several years working experiences in the business sector in Thailand and overseas combined with the researcher's passion for being part of an education reform in Thailand. This research aims to compare and contrast the two theories, Theory of Constraints (TOC) and Appreciative Inquiry (AI) by applying them to improve VUW Masters thesis students' performance.

Firstly the research aims to answer questions of a methodological nature. The Theory of Constraints (TOC) provides Thinking Processes (TP) and tools to find out, analyze, and manage the root cause of a problem, while Appreciative Inquiry (AI) offers positive questions together with its AI 4-D Cycle to seek and exploit the root cause of success. Despite approaching an issue from different ends, the two theories, TOC and AI, appear to have a common goal: striving for the best improvement. So there are also similarities within the two diametrically-opposed approaches. Can we utilize those similarities and differences? Can a hybrid approach be used as an alternative? And how effective are these approaches in dealing with a similar issue and context?

Many students fail to complete their theses on time. Of all New Zealand domestic students starting an Honours/Masters qualification at public providers in 1998, by the end of 2002 (5 years later), only 59% had completed their degrees successfully, 2% were still studying towards completion and 39% had left without completing (Scott, 2004). To fail in thesis completion is a waste of time and resource on both the university's and student's part. The low completion rate will impact the university's ranking and earnings. Low ranking universities may not attract desired levels of students enrolment, which may jeopardize funding subsidized by the government. Additionally students who have not had a job and/or live on a student loan, are finding their debts increasing through extension fees and their own living expenses. Thesis students typically experience high and low points in their research, which provide the impetus for the study using the two methodologies.

3. Theory of Constraints (TOC)

Goldratt's Theory of Constraints (TOC) provides Thinking Processes (TPs) to address the root cause of a problem. In TOC's problem-solving process, problems are traced back to an unresolved conflict. The conflict can be within an organization, between two or more people, and/or a personal conflict within oneself. Goldratt (1990, 4) characterizes constraints as "anything that limits a system from achieving higher performance versus its goal", and in the majority of cases these are found to be policy constraints rather than physical constraints – the latter being generally easier to resolve. The TOC TPs have been developed by Goldratt and other TOC scholars since the late 1980s specifically to handle policy constraints (Kim et al, 2008).

The TOC TPs include four critical questions, thinking process steps, and tools. The original three critical questions introduced by Goldratt (1990) are: what to change? what to change to? and how to cause change? The latest TOC version developed by Dettmer (2007) proposes an initial critical question, why is change needed? before the original three questions. What Dettmer has added is in accordance with Goldratt's TOC goal-oriented philosophy. The TOC TP steps are stated as identify the system goal, identify the core problem and the linkages to the problem, frame the core problem, construct and test the solution, identify the obstacles and intermediate objectives, prepare buy-in, implementation plan and activity plan (Cox, Blackstone, and Schleier, 2003, and Dettmer, 2007). The TOC TP tools are: Intermediate Objective (IO) Map, Evaporating Cloud (EC), Current Reality Tree (CRT), Future Reality Tree (FRT), Prerequisite Tree (PRT), Transition Tree (TT) and Categories of Legitimate Reservation (Goldratt, 1990, Noreen et al., 1995, Scheinkopf, 1999, Cox et al., 2003, and Dettmer, 2007). The collaboration of these tools assists and enhances the TOC TPs to provide an answer to the four critical questions of the TOC.

4. Appreciative Inquiry (AI)

Within the same situation, how can we also make use of the contrasting high points in the thesis experience? Interestingly, Appreciative Inquiry (AI) is based on the simple assumption that every organization has something that works well, and those strengths can be the starting point for creating positive change (Cooperrider, Whitney, and Stavros, 2008, 3). In order to improve a system's performance, Cooperrider's Appreciative Inquiry (AI) provides a 4-D Cycle to address the root cause of success (Cooperrider & Whitney, 2005, 12) and avoid "problem solving", which is different from TOC. AI seeks to accentuate the positive rather than eliminate the negative (Hayes, 2007, 295). The AI 4-D Cycle comprises Discovery, Dream, Design, and Destiny. Before employing the AI 4-D Cycle, an affirmative topic choice based on past and/or current success of a system needs to be constructed. In the "Discovery" phase (Appreciating what gives life), the system's members are invited/challenged/required to

discover and value positive exceptions, successes, and more vital or alive moments (Cooperrider, Whitney, and Stavros, 2008, 6). By doing this, the members are collectively appreciating their system's achievements before embarking on the next step.

The second step of AI 4-D Cycle is a Dream (Envisioning what might be), where system's members and stakeholders collectively explore their hopes and dreams in order to envision possibilities that are big, bold, and beyond the boundaries of what has been in the past (Whitney & Trosten-Bloom, 2003, 8).

The last two phases of AI are Design and Destiny. The Design phase (third phase) is determining what will be. System members and stakeholders are encouraged to combine what they have appreciated in the first phase and envisioned in the second phase to construct a provocative proposition. Reed (2007, 33) explains that the provocative proposition is a statement about what the organization wants to achieve. Then, the last phase of AI 4-D Cycle is Destiny (Planning what will be). The activity plan to achieve the provocative proposition is introduced and implemented by system members and owners.

5. Research Questions

Main research question: How effective are TOC, AI and Hybrid in dealing with performance issues of Masters thesis students?

6. Research Strategy

The research strategy is a general plan of how the researcher intends to go about answering the research question(s) (Saunders et al., 2000, 92). In this research, the researcher has conducted a web-based survey, interviews and action research in order to answer the research question. The web-based survey provided preliminary data, on the issues encountered by Masters thesis students at VUW. Next, the researcher recruited interviewees from participants who took part in the web-based survey. The researcher is conducting semi-structured individual interviews with 15 students: 5 students for TOC interview, 5 students for AI interview, and 5 students for Hybrid interview. At the end of this research, action research is being employed. According to Cardno (2003, 1) the term "action research" creates the expectation that those involved will be researching a particular situation with the intention of taking action that will make a difference – that is, bring change or improvement. The aim of this research is to understand and improve Masters thesis students' performance. To understand the students' performance, after receiving preliminary data, problematic issues identified by the web-based respondents and the semi-structured interview with each student by using a range of TOC or AI or Hybrid set of questions, the researcher employs action research as the final stage of this research. The action research includes an individual interview with 3 selected students from each group, TOC/AI/Hybrid, who experienced the same issues as per the web-based survey results.

7. The latest research findings

7.1 Web-based survey

The top ten issues encountered by VUW Masters thesis students based on the 2009 web-based survey results comparing with 2004 survey conducted by the researcher. The students were asked to rate the degree of difficulty from the list of issues provided by the Association for Support of Graduate Students (ASGS). Then the researcher summarized the top ten issues based on the degree of difficulty rated by the research participants (Table 2).

7.2 TOC Interview results

The researcher employed the TOC eight questions provided by Cox, Blackstone, and Schleier (2003) to conduct the individual interviews and added Intermediate Objectives (IO) Map (Dettmer, 2007) on top of the eight questions. Then the answers from each interviewee were composed into a storyline. From the storyline, the researcher applied TOC tools step by step, IO Map, CRT, EC, NBR, and PRT to the storyline told by the interviewee. Due to the limitation of page numbers in this paper, the researcher can only present IO Map, partial CRT, EC, and PRT of Masters thesis student, Cindy (not the interviewee's real name) in Figure 1 – 4. The 3 major issues experienced by Cindy are “Keeping your deadlines/timelines”, “Not knowing how to get started”, and “Feeling your study valuable/worthwhile”.

Masters thesis students' performance issues	Ranking	
	2009	2004
Keeping the deadlines	1	1
Knowing when to stop reading the literature	2	6
Designing your study	3	4
Knowing how to get started	4	5
Organizing the literature found	5	10
Staying motivated	6	8
Feeling supported	7	7
Writing the proposal	8	
Gathering the information for the literature review	9	
Feeling your study is valuable/worthwhile	10	
Keeping healthy/fit		2
Finding time for your thesis		3
Meeting social demands		9

Table 2: The top ten issues comparison between VUW 2009 and 2004 survey (Pongsart, 2009 & 2005)

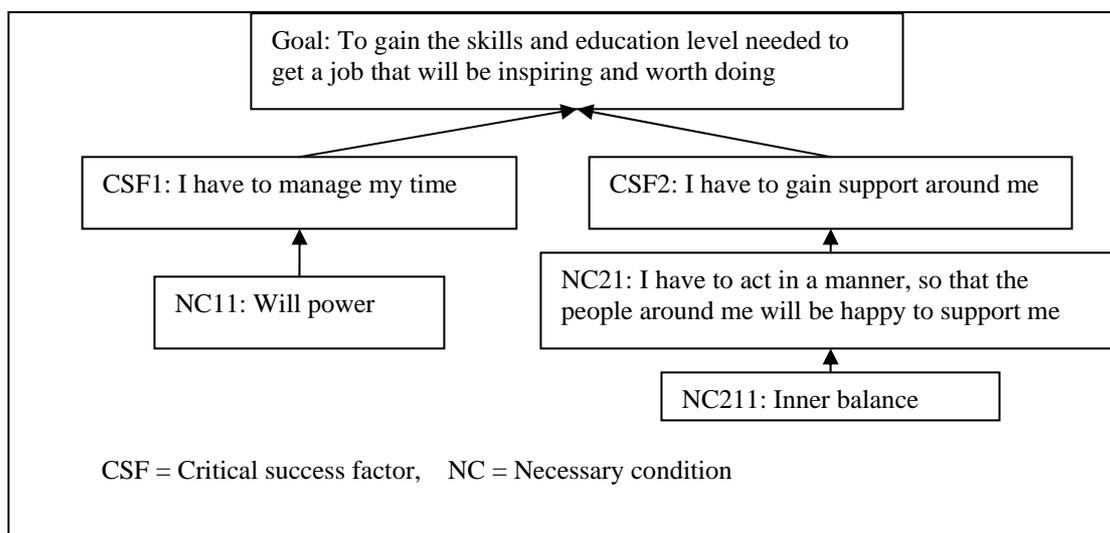


Figure 1: Intermediate Objectives (IO) Map of Cindy

According to the IO Map, Cindy identified her goal of pursuing a Masters Degree thesis including Critical Success Factors (CSF), and Necessary Conditions (NC) (Figure 1). The partial current situation of Cindy related to her major issues is demonstrated in the format of CRT (Figure 2). The CRT discloses the effect-cause-effect entities from 205 which constitutes an unacceptable deviation from expectations (Dettmer, 2007, 101) when compared to her IO Map (Figure 1).

7.2.1 What to change?

From CRT, the researcher selected entity 101 (I am very bad in making decisions) as a critical root cause of the problem to be addressed as it is within Cindy's span of control

and sphere of influence: areas where you have authority to control or influence the change (Dettmer, 2007, 70). Furthermore, as TOC usually frames a problem as a conflict. There is a conflict behind the entity 101. From one of Cindy’s answers, she said “*I felt that the longer I took to decide the less time I have to actually work on what I have decided*”. This is the existing conflict whenever Cindy has to make decisions. In making decisions, if she spends too much time on making decisions on any of her thesis’ issues she will not have much time left to work on the activities related to her decisions.

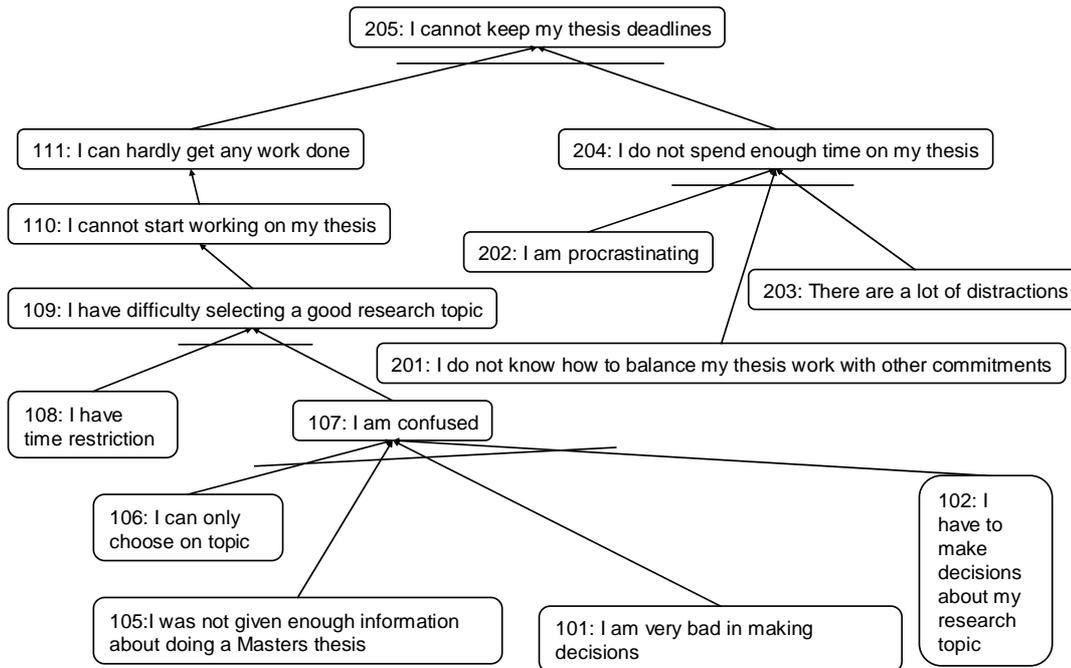


Figure 2: Partial Current Reality Tree (CRT) of Cindy

7.2.2 What to change to?

The next step is to employ Evaporating Cloud (EC) to demonstrate the existing conflict with underlying assumptions and find a solution to eliminate the conflict (Figure 3). The EC presents the two sides of Cindy’s conflict: D and D’ (Prerequisites). However, Cindy’s objective (A) in doing her Masters thesis is to be a successful student who can complete her thesis on time. In order to achieve A Cindy must complete B or C (Requirements). In order to achieve B or C Cindy must have done D or D’ respectively. Between each entity, there are underlying assumptions: AB, AC, BD, CD’, and DD’. In order to dissolve the conflict Cindy has to find an injection (solution) to make any of the existing assumption invalid. TOC suggests TOC practitioners to find a simple solution that can cause a huge impact on improvements (Goldratt, 1990). In this case, Cindy who is doing a Masters thesis should not spend time on her own in making decisions on her thesis issues. She should discuss and consult with her supervisor who usually plays an important role to support his/her students. This solution (to meet and discuss with her supervisor) can make the assumption BD invalid.

7.2.3 How to cause the change?

From the solution to the issues encountered by Cindy in sections 7.2.1 and 7.2.2, TOC also provides the steps and tool to verify and examine if any negative effects occurred by using the Negative Branch Reservation (NBR) method. However, due to the limitation of space in this paper, the researcher will not demonstrate the NBR process.

In this section, to answer one of the TOC critical questions, how to cause change?, TOC offers the Prerequisite Tree (PRT) to assist the solution's implementation (Figure 4). PRT requires an identification of a clear objective of all activities that we want to implement including Obstacles (O) and Intermediate Objectives (IO) to overcome those obstacles.

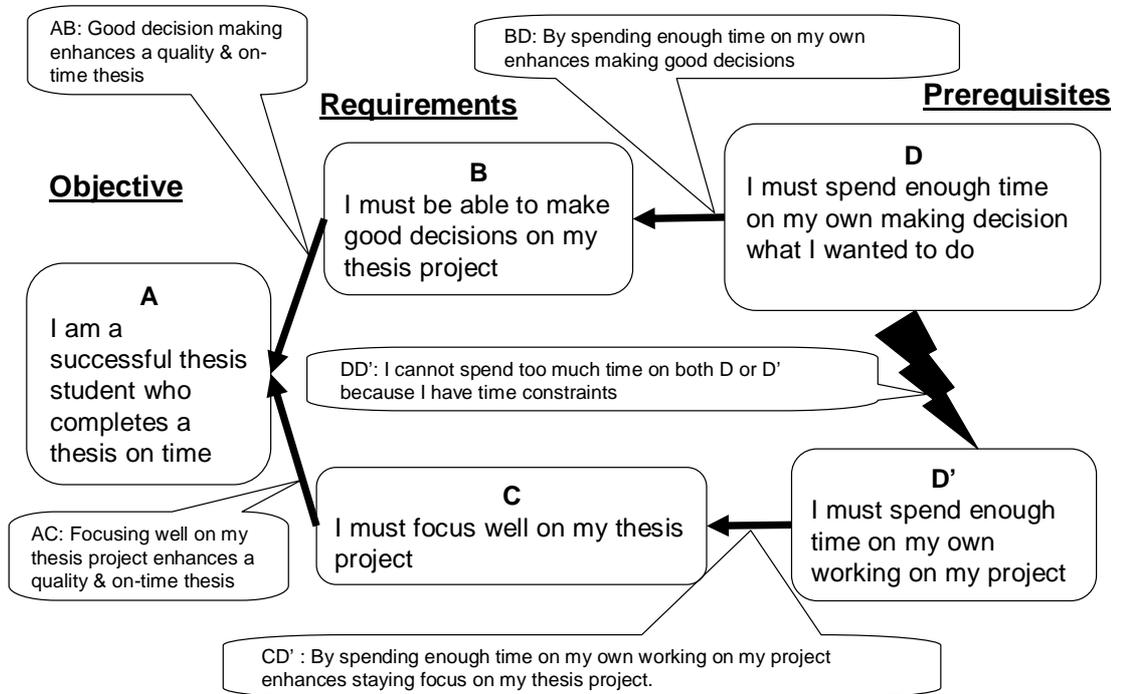


Figure 3: Evaporating Cloud (EC) of Cindy

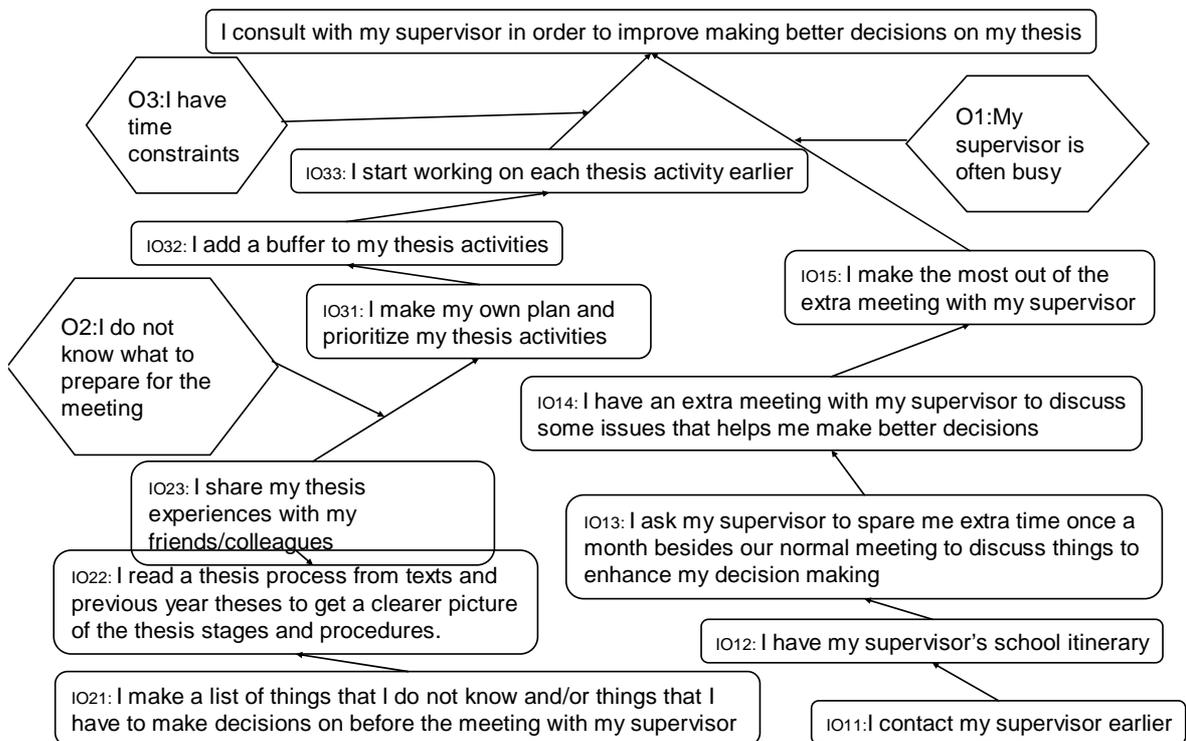


Figure 4: Prerequisite Tree (PRT) of Cindy.

7.3 AI Interview

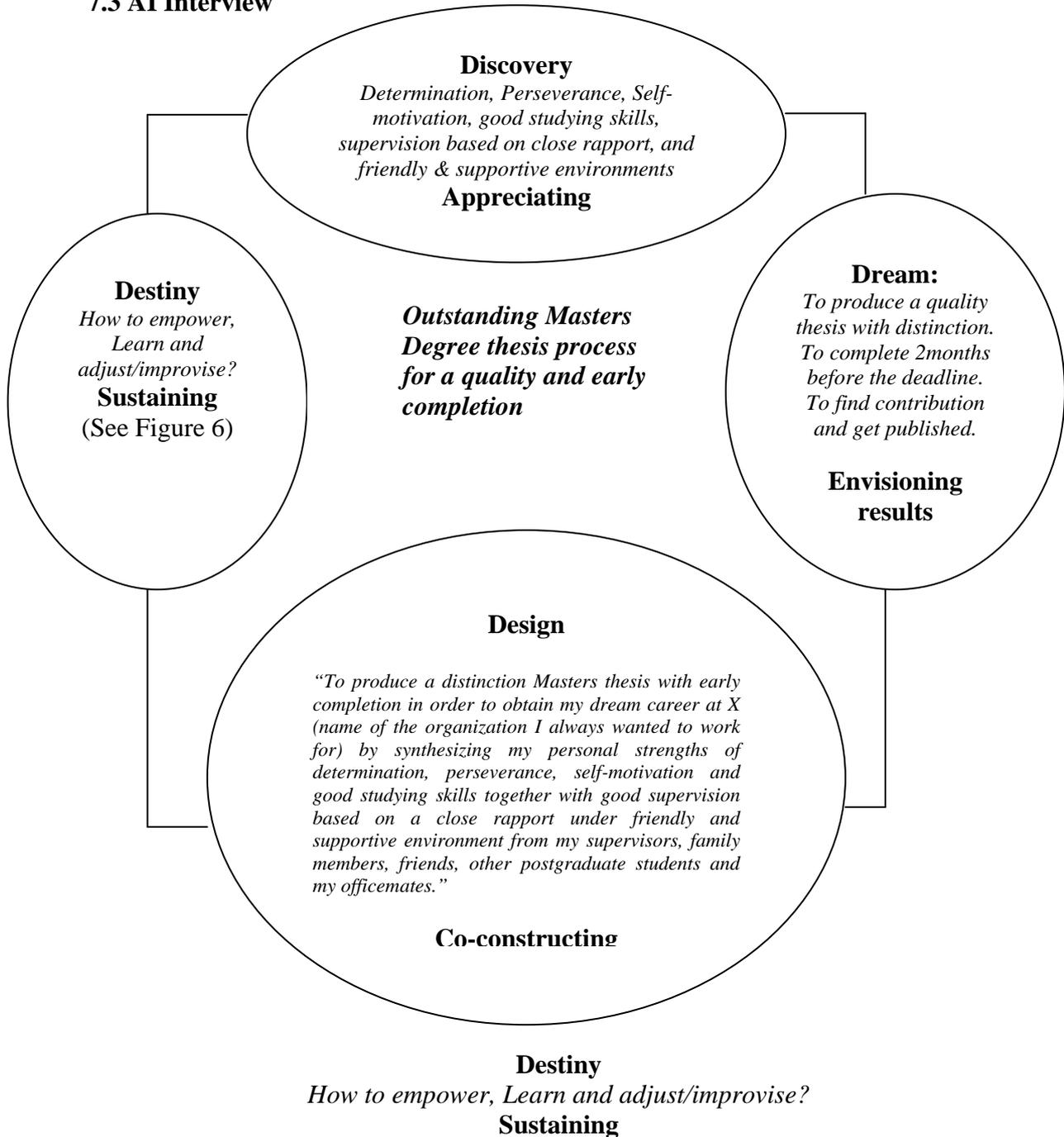


Figure 5: AI 4-D Cycle of Emma (Adapted from Cooperrider, Whitney & Stavros, 2008)

The Researcher recruited Emma (not her real name), another Masters thesis student who experienced the same 3 major issues (“Keeping your deadlines/timelines”, “Not knowing how to get started”, and “Feeling your study valuable/worthwhile”) as Cindy. By applying AI 4-D cycle to the issues found, the researcher employed AI positive questions to encourage Emma to talk about the strengths of her studying skills from her Honours year prior to pursuing a Masters thesis. The examples of those positive questions adapted from Cooperrider, Whitney and Stavros (2008) are:

- What would you describe as being a high-point experience in your university days when you were most alive and engaged? What happened? How was it? What are the key success factors that enabled you to the achievements?
- What the 3 wishes do you have to enhance the vitality of your Masters thesis?

From Emma's answers to AI positive questions, the researcher constructed a storyline and an AI 4-D Cycle (Figure 5). From Emma's story, her affirmative topic for pursuing a Masters thesis is "Outstanding Masters thesis process for a quality and early completion". The provocative proposition constructed from Emma's strengths combined with her dreams is shown in "Design" (Figure 5). Then, Figure 6 demonstrates a list of activities to be implemented in accordance with the provocative proposition in order to improve Emma's performance in the Destiny phase.

- *Outstanding Masters thesis: to discuss the criteria of a good thesis as well as the examiner's from my supervisors and related sources and to perform accordingly.*
- *Good studying skills: to use good reading techniques from my Honours to read articles or textbooks, good writing and analytical techniques, and to improve analytical skills by studying from the previous year outstanding theses and by submitting works to my supervisors and to learn from their constructive feedback.*
- *Good supervision based on close rapport: to maintain the same good supervision, to submit works to my supervisors regularly and get constructive feedback & criticism, to update work in progress to my supervisors, and to feel free to discuss any issues related to my thesis with my supervisors at anytime.*
- *Meeting the thesis deadlines: to motivate myself by thinking of the achievement with an aim to produce the outstanding thesis so that I can use this qualification to apply and get a job at the organization that I always want to work with in my home country, to work well under pressure, be prepare that I have had enough readings/get timely & constructive feedback from my supervisors, to stay focus on my thesis, and to set up a time table in order to submit my thesis chapters earlier.*
- *Friendly and supportive environment: to visit my office regularly and share research experience and/or discuss some issues with my friends, other postgraduate students, my officemates (to give and take willingly), to call home and get support from my family members.*

Figure 6 Destiny: A list of activities for Emma to be implemented in order to improve her performance

8. Summary

TOC provides Thinking Processes and tools step by step to address the root cause of a problem, starting from the IO Map, to find solutions to answer the TOC critical questions, what to change, to what to change, and how to cause change, in order to improve the situation and achieve the goal. AI provides positive questions and a 4-D Cycle to identify the root cause of success to achieve the best possible practices.

9. Reflection

To improve Masters thesis students' performance by using the two contrasting methodologies, TOC and AI, based on the interviews with the two students (Cindy and Emma) individually, the researcher is keen to reflect an improvement and change in students' performance as follows:

9.1 Goal versus the Affirmative Topic and Provocative Statement:

To pursue a quality and on-time or early completion Masters thesis, students must ask themselves why they want to do a thesis and set up a clear goal what they want to achieve. TOC, known as goal oriented theory, provides the IO Map not only to establish

a goal, but to identify the Critical Success Factors (CSF) and Necessary Conditions (NC) in order to achieve the set goal. The AI affirmative topic acts as a theme of the main focused activity for AI practitioners before embarking any activities. Furthermore, the first two “D”, Discovery and Dream, aims to lead those who are using AI to dream beyond boundaries after appreciating their own strengths. The Design phase produces a provocative proposition based on Discovery and Dream as a mission for AI users to move further or cause change. The accomplishment in this stage is that Masters thesis students have an explicit guideline from TOC to set up the IO Map, but need to spend enough time (for AI) to appreciate their own related strengths and dare to dream high in order to create motivating factors for them to pursue a quality and early or on-time completion thesis. Furthermore, AI’s success may depend highly on the right positive questions to the right interviewees.

9.2 The improvement process and tools:

While pursuing a Masters thesis, many students encounter several problematic issues. TOC provides Thinking Processes and tools to help address at the root cause of a problem in order to improve the problematic situations (see the examples in Figure 1-4). Masters thesis students who are having difficulties can see rich current pictures clearly from the CRT and can decide to address the core problem (rather than symptoms) which, once solved, will improve their performance significantly. EC, NBR and PRT are very helpful for the students to find the right solutions to the problem they are facing and to implement good actions to enhance their success.

AI does not pay much attention to problem solving; instead AI focuses on the root cause of success by providing positive questions with 4-D Cycle to cause positive changes and improvements or aiming for the best practices by using strengths based method. By employing AI to interview Masters thesis students, the interviewer may need to share positive thesis experience with the interviewee in order to facilitate the appreciating moment for the interviewee. Moreover, when asking positive questions for “Dream”, it may be helpful if the interviewee knows the objective of asking and answering the questions to enhance constructing a provocative proposition.

9.3 Effective Solutions (Figure 4 and 6):

The researcher applied TOC to Cindy’s problematic issues and addressed how to improve decision making from the CRT (Figure 2). In order to improve Cindy’s decision making while pursuing her Masters thesis (within a limited timeframe), it is apparent that Cindy must work closely with her supervisor and must have an extra meeting to improve her decision making. The PRT provides the list of activities for Cindy to prepare herself and make the most out of the meeting with her supervisor. By applying TOC, Cindy can see the effect-cause-effect diagrams that currently lead to the negative impact on her goal. This becomes one of the motivating factors for Cindy to continue working on overcoming the problematic issues in order to achieve her goal (IO map). The activities yielded from the PRT will help Cindy to eliminate the related problems.

On the contrary, when AI is applied to Emma’s issues (same issues as Cindy), the activities (Figure 6) that Emma needs to implement are based on the provocative proposition which was constructed from Emma’s strengths and dreams. AI has guided Emma to use her past achievements and strengths, determination, perseverance, self-motivation, good studying skills and close relationship with her supervisor, and supportive environments (friends, family members and supervisor) to embark on Emma’s Masters thesis. Furthermore, Emma will need to utilize her strengths and think of her high dreams to overcome any obstacles and to complete her thesis prior to the deadlines. The implementation plan (activity plan) yielded from AI does not aim to eliminate the problematic issues, but to achieve what Emma has been dreaming for.

10. Conclusion

The researcher has just completed interviewing 15 VUW Masters thesis students and applied 3 approaches, TOC, AI, and Hybrid, with 5 students for each approach. However, due to space limitations, the Hybrid findings and analysis are not included here. Analysis from the research findings is being completed group by group, with a preliminary analysis of the first group of interviewees already completed. The researcher is shortly to conduct action research with Cindy (TOC), Emma (AI) and one more student (Hybrid), as per the research design. In this way the researcher hopes to understand Masters thesis students' problems and success in greater depth and over time and to produce some tested guidelines to enhance Masters thesis students' performance and success as well as to contribute to the methodologies and research fields.

11. Acknowledgement

Thanks to VUW Post-Graduate Students' Association (PGSA), all VUW postgraduate coordinators, staffs and friends who helped recruit my research interviewees, and to my interviewees in particular. Thanks to VMS for giving an opportunity to pursue my dream. Special thanks to my supervisors who always support me and make my life a lot easier while pursuing a PhD at VUW, NZ which is far from my homeland.

12. References

- Cardno, Carol E.M. (2003): *Action Research: A Development approach*, Ellington: New Zealand Council for Educational Research.
- Cooperrider, D.L., Whitney, D., and Stavros, J.M. (2008): *Appreciative Inquiry Handbook: For Leaders of Change 2nd edition*, Brunswick: Crown Custom Publishing, Inc.
- Cooperrider, D. and Whitney, D. (2005): *Appreciative Inquiry: a Positive Revolution in Change*, San Francisco: Berrett-Koehler Publishers, Inc.
- Cox, J.F., Blackstone, Jr., J.H. and Schleier, Jr., J.G. (2003): *Managing Operations: A focus on excellence volume I*, Great Barrington: The North River Press Publishing Corporation.
- Dettmer, W.H. (2007): *The Logical Thinking Process: A Systems Approach to Complex Problem Solving*, Milwaukee: ASQ Quality Press.
- Goldratt, E.M. (1990): *What is the thing called Theory of Constraints and how should it be implemented?*, Croton-on-Hudson: The North River Press.
- Hayes, J. (2007): *The Theory and Practice of Change Management*, New York: Palgrave Macmillan.
- Kim, S., Mabin, V.J., Davies, J. (2008): The Theory of Constraints Thinking Processes: Retrospect and Prospect. *International Journal of Operations and Production Management* 28(2): 155-184.
- Noreen, E., Smith, D., and Mackey, T. (1995): *The Theory of Constraints and its implication for Management Accounting*, Great Barrington: The North River Press.
- Pongsart, G. (2005): *Postgraduate students' constraints in doing a Masters Degree thesis*, Wellington: Victoria University of Wellington
- Reed, J. (2007): *Appreciative Inquiry: Research for Change*, Thousand Oaks: Sage Publications, Inc.
- Saunders, M., Lewis, P. and Thornhill A. (2000): *Research Methods for Business Students*, Essex: Pearson Education limited.

- Scheinkopf, L.J. (1999): *Thinking for a Change: Putting the TOC Thinking Processes to Use*, Boca Raton: CRC Press LLC.
- Scott, D. (2004): *Retention, Completion and Progression in Tertiary Education 2003*, the Ministry of Education, [online]. Available: <http://www.minedu.govt.nz> [March 2004].
- Whitney, D. and Trosten-Bloom, A. (2003): *The Power of Appreciative Inquiry*, San Francisco: Berrett-Koehler Publishers, Inc.

Using phone logs to analyse call centre performance

John Paynter
Auckland Institute of Studies St Helens, Auckland
New Zealand
johnp@ais.ac.nz

Abstract

Call Centres are of increasing importance in a society expecting 24x7 services. Most utility providers run call centres to support in-bound calls. In addition, a large number of organisations support in-bound calls in response to advertising and similar campaigns or out-bound calls associated with market research and social marketing work. Despite improvements in telecommunications, and ICT in general, the cost of such services is increasing in terms of staffing costs. Thus, increasingly New Zealand organisations are outsourcing call centres to Asian countries, such as India and the Philippines.

It is necessary to run efficient operations to gain competitive work and to save jobs. A good metrics collection program is needed to understand call centre dynamics and to competitively price contracts and to run the different operations. In this paper we analysis a call centre's phone logs to understand the various factors that influence the efficiency of handling out-bound calls. The variables examined include the type of campaign, the call length, the response rate, the call outcome, individual differences between operators and learning effects during the course of the campaign.

This information can be used to price the calls per completion and to fine-tune the call centre operation during individual campaigns.

A Multi-plan Method for Radiotherapy Treatment Design via Finite Representation of the Non-dominated Set of Multi-objective Linear Programmes

Matthias Ehrgott and Lizhen Shao
Department of Engineering Science
The University of Auckland
New Zealand
m.ehrgott@auckland.ac.nz

Abstract

The choice of a plan for radiotherapy treatment for an individual cancer patient requires the careful trade-off between the goals of delivering a sufficiently high radiation dose to the tumour and avoiding irradiation of critical organs and normal tissue. This problem can be formulated as a multi-objective linear programme (MOLP). In this talk we present a method to compute a finite set of non-dominated points that can be proven to uniformly cover the complete non-dominated set of an MOLP (a finite representation). This method generalises and improves upon two existing methods from the literature. We apply this method to the radiotherapy treatment planning problem, showing some results for clinical cases. We illustrate how the method can be used to support clinician's decision making when selecting a treatment plan. The treatment planner only needs to specify a threshold for recognising two treatment plans as different and is able to interactively navigate through the representative set without the trial-and-error process often used in practice today.

Key words: Radiation therapy, multi-objective optimisation, linear programming, finite representation.

1 Radiation Therapy for Cancer

Apart from surgery and chemotherapy, radiation therapy is a major treatment mode for cancer. Ionising radiation damages the deoxyribonucleic acid (DNA) of cells. Although this affects both healthy and cancerous cells, non-cancerous cells are able to reproduce even with slightly damaged DNA, whereas even small amounts of DNA damage renders cancerous cells incapable of reproducing. Radiation therapy exploits this therapeutic advantage to focus radiation so that enough dose is delivered to the targeted region to damage the cancerous cells while sparing surrounding anatomical structures. In intensity modulated radiation therapy (IMRT) based on photon or electron beams, which we are concerned with in this paper, several beams are

focused on the tumour from usually between three and nine directions. The cumulative effect of intersecting beams enables a high dose delivered to the tumour while spreading the dose to healthy organs out to keep it low. The intensity of radiation (the irradiation time) can be modulated across each beam using a mechanical device called multi-leaf collimator (MLC) that blocks out areas of the beam by moving metal leaves into the beam, essentially decomposing the beam into a large number of sub-beams (also called beamlets or bixels). This technique allows conformation of the beam to the shape of the tumour and further reduction of dose to healthy tissues. An IMRT treatment plan needs to specify the beam directions, the intensity for each bixel of each beam (called intensity patterns or fluence maps), and a schedule for the movements of the collimator leaves to deliver the optimised fluence maps. While IMRT allows much more precise and higher quality treatments than conventional open field or conformal radiotherapy it also increases the complexity of the planning process due to the very large number of parameters that need to be specified. This gives rise to the optimisation problems of choosing optimal beam directions, intensities, and delivery schedules. Because of this, operations research methods have increasingly been applied to IMRT in the last decade, as the survey by Ehrgott *et al.* (2008) shows. In this paper we deal with the determination of optimal intensities.

Knowing the intensities of all bixels of all beams, the radiation dose (measured in gray, Gy, 1 Gy = 1 J/kg) delivered to a point in the patient body can be calculated. To that end the patient body is discretised into 3D volume elements (voxels) and dose is calculated at one point per voxel. We write the intensities as a vector of variables $x = (x_j)_{j=1,\dots,n}$ and let a_{ij} denote the dose delivered to voxel i with unit intensity applied at bixel j . The values a_{ij} can be computed using models of the physical behaviour of radiation as it interacts with matter. They define a dose deposition matrix $A = (a_{ij})_{i=1,\dots,m;j=1,\dots,n}$ and the dose distribution vector $d = (d_i)_{i=1,\dots,m}$ can be calculated as $d = Ax$ (see e.g. Ehrgott *et al.* (2008) and references therein).

At the beginning of the radiotherapy treatment process an oncologist will prescribe a dose to be delivered to the tumour. To maximise tumour control probability, this dose is to be uniformly delivered to all tumour voxels. On the other hand, the oncologist will also prescribe tolerable dose levels for critical organs close to the tumour and other normal tissue. To minimise the probability of complications from radiotherapy treatment, these dose levels ought not to be exceeded. However, radiation has to travel through normal tissue to reach the tumour site and it is usually not possible to exactly achieve the prescribed dose levels. Hence the determination of intensities pursues contradictory goals. Below, we formulate the beam intensity optimisation problem as a multi-objective linear programme (MOLP):

$$\begin{array}{ll}
\min & (\alpha, \beta, \gamma) \\
\text{s.t.} & TLB - \alpha e \leq A_T x \leq TUB \\
& A_C x \leq CUB + \beta e \\
& A_N x \leq NUB + \gamma e \\
& 0 \leq \alpha \leq \alpha_u \\
& -\min CUB \leq \beta \leq \beta_u \\
& 0 \leq \gamma \leq \gamma_u \\
& 0 \leq x,
\end{array} \tag{1}$$

where e is the vector in which each entry is 1, A_T, A_C, A_N are sub-matrices of A consisting of the rows of A pertaining to voxels in the tumour, critical organs, and

normal tissue, respectively. TLB, TUB, CUB, NUB are vectors of lower and upper bounds on the dose delivered to tumour, critical organ, and normal tissue voxels, respectively. These are derived from the oncologists prescription doses.

Formulation (1) is a modification of the elastic constraint linear programme of Holder (2003) where we omit weighting factors for the objectives but include upper bounds on the objective function values. The objective functions are to minimise the maximum deviations α, β, γ from tumour lower bounds, critical organ upper bounds, and normal tissue upper bounds. The constraints ensure that α, β, γ are defined properly, that they do not exceed clinically relevant values α_u, β_u , and γ_u and that the physical constraint of non-negative intensity is met. Notice that negative values of β are possible and encouraged as they mean that the dose delivered to critical structures is below the tolerable limit. The size of problem (1) in a clinical case can be very large. Modern MLCs allow 1,600 bixels per beam. With nine beams this means about 15,000 variables and possibly in the order of 100,000 constraints, depending on the voxel resolution. Solving (1) is therefore a challenge.

In Section 2 we summarise existing methods to solve multi-objective linear programmes such as (1) and the ways treatment plans are computed in clinical practice today. We argue that the algorithms may not be suitable for clinical practice, while clinical treatment planning methods do not appropriately account for the mathematical nature of the optimisation model (1). Hence, in Section 3 we present a new approach which addresses this mismatch between theory and practice. Our method is based on computing a finite set of solutions of an MOLP whose objective function vectors “represent” the infinite set of non-dominated points in the objective space of an MOLP in the sense that they are uniformly distributed over the whole non-dominated set. In the radiotherapy treatment planning problem, each of these solutions represents a possible treatment plan that can be presented to a treatment planner. Hence, the treatment planner can easily navigate among the finite set of treatment plans to identify the most suitable one for the individual patient. Nevertheless, there is a (mathematical) assurance that the range of possible trade-offs between the goals of tumour control and healthy tissue protection is represented in the computed treatment plans. In Section 4 we apply our method to a radiotherapy treatment example and illustrate the process of selecting a plan. Finally, in Section 5 we draw some conclusions and point out directions of future research.

2 Solving Multi-objective Linear Programmes and Finding Treatment Plans

A multi-objective linear programme can be written as the following optimisation problem:

$$\min\{Cx : Ax = b, x \geq 0\}, \quad (2)$$

where C is a $p \times n$ matrix of objective function coefficients, x is a vector of decision variables of length n , A is an $m \times n$ matrix of constraint coefficients, and b is a vector of right hand side values of length m . The aim of (2) is to simultaneously minimise $p \geq 2$ objective functions. Thus, because vectors in \mathbb{R}^p are not always comparable, the following definition of minimisation is used. A feasible solution \hat{x} of (2) is called efficient if there is no other feasible solution x such that $Cx \leq C\hat{x}$ and $Cx \neq C\hat{x}$. If \hat{x} is efficient then $\hat{y} = C\hat{x}$ is called non-dominated. Solving (2) therefore means finding the set of efficient solutions X_E or the set of non-dominated points Y_N . Since $X = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ is a polyhedron, so is $Y = \{Cx : x \in X\}$. Hence,

unless in trivial cases, X_E and Y_N are connected (but non-convex) subsets of the surface of polyhedra X and Y , respectively.

MOLPs have been studied since the 1960s and a variety of solution approaches have been developed. We shall briefly review the main approaches here, but refer to Ehrgott and Wiecek (2005) and references therein for more details. Most attention has been given to algorithms based on extensions of the simplex method of linear programming. These rely on the fact that basic feasible solutions of (the constraints of) (2) correspond to extreme points of X . Basic feasible solutions corresponding to efficient extreme points of X are hence called efficient. It can be shown that efficient basic feasible solutions can be identified using some generalised reduced cost criterion, and that efficient basic feasible solutions are connected via so-called efficient pivots. Hence, the single objective simplex algorithm can be extended to pivot among efficient basic feasible solutions and thereby identify X_E . The major drawback of the simplex approach is the possibly very large number of efficient extreme points. It is easy to construct examples, where this number increases exponentially in both the number of objectives p and the number of variables n .

The advent of interior point methods for linear programming also stimulated some interest in interior point methods for MOLP. There is, however, a fundamental problem with the applicability of the interior point paradigm. Interior point methods converge to a single solution. To date, no algorithm has been proposed that can identify the whole set X_E – it seems that at best a face of X_E can be identified (Abhyankar *et al.*, 1990).

The fact that the number of efficient extreme points can be very large together with the observations that p is in general much smaller than n , and that often many efficient solutions map to the same non-dominated point has given rise to another approach, which attempts to solve (2) in objective space. Such objective space methods focus on finding Y_N . For any element of Y_N a corresponding solution $x \in X_E$ can be calculated via solving a linear programme, if necessary. Skipping the historical development, we just mention Benson’s algorithm (Benson, 1998). This algorithm first defines a polyhedron S containing Y then proceeds to compute supporting hyperplanes to Y , adding them to the description of S , until Y_N is known. Recent advances in duality theory for MOLP (Heyde and Löhne, 2008) have also made it possible to develop a dual version of Benson’s algorithm (Ehrgott *et al.*, 2007).

Despite the advantage of objective space methods over simplex based algorithms, the application to the very large problems arising in radiation therapy ((1) may have hundreds of thousands of constraints and thousands of variables), they may still result in unacceptable computation times. Shao and Ehrgott (2008a) and Shao and Ehrgott (2008b) have previously developed approximation algorithms based on Benson’s algorithm and its dual variant. These allow to compute Y_N to some specified accuracy, which leads to considerable savings in computation time, see also Shao and Ehrgott (2008c).

We can conclude here that algorithms to solve (2) are designed to compute X_E or Y_N (or to approximate these sets). But is that what is needed in radiotherapy treatment planning in practice? After all, the radiotherapist needs a treatment plan (defined by a feasible solution x of (1)). Naturally, this plan should be an efficient solution. Realising that simplex methods are not applicable to very large scale problems and that interior point methods appear to be not useful at all one would need to resort to (exact or approximate) objective space method. For the treatment planner this entails selecting the most preferable treatment plan by selecting y from Y_N , i.e.

based on the outcome (dose distribution) of a treatment and then implementing the corresponding plan x with $Cx = y$. This is quite reasonable, as planners are used to judging treatments by the resulting dose distribution (i.e. $d = Ax$) rather than its intensity x . However, Y_N is an infinite continuous set (a subset of the faces of a polyhedron). So the issue of how to choose y from Y_N needs to be addressed.

Let us now consider the practice of radiotherapy planning. In “the old days” a forward approach was used. I.e. radiation intensities were selected and the resulting dose distribution calculated. This process was repeated until a satisfactory treatment was found. Improvements in technology made this approach obsolete and computerised treatment planning systems became the norm. These systems are based on a forward approach, i.e. a desired dose distribution is specified (e.g. via the lower and upper bounds incorporated in (1)) and a matching intensity x is calculated. This is usually done by formulating and solving an optimisation problem with the objective of minimising a weighted deviation between the calculated and desired dose distribution, in other words, in the LP setting we use in this paper, minimising $w_1\alpha + w_2\beta + w_3\gamma$ with the constraints of (1), see Holder (2003). Thus, “importance weights” w_1, w_2, w_3 are selected, the optimal x is calculated, and if necessary the process is repeated with different weights until a satisfactory treatment plan is found. There is some justification for this approach via Isermann’s theorem (Isermann, 1974) which states that a feasible solution \hat{x} of (2) is efficient if and only if there is a vector $w \in \mathbb{R}^p, w > 0$ such that \hat{x} is an optimal solution of the single objective LP

$$\min\{w^T Cx : Ax = b, x \geq 0\}.$$

So the current clinical approach guarantees that an efficient solution is chosen. However, there is a problem with the trial-and-error method of selecting w . It is well known mathematically that widely different weights w may result in one and the same efficient solution; it is equally well known that very similar weights w may result in drastically different efficient solutions. In short, too little information is available about the relationship between positive weights w and efficient solution x to be used in a trial-and-error process in clinical practice.

To conclude, we observe a mismatch between the mathematical algorithms to solve MOLPs, which are not necessarily useful for practical application in a clinical context, and current clinical practice which lacks mathematical justification. In the next section we present an approach to address this mismatch.

3 Finite Representation of Non-dominated Sets

In this section we present a different approach to solve MOLPs. Rather than attempting to compute X_E or Y_N we attempt to find a finite subset R of Y_N (and the associated efficient solutions) that “represents” Y_N , i.e. we require the cardinality of R to be “reasonable”, that there is no large area of Y_N which does not contain an element of R and that the points of R are not too close together. We formalise this next.

Definition 1 *A finite representation of Y_N is a subset R of Y_N such that $|R| < \infty$.*

To formalise the above mentioned quality criteria we follow Sayin (2000) who first introduced them. Let d be a metric on \mathbb{R}^n . The coverage error ε of the representation R of the non-dominated set Y_N is defined as

$$\varepsilon := \max_{y \in Y_N} \min_{r \in R} d(y, r).$$

The uniformity level δ of the representation R of the non-dominated set Y_N is defined as

$$\delta := \min_{r^1, r^2 \in R} d(r^1, r^2).$$

A good representation R of Y_N has small cardinality $|R|$, small coverage error ε and a high uniformity level δ .

Definition 2 Let R be a representation of Y_N , d a metric and $\varepsilon > 0$ and $\delta > 0$ be real numbers.

- R is called a d_ε -representation of Y_N if for any $y \in Y_N$, there exists $r \in R$ such that $d(y, r) \leq \varepsilon$.
- R is called a δ -uniform d_ε -representation if $\min_{r^1, r^2 \in R, r^1 \neq r^2} \{d(r^1, r^2)\} \geq \delta$.

There are several methods in the literature that describe algorithms to compute finite representations. The global shooting method (Benson and Sayin, 1997) is guaranteed to cover the whole non-dominated set, but it is possible to construct examples for which the uniformity level is arbitrarily bad. The normal boundary intersection method of (Das and Dennis, 1998) on the other hand produces representations with good uniformity, but it can be shown that the majority of Y_N may not be represented in problems with $p \geq 3$ objectives. The normal constraint method (Messac *et al.*, 2003) has been reported to result in good coverage and good uniformity, but there are no theoretical results ensuring this.

Our *revised boundary intersection method* – illustrated for an MOLP with two objectives in Figure 1 – combines features of both global shooting and normal boundary intersection and avoids their disadvantages. We summarise the description given in Shao and Ehrgott (2007) in Algorithm 1. For simplicity of exposition we assume that Y_N is bounded.

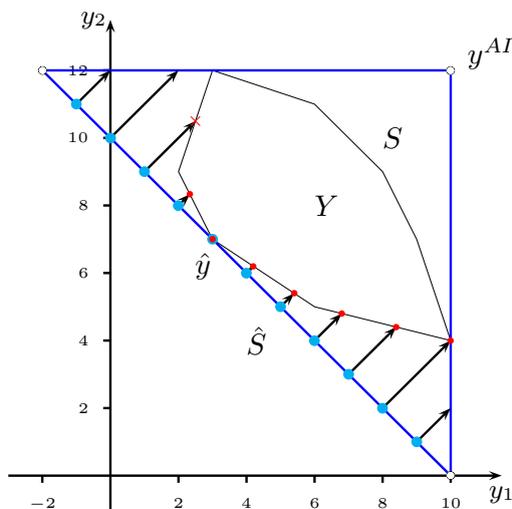


Figure 1: The simplex S (triangle), defined by y^{AI} and the non-dominated point \hat{y} contains the feasible set Y . Equidistantly spaced reference points are located on the reference plain \hat{S} supporting Y in \hat{y} (diagonal line). The reference points are projected to Y (indicated by arrows). Some projections do not yield points in Y (top left) and some may be dominated (indicated by an x). The non-dominated points (dots on the boundary of Y) are uniformly distributed and cover all of Y_N .

Note here that the only input parameter (apart from the MOLP data) required for the procedure is the distance of the reference points ds . This determines the number of reference points, which is an upper bound on $|R|$. The following theorem asserts the uniformity level of R . We use the Euclidean distance to measure distances.

Theorem 1 Let R be the representation of Y_N obtained Algorithm 1 and let q^1, q^2 be two reference points with $d(q^1, q^2) = ds$ that yield non-dominated representative points r^1, r^2 . Then $ds \leq d(r^1, r^2) \leq \sqrt{p}ds$. Hence R is a ds -uniform representation of Y_N .

Algorithm 1 Revised Normal Boundary Intersection Method

- 1: **Input:** MOLP data A, b, C and $ds > 0$.
- 2: Find y^{AI} defined by $y_k^{AI} = \max\{y_k : y \in Y\}, k = 1, \dots, p$.
- 3: Find a non-dominated point \hat{y} by solving the LP $\phi := \min\{e^T y : y \in Y\}$.
- 4: Compute $p + 1$ points $v^k, k = 0, \dots, p$ in \mathbb{R}^p as follows:

$$v_l^k = \begin{cases} y_l^{AI}, & \text{if } l \neq k, \\ \phi + \hat{y}_k - e^T v^0 & \text{if } l = k, \end{cases} \quad \text{for } l = 1, 2, \dots, p.$$

- The convex hull S of $\{v^0, \dots, v^p\}$ is a simplex containing Y . The convex hull \hat{S} of $\{v^1, \dots, v^p\}$ is a hyperplane with normal e supporting Y in \hat{y} .
- 5: Compute equally spaced reference points q^i with distance ds on \hat{S} .
 - 6: For each reference point q solve the LP $\min\{t : q + te \in Y, t \geq 0\}$. If this LP is infeasible, the ray $q + te$ does not intersect Y , otherwise for the optimal value \hat{t} , $q = \hat{t}y \in Y$. Note that the LP cannot be unbounded.
 - 7: For each weakly non-dominated point \hat{y} found in the previous step solve the LP $\min\{e^T y : y \in Y, y \leq \hat{y}\}$. It holds that \hat{y} is non-dominated if and only if it is an optimal solution of this LP.
 - 8: **Output:** Representation R consisting of the non-dominated points confirmed in Step 7.
-

To assure the coverage we note that Y_N is the union of maximal nondominated faces Y^j of Y . Let S^j be the projection of Y^j on \hat{S} in direction e , i.e. the projection of Y_N on \hat{S} is $Y_N^p = \cup_j S^j$. Let the width $w(S)$ of a subset S of \hat{S} be the smallest distance between any two parallel supporting hyperplanes of S .

Theorem 2 *Let R be the representation of Y_N obtained from Algorithm 1 and assume that $w(S^j) \geq ds$ for each subset S^j of the projection of Y_N on \hat{S} . Then R is a ds -uniform $d_{\sqrt{p}ds}$ -representation of Y_N .*

4 Application to a Radiotherapy Treatment Case

We illustrate our method using a (simplified) clinical case, namely an acoustic neuroma, see Figure 2. The representative set is depicted in Figure 3.



Figure 2: The acoustic neuroma is a brain tumour. On the simplified CT image it is shown in light gray and borders the brain stem in the centre. The problem (based on 3 CT images and 3mm voxel size) has 78 tumour voxels, 472 critical organ voxels, 6778 normal tissue voxels and 597 bixels. The lower and upper bounds for the tumour dose are 57.58 and 61.14, respectively. The upper bound for the dose to critical organs is 50 Gy for the brain stem and 5 Gy for the eyes. The upper bound for normal tissue is set to zero. The resulting MOLP (1) has dimension $m = 7.410, n = 600, p = 3$.

At the beginning of a planning session, one of the non-dominated points is chosen. We choose a point that is centrally located in R . This point is shown as \odot in Figure 3 and has objective function values $\alpha = 3.88, \beta = 2.37, \gamma = 36.35$. This

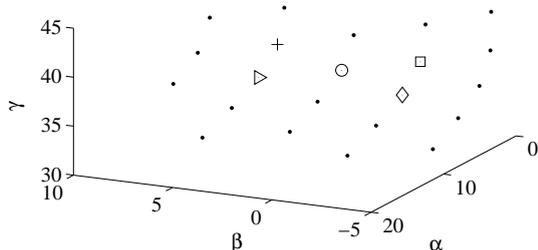


Figure 3: The distance of the reference points is $ds = 3.47$, equivalent to 153 reference points. 22 nondominated points are found in 140 seconds of computation time. The points marked by different symbols are referred to in the simulated planning session.

means that there is a tumour voxel that receives 3.88 Gy less than the lower bound of 57.58, a critical organ voxel that receives 2.37 Gy more than its desired upper bound and a normal tissue voxel that receives 26.53 Gy. Figure 4 illustrates the dose distribution and dose volume histograms (DVHs) for this plan. The dose distribution shows isodose curves overlaid on contours of tumour and organs. The dose volume histogram shows the dose (as a fraction of prescribed tumour dose) by percentage of organ volume, i.e. a point (a, b) on one of the curves means that $a\%$ of the volume receive a dose of $b\%$ of the prescribed tumour dose or more.

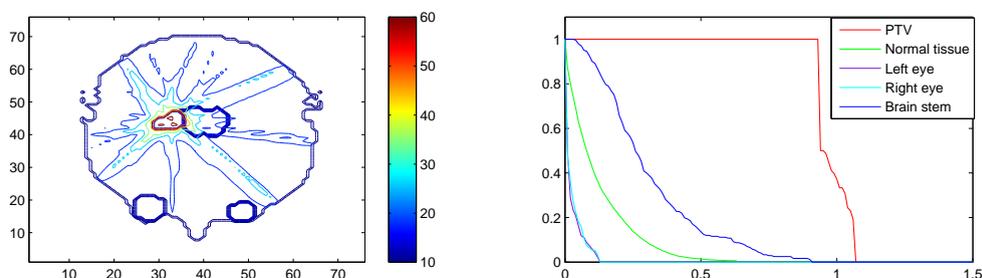


Figure 4: Plan 1, objective values (3.882, 2.366, 36.354) depicted by \odot in Figure 3.

Considering an improvement of the objective value for the tumour, which is achievable with tolerable deteriorating to the objective values for the critical organs and normal tissue, the planner may decide on the plan which is depicted by $+$ in Figure 3 with the objectives (2.663, 6.048, 37.585). Dose distribution and DVHs are shown Figure 5. Alternatively, the planner may consider a plan with better objective value for the critical organs and the normal tissue, such as the plan depicted by \diamond in Figure 3 with objectives (5.231, -1.185, 35.253). Its dose distribution and DVHs are shown in Figure 6.

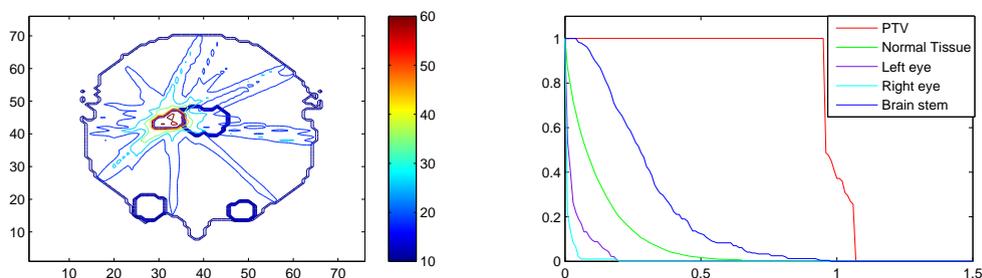


Figure 5: Plan 2, objective values (2.663, 6.048, 37.585) depicted by $+$ in Figure 3.

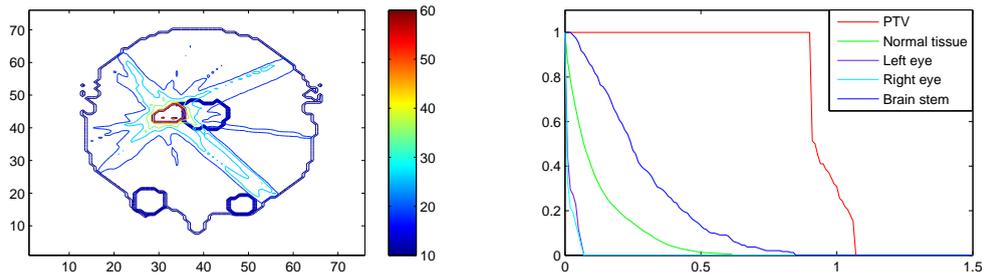


Figure 6: Plan 3, objective values (5.231, -1.185, 35.253) depicted by \diamond in Figure 3.

Additionally, if the planner wants a better objective value for the tumour by deteriorating the objective value for the normal tissue, he may consider the plan depicted by \square in Figure 3 with objective values (2.770, -1.196, 37.693). Dose distribution and DVHs are shown in Figure 7. If he wants a better objective value for the normal tissue, he may also consider the plan depicted by \triangleright in Figure 3 with the objectives (5.148, 6.083, 35.170). Its dose distribution and DVHs are shown in Figure 6.

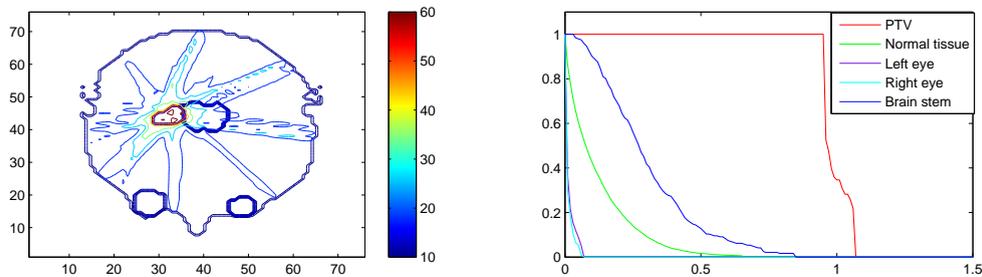


Figure 7: Plan 4, objective values (2.770, -1.196, 37.693) depicted by \square in Figure 3.

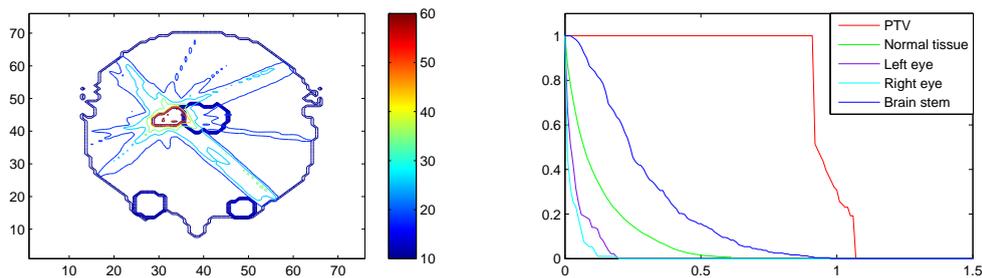


Figure 8: Plan 5, objective values (5.148, 6.083, 35.170) depicted by \triangleright in Figure 3.

In this simulated treatment session, a treatment planner's workflow does not change significantly from current practice. The optimisation problem is set up and plans are calculated by the treatment planning system. The planner visualises and assesses the suitability of treatment plans exactly as in current practice. However, there are significant differences in the support for the workflow:

- Computation and selection are de-coupled because multiple plans are calculated in parallel, rather than sequentially. Hence the trial-and-error process of selecting weights and re-optimising is eliminated and replaced with an on-line plan selection tool, reducing planning time.

- There is a guarantee that the whole nondominated set is covered, hence a better chance to find the best plan for the patient using a systematic exploration of efficient treatment plans.
- The price for this is the specification of the parameter δ , which can be interpreted as a measure of difference between dose distributions (and has unit Gy). This threshold is easily understood and specified by planners.

5 Conclusion

In this paper we have presented a new approach for computing a finite set of points representing the non-dominated set of an MOLP in objective space. We have applied this method to the problem of finding a suitable treatment plan for radiotherapy of cancer and demonstrated that this approach allows a multi-plan approach to the treatment planning problem that is appropriate to the multi-objective nature of the problem and close to treatment planners current practice, while also mathematically justified.

Further challenges include the solution of large scale problems that occur in clinical cases and the numerical issues arising from that. In practice, the intensity problem is closely related to the problem of finding optimal beam directions and delivery schedules on the MLCs. Integrating the beam intensity problem with either of those would be an important step towards an ultimate goal of solving the radiotherapy treatment problem as an integrated optimisation problem including beam selection, intensity optimisation, and delivery planning.

References

- Abhyankar, S., Morin, T., and Trafalis, T. (1990). Efficient faces of polytopes: Interior point algorithms, parametrization of algebraic varieties, and multiple objective optimization. *Contemporary Mathematics*, **114**, 319–341.
- 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**, 1–24.
- Benson, H. P. and Sayin, S. (1997). Towards finding global representations of the efficient set in multiple objective mathematical programming. *Naval Research Logistics*, **44**, 47–67.
- Das, I. and Dennis, J. E. (1998). Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, **8**(3), 631–657.
- Ehrgott, M. and Wiecek, M. (2005). Multiobjective programming. In J. Figueira, S. Greco, and M. Ehrgott, editors, *Multicriteria Decision Analysis: State of the Art Surveys*, volume 78 of *International Series in Operations Research & Management Science*, chapter 17, pages 667–722. Springer New York.
- Ehrgott, M., Löhne, A., and Shao, L. (2007). A dual variant of benson’s outer approximation algorithm. School of Engineering Report 654, Department of Engineering Science The University of Auckland.
- Ehrgott, M., Güler, C., Hamacher, H. W., and Shao, L. (2008). Mathematical optimization in intensity modulated radiation therapy. *4OR*, **6**(3), 199–262.

- Heyde, F. and Löhne, A. (2008). Geometric duality in multi-objective linear programming. *SIAM Journal on Optimization*, **19**(2), 836–845.
- Holder, A. (2003). Designing radiotherapy plans with elastic constraints and interior point methods. *Health Care Management Science*, **6**, 5–16.
- Isermann, H. (1974). Proper efficiency and the linear vector maximum problem. **22**, 189–191.
- Messac, A., Ismail-Yahaya, A., and Mattson, C. A. (2003). The normalized constraint method for generating the pareto frontier. *Structural Multidisciplinary Optimization*, **25**, 86–98.
- Sayin, S. (2000). Measuring the quality of discrete representations of efficient sets in multiple objective mathematical programming. *Mathematical Programming*, **87**, 543–560.
- Shao, L. and Ehrgott, M. (2007). Finding representative nondominated points in multi-objective linear programming. In P. Bonissone, editor, *IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, Honolulu 1-5 April 2007, Proceedings*, pages 245–252. IEEE.
- Shao, L. and Ehrgott, M. (2008a). Approximately solving multiobjective linear programmes in objective space and an application in radiotherapy treatment planning. *Mathematical Methods in Operations Research*, **68**(2), 257–276.
- Shao, L. and Ehrgott, M. (2008b). Approximating the nondominated set of an MOLP by approximately solving its dual problem. *Mathematical Methods in Operations Research*, **68**, 469–492.
- Shao, L. and Ehrgott, M. (2008c). Solving the intensity problem of radiotherapy treatment planning. In J. Haywood, editor, *Proceedings of the 43rd Annual Conference of the Operational Research Society of New Zealand*, pages 177–186. Operational Research Society of New Zealand.

A simulated annealing approach to the inventory routing problem

Mohammad J. Tarokh

Associate Professor, Dep. of Industrial Engineering, K.N.T University of
Technology, Tehran, Iran

mjtarokh@kntu.ac.ir

Nooraddin Dabiri

PhD student, Dep. of Industrial Engineering, K.N.T University of Technology,
Tehran, Iran

dabiri@dena.kntu.ac.ir

Mehdi Alinaghian

PhD student, Dep. of Industrial Engineering, Iran University of Science and
Technology, Tehran, Iran

alinaghian@iust.ac.ir

Abstract

Inventory Routing Problems (IRP) deal with how to manage the activity of supplying one or several goods from one or several origins to one or several destinations during some finite or infinite time horizon, considering both routing and inventory issues. The IRP is formulated as a mixed integer program with the objective of minimizing the costs associated with making deliveries in a specific time period to a widely dispersed set of customers. In this paper, we introduce a modified simulated annealing algorithm (SA) approach for the inventory routing problem. We present the suitable solution representation and moving method to improve the performance of SA algorithm. Finally, we compare our algorithm with genetic algorithm approach from literature. Our findings shows SA based method make some advantages.

Discovering Relationships between Scheduling Problem Structure and Heuristic Performance

Kate A. Smith-Miles¹, Ross J. W. James², John W. Giffin² and Yiqing Tu³

¹School of Mathematical Sciences, Monash University, Melbourne, Australia

kate.smith-miles@sci.monash.edu.au,

²Department of Management, University of Canterbury, Christchurch, New Zealand

ross.james@canterbury.ac.nz, john.giffin@canterbury.ac.nz

³School of Engineering and Information Technology, Deakin University, Burwood

Australia

ytu@deakin.edu.au

Abstract

Using a knowledge discovery approach, we seek insights into the relationships between problem structure and the effectiveness of scheduling heuristics. A large collection of 75,000 instances of the single machine early/tardy scheduling problem is generated, characterized by six features, and used to explore the performance of two common scheduling heuristics. The best heuristic is selected using rules from a decision tree with accuracy exceeding 97%. A self-organizing map is used to visualize the feature space and generate insights into heuristic performance. This paper argues for such a knowledge discovery approach to be applied to other optimization problems, to contribute to automation of algorithm selection as well as insightful algorithm design.

1 Introduction

It has long been appreciated that knowledge of a problem's structure and instance characteristics can assist in the selection of the most suitable algorithm or heuristic [1, 2]. The No Free Lunch theorem [3] warns us against expecting a single algorithm to perform well on all classes of problems, regardless of their structure and characteristics. Instead we are likely to achieve better results, on average, across many different classes of problem, if we tailor the selection of an algorithm to the characteristics of the problem instance.

As early as 1976, Rice [1] proposed a framework for the algorithm selection problem. There are four essential components of the model:

- the problem space P represents the set of instances of a problem class;
- the feature space F contains measurable characteristics of the instances generated by a computational feature extraction process applied to P ;
- the algorithm space A is the set of all considered algorithms for tackling the problem;
- the performance space Y represents the mapping of each algorithm to a set of performance metrics.

In addition, we need to find a mechanism for generating the mapping from feature space to algorithm space. The Algorithm Selection Problem can be formally stated as: For a given problem instance $x \in P$, with features $f(x) \in F$, find the selection mapping $S(f(x))$ into algorithm space A , such that the selected algorithm $\alpha \in A$ maximizes the performance mapping $y(\alpha(x)) \in Y$. The collection of data describing $\{P, A, Y, F\}$ is known as the meta-data.

There have been many studies in the broad area of algorithm performance prediction, which is strongly related to algorithm selection in the sense that supervised learning or regression models are used to predict the performance ranking of a set of algorithms, given a set of features of the instances. In the AI community, most of the relevant studies have focused on constraint satisfaction problems like SAT, using solvers like CPLEX or heuristics (A), and building a regression model (S) to use the features of the problem structure (F) to predict the run-time performance of the algorithms (Y). Studies of this nature include Leyton-Brown and co-authors [5-7]. In recent years these studies have extended into the algorithm portfolio approach [4] and a focus on dynamic selection of algorithm components in real-time [8, 9].

In the machine learning community, research in the field of meta-learning has focused on classification problems (P), solved using typical machine learning classifiers such as decision trees, neural networks, or support vector machines (A), where supervised learning methods (S) have been used to learn the relationship between the statistical and information theoretic measures of the classification instance (F) and the classification accuracy (Y). The term meta-learning [10] is used since the aim is to learn about learning algorithm performance.

For many NP-hard optimization problems, such as scheduling, there is a great deal we can discover about problem structure which could be used to create a rich set of features. Landscape analysis (see [11-12]) is one framework for measuring the characteristics of problems and instances, and there have been many relevant developments in this direction, but the dependence of algorithm performance on these measures is yet to be completely determined [13].

In this paper we present a methodology encompassing both supervised and unsupervised knowledge discovery processes on a large collection of meta-data to explore the problem structure and its impact on algorithm suitability. The problem considered is the early/tardy scheduling problem, described in Section 2. The methodology and meta-data is described in section 3, comprising 75,000 instances (P) across a set of 6 features (F). We compare the performance of two common heuristics (A), and measure which heuristic produces the lowest cost solution (Y). The mapping S is learned from the meta-data $\{P, A, Y, F\}$ using knowledge derived from self-organizing maps, and compared to the knowledge generated and accuracy of the performance predictions using the supervised learning methods of neural networks and decision trees. Section 4 presents the results of this methodology, including decision tree rules and visualizations of the feature space, and conclusions are drawn in Section 5.

2 The Early/Tardy Machine Scheduling Problem

Research into the various types of E/T scheduling problems was motivated, in part, by the introduction of Just-in-Time production, which required delivery of goods to be made at the time required. Both early and late production are discouraged, as early

production incurs holding costs, and late delivery means a loss of customer goodwill. A summary of the various E/T problems was presented in [14] which showed the NP-completeness of the problem.

2.1 Formulation

The E/T scheduling problem we consider is the single machine, distinct due date, early/tardy scheduling problem where each job has an earliness and tardiness penalty and due date. Once a job is dispatched on the machine, it runs to completion with no interruptions permitted. The objective is to minimise the total penalty produced by the schedule. The objective of this problem can be defined as follows:

$$\min \sum_{i=1}^n \left(\alpha_i |d_i - c_i|^+ + \beta_i |c_i - d_i|^+ \right). \quad (1)$$

where n is the number of jobs to be scheduled, c_i is the completion time of job i , d_i is the due date of job i , α_i is the penalty per unit of time when job i is produced early, β_i is the penalty per unit of time when job i is produced tardily, and $|x|^+ = x$ if $x > 0$, or 0 otherwise. We also define p_i as the processing time of job i . The decision variable is the completion time c_i of job i .

The objective of this problem is to schedule the jobs as closely as possible to their due dates; however the difficulty in formulating a schedule occurs when it is not possible to schedule all jobs on their due dates, which also causes difficulties in managing the many tradeoffs between jobs competing for processing at a given time [15]. Two of the simplest and most commonly used dispatching heuristics for the E/T scheduling problem are the Earliest Due Date and Shortest Processing Time heuristics.

2.2 Earliest Due Date (EDD) heuristic

The EDD heuristic orders the jobs based on the date the job is due to be delivered to the customer. The jobs with the earliest due date are scheduled first, while the jobs with the latest due date are scheduled last. After the sequence is determined, the completion times of each job are then calculated using the optimal idle time insertion algorithm of Fry, Armstrong and Blackstone [16]. For single machine problems, EDD is known to be the best rule to minimise the maximum lateness, and therefore tardiness, and also the lateness variance.

2.3 Shortest Processing Time (SPT) heuristic

The SPT heuristic orders the jobs based on their processing time. The jobs with the smallest processing time are scheduled first, while the jobs with the largest processing time are scheduled last; this is the fastest way to get most of the jobs completed quickly. Once the SPT sequence has been determined, the job completion times are calculated using the optimal idle time insertion algorithm [16]. The “weighted” version of the SPT heuristic, where the order is determined by p_i/β_i , is used in part by many E/T heuristics, as this order can be proven to be optimal for parts of a given schedule.

3 Methodology

In this section we describe the meta-data for the E/T scheduling problem in the form of {P, A, Y, F}. We also provide a description of the machine learning algorithms applied to the meta-data to produce rules and visualizations of the meta-data.

3.1 Meta-Data for the E/T Scheduling Problem

The most critical part of the proposed methodology is identification of suitable features of the problem instances that reflect the structure of the problem and the characteristics of the instances that might explain algorithm performance. Generally there are two main approaches to characterizing the instances: the first is to identify problem dependent features based on domain knowledge of what makes a particular instance challenging or easy to solve; the second is a more general set of features derived from landscape analysis [17]. In the case of the generalised single machine E/T scheduling problem however, there is sufficient differentiation power in a small collection of problem-dependent features that we can derive rules explaining the different performance of the two heuristics.

In this paper, each n -job instance of the generalised single machine E/T scheduling problem has been characterized by the following features.

1. Number of jobs to be scheduled in the instance, n
2. Mean Processing Time \bar{p} : The mean processing time of all jobs in an instance
3. Processing Time Range p_σ : The range (max – min) of the processing times of all jobs in the instance
4. Tardiness Factor τ : Defines where the average due date occurs relative to, and as a fraction of the total processing time of all jobs in the instance. A positive tardiness factor indicates the proportion of the schedule that is expected to be tardy, while a negative tardiness factor indicates the amount of idle time that is expected in the schedule as a proportion of the total processing time of all jobs in the sequence. Mathematically the tardiness factor was defined by Baker and Martin [18] as:
$$\tau = 1 - \frac{\sum d_i}{n \sum p_i}$$
5. Due Date Range factor D_σ : Determines the spread of the due dates from the average due date for all jobs in the instance, normalized by the size of processing times. It is defined as $D_\sigma = \frac{(b - a)}{\sum p_i}$, where b is the maximum due date in the instance and a is the minimum due date in the instance, and is a fraction of the total processing time needed for the instance
6. Penalty Ratio ρ : The maximum over all jobs in the instance of the ratio of the tardy penalty to the early penalty.

Any instance of the problem, whether contained in the meta-data set or generated at a future time, can be characterized by this set of six features. It is not the only possible set of features but, as the results presented later in this paper demonstrate, it captures the essential variation in instances needed to accurately predict heuristic performance. Since we are comparing the performance of only two heuristics, we can create a single binary variable to indicate which heuristic performs best for a given problem instance. Let $Y_i=1$ if EDD is the best performing heuristic (lowest objective function) compared to SPT for problem instance i , and $Y_i=0$ otherwise (SPT is best). The meta-data then comprises the set of six-feature vectors and heuristic performance measure (Y), for a large number of instances, and the task is to learn the relationship between features and heuristic performance.

In order to provide a large and representative sample of instances for the meta-data, an instance generator was created to span a range of values for each feature. Problem instances were then generated for all combinations of parameter values. Note that these

parameters are targets for the instances and the random generation process may create slight variation from these target values. The parameter settings used were:

- problem size (number of jobs, n): 20-100 with increments of 20 (5 levels)
- target processing time range p_σ : processing times randomly generated with a range ($p_{\max} - p_{\min}$) of 2-10 with increments of 2 (5 levels).
- target due date range factor D_σ as a proportion of total processing time: ranges from 0.2 to 1 in increments of 0.2 (5 levels)
- target tardiness factor τ as a proportion of total processing time: ranges from 0 (all jobs should complete on time) to 1 (all jobs should be late) in increments of 0.2 (6 levels)
- penalty ratio ρ : 1-10 with increments of 1 (10 levels)

From these parameters the following instance data can be generated:

- processing times p_i : calculated within the processing time range.
- processing time means \bar{p} : calculated from the randomly generated p_i
- due dates d_i : due dates randomly generated within the due date range and offset by the tardiness factor.

To calculate the actual p_σ , actual D_σ and actual τ we use the actual p_i , d_i of the problem rather than the target values. Ten instances using each parameter setting were then generated, giving a total of 5 (size levels) x 5 (processing time range levels) x 6 (tardiness factor levels) x 5 (due date range factor levels) x 10 (penalty ratio levels) x 10 (instances) = 75,000 instances.

A correlation analysis between the instance features and the Y values across all 75,000 instances reveals that the only instance features that appear to correlate (linearly) with heuristic performance are the tardiness factor (correlation = -0.59) and due date range factor (correlation = 0.44). None of the other instance features appear to have a linear relationship with algorithm performance. Clearly due date range factor and tardiness factor correlate somewhat with the heuristic performances, but it is not clear if these are non-linear relationships, and if either of these features with combinations of the others can be used to seek greater insights into the conditions under which one heuristic is expected to outperform the other.

Using Rice's notation, our meta-data can thus be described as:

- P = 75,000 E/T scheduling instances
- A = 2 heuristics (EDD and SPT)
- Y = binary decision variable indicating if EDD is best compared to SPT (based on objective function which minimizes weighted deviation from due dates)
- F = 6 instance features (problem size, processing time mean, processing time range, due date range factor, tardiness factor and penalty ratio).

3.2 Knowledge Discovery on the Meta-Data

When exploring any data-set to discover knowledge, there are two broad approaches. The first is supervised learning (aka directed knowledge discovery) which uses training examples – sets of independent variables (inputs) and dependent variables (outputs) - to learn a predictive model which is then generalized for new examples to predict the dependent variable (output) based only on the independent variables (inputs). This approach is useful for building models to predict which algorithm or heuristic is likely to perform best given only the feature vector as inputs. The second broad approach to knowledge discovery is unsupervised learning (aka undirected knowledge discovery) which uses only the independent variables to find similarities and differences between

the structure of the examples, from which we may then be able to infer relationships between these structures and the dependent variables. This second approach is useful for our goal of seeking greater insight into *why* certain heuristics might be better suited to certain instances and, rather than just building predictive models of heuristic performance.

In this section we briefly summarise the machine learning methods we have used for knowledge discovery on the meta-data.

Neural Networks.

As a supervised learning method [19], neural networks can be used to learn to predict which heuristic is likely to return the smallest objective function value. A training dataset is randomly extracted (80% of the 75,000 instances) and used to build a non-linear model of the relationships between the input set (features F) and the output (metric Y). Once the model has been learned, its generalisation on an unseen test set (the remaining 20% of the instances) is evaluated and recorded as a percentage accuracy in predicting the superior heuristic. This process is repeated ten times for different random extractions of the training and test sets, to ensure that the results were not simply an artifact of the random number seed.

Decision Tree

A decision tree [20] is also a supervised learning method, which uses the training data to successively partition the data, based on one feature at a time, into classes. The goal is to find features on which to split the data so that the class membership at lower leaves of the resulting tree is as “pure” as possible. In other words, we strive for leaves that are comprised almost entirely of one class only. The rules describing each class can then be read up the tree by noting the features and their splitting points. Ten-fold cross validation is also used in our experiments to ensure the generalisation of the rules.

Self-Organizing Maps

Self-Organizing Maps (SOMs) are the most well-known unsupervised neural network approach to clustering. Their advantage over traditional clustering techniques such as the k-means algorithm lies in the improved visualization capabilities resulting from the two-dimensional map of the clusters. Often patterns in a high dimensional input space have a very complicated structure, but this structure is made more transparent and simple when they are clustered in a lower dimensional feature space. Kohonen [21] developed SOMs as a way of automatically detecting strong features in large data sets. SOMs find a mapping from the high dimensional input space to low dimensional feature space, so any clusters that form become visible in this reduced dimensionality.

For our experiments we randomly split the 75000 instances into training data (50000 instances) and test data (25000 instances). We use the Viscovery SOMine software (www.eudaptics.com) to cluster the instances based only on the six features as inputs. After the clustering of the training instances, the distribution of Y values is examined within each cluster, and knowledge about the relationships between instance structure and heuristic performance is inferred and evaluated on the test data.

4 Experimental Evaluation

4.1 Supervised Learning Results

Both the neural network and decision tree algorithms were able to learn the relationships in the meta-data, achieving greater than 97% accuracy (on ten-fold cross-validation test sets) in predicting which of the two heuristics would be superior based

only on the six features (inputs). These approaches have an overall classification accuracy of 97.34% for the neural network and 97.13% for the decision tree. While the neural network can be expected to learn the relationships in the data more powerfully, due to its nonlinearity, its limitation is the lack of insight and explanation of those relationships. The decision tree's advantage is that it produces a clear set of rules, which can be explored to see if any insights can be gleaned. The decision tree rules are presented in the form of pseudo-code in Figure 1, with the fraction in brackets showing the number of instances that satisfied both the condition and the consequence (decision) in the numerator, divided by the total number of instances that satisfied the condition in the denominator. This proportion is equivalent to the accuracy of the individual rule.

The results allow us to state a few rules with exceptionally high accuracy:

- 1) If the majority of jobs are expected to be scheduled early (tardiness factor ≤ 0.5) then EDD is best in 99.8% of instances
- 2) If the majority of the jobs are expected to be scheduled late (tardiness factor > 0.7) then SPT is best in 99.5% of instances
- 3) If slightly more than half of the jobs are expected to be late (tardiness factor between 0.5 and 0.7) then as long as the tardiness penalty ratio is no more than 3 times larger than the earliness penalty ($\rho \leq 3$), then EDD is best in 98.9% of the instances with a due date range factor greater than 0.2.

The first two rules are intuitive and can be justified from what we know about the heuristics - EDD is able to minimise lateness deviations when the majority of jobs can be scheduled before their due date, and SPT is able to minimise the time of jobs in the system and hence tardiness when the majority of jobs are going to be late. The third rule reveals the kind of knowledge that can be discovered by adopting a machine learning approach to the meta-data. Of course other rules can also be explored from Figure 1, with less confidence due to the lower accuracy, but they may still provide the basis for gaining insight into the conditions under which different algorithms can be expected to perform well.

```

If ( $\tau \leq 0.7$ ) Then
  If ( $\tau \leq 0.5$ ) Then EDD best (44889/45000 = 99.8%)
  If ( $\tau > 0.5$ ) Then If ( $D_\sigma \leq 0.2$ ) Then If ( $\rho \leq 3$ ) Then EDD best (615/750 = 82.0%)
    Else SPT best (1483/1750 = 84.7%)
    Else If ( $\rho \leq 3$ ) Then EDD best (5190/5250 = 98.9%)
      Else If ( $\tau \leq 0.6$ ) Then EDD best (8320/8750 = 95.1%)
        Else If ( $\bar{p} \leq 2$ ) Then EDD best (556/700 = 79.4%)
          Else If ( $n \leq 60$ ) Then SPT best (1150/1680 = 68.4%)
            Else EDD best (728/1120 = 65%)
          Else SPT best (9950/10000 = 99.5%)

```

Figure 1. Pseudo-code for the decision tree rule system, showing the accuracy of each rule

4.2 Unsupervised Learning Results

After training the SOM, the converged map shows 5 clusters, each of which contains similar instances defined by Euclidean distance in feature space. Essentially, the six-dimensional input vectors have been projected onto a two-dimensional plane, with topology-preserving properties. The clusters can be inspected to understand what the

instances within each cluster have in common. The statistical properties of the 5 clusters can be seen in Table 1. The distribution of the input variables (features), and additional variables including the performance of the heuristics, can be visually explored using the maps shown in Figure 2.

Looking first at the bottom row of Table 1, it is clear that clusters 1, 2 and 3 contain instances that are best solved using EDD ($Y=1$). These clusters are shown visually in the bottom half of the 2-d self-organizing map (see Figure 2a for cluster boundaries, and Figure 2b to see the distribution of Y across the clusters). These three clusters of instances account for 70.2% of the 75,000 instances (see Table 1). The remaining clusters 4 and 5 are best solved, on average, by SPT. The maps shown in Figure 2c – 2h enable us to develop a quick visual understanding of how the clusters differ from each other, and to see which features are prominent in defining instance structure.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	All Data
instances	17117 (8483)	10454 (5236)	7428 (3832)	8100 (4000)	6901 (3449)	50000 (25000)
instances (%)	34.23 (33.93)	20.91 (20.94)	14.86 (15.33)	16.2 (16.0)	13.8 (13.8)	100 (100)
n	60.65 (61.03)	59.73 (59.73)	58.73 (58.96)	57.8 (57.7)	63.39 (61.56)	60.0 (59.97)
\bar{p}	2.77 (2.76)	5.24 (5.22)	5.08 (5.07)	5.12 (5.11)	2.70 (2.71)	4.0 (3.99)
$p\sigma$	3.54 (3.52)	8.48 (8.45)	8.16 (8.13)	8.24 (8.21)	3.41 (3.41)	6.0 (5.99)
τ	0.31 (0.31)	0.36 (0.35)	0.21 (0.21)	0.72 (0.73)	0.72 (0.72)	0.43 (0.42)
$D\sigma$	0.70 (0.70)	0.88 (0.88)	0.38 (0.38)	0.40 (0.39)	0.40 (0.40)	0.6 (0.59)
ρ	5.89 (5.88)	4.93 (4.99)	5.37 (5.41)	5.24 (5.19)	5.87 (5.72)	5.5 (5.49)
Y	1.00 (0.99)	1.00 (1.00)	0.99 (0.99)	0.36 (0.36)	0.42 (0.41)	0.82 (0.82)

Table 1. Cluster statistics for training data (test data in brackets) - mean values of input variables, and heuristic performance variable Y , as well as cluster size.

By inspecting the maps shown in Figure 2, and the cluster statistics in Table 1, we can draw some conclusions about whether the variables in each cluster are above or below average (compared to the entire dataset), and look for correlations with the heuristic performance metric Y . For instance, cluster 2 is characterized by instances with above average values of processing time mean and range, below average tardiness factor, and above average due date range factor. The EDD heuristic is always best under these conditions ($Y=1$). Instances in cluster 3 are almost identical, except that the due date range factor tends to be below average. Since cluster 3 instances are also best solved by the EDD heuristic, one could hypothesize that the due date range factor does not have much influence in predicting heuristic performance. An inspection of the maps, however, shows this is not the case.

The distribution of Y across the map (Figure 2b) shows a clear divide between the clusters containing instances best solved using EDD (bottom half) and the clusters containing instances best solved using SPT (top half). Inspecting the distribution of features across this divide leads to a simple observation that, if the tardiness factor τ is below average (around 0.5 represented by white to mid-grey in Figure 2f), then EDD will be best. But there are small islands of high Y values in clusters 4 and 5 that overlay nicely with the medium grey values of due date range factor. So we can observe another rule that EDD will also be best if the tardiness factor is above average and the due date range factor is above average. Also of interest, from these maps we can see that problem size and the penalty ratio do not influence the relative performance of these heuristics. As neither heuristic considers the penalty ratio (it is used within the optimal idle time

insertion algorithm [16], common to both heuristics, but not used by the EDD or SPT heuristics themselves), its not being a factor in the clusters is not surprising.

Within Viscovery SOMine, specific regions of the map can be selected, and used as the basis of a classification. In other words, we can define regions and islands to be predictive of one heuristic excelling based on the training data (50,000 instances). We can then test the generalization of the predictive model using the remaining 25,000 instances as a test set, and applying the k-nearest neighbour algorithm to determine instances that belong to the selected region. We select the dark-grey to black regions of the Y map in Figure 2b, and declare that any test instances falling in the selected area are classified as $Y=1$, while any instances falling elsewhere in the map are classified as $Y=0$. The resulting accuracy on the test set is 95% in predicting which heuristic will perform better. The self-organizing map has proven to be useful for both visualization of feature space and predictive modeling of heuristic performance, although the accuracy is not quite as high as the supervised learning approaches.

5 Conclusions and Future Research

In this paper we have illustrated how the concepts of Rice's Algorithm Selection Problem can be extended within a knowledge discovery framework, and applied to the domain of heuristics in order that we might gain to insights into heuristic performance. While only two very simple heuristics have been used to illustrate the approach, we expect full generalization of the methodology to consider a broader range of complex heuristics and meta-heuristics. Both supervised and unsupervised learning approaches have been presented, each with their own advantages and disadvantages made clear by the empirical results. The neural network obtained the highest accuracy for performance prediction, but its weakness is the lack of explanation or interpretability of the model. Our goal is not merely performance prediction, but to gain insights into the characteristics of instances that make solution by one heuristic superior than another. Decision trees are also a supervised learning method, and the rules produced demonstrate the potential to obtain both accurate performance predictions and some insights. Finally, the self-organizing map demonstrated its benefits for visualization of the meta-data and relationships therein.

One of the most important considerations for this approach to be successful for any arbitrary optimization problem is the choice of features used to characterize the instances. These features need to be carefully chosen in such a way that they can characterize instance and problem structure as well as differentiate algorithm performance.

There is little that will be learned via a knowledge discovery process if the features selected to characterize the instances do not have any differentiation power. The result will be supervised learning models of algorithm performance that predict average behaviour with accuracy measures no better than the default accuracies one could obtain from using a naïve model. Likewise, the resulting self-organizing map would show no discernible difference between the clusters when superimposing Y values (unlike in Figure 2b where we obtain a clear difference between the top and bottom halves of the map). Thus the success of any knowledge discovery process depends on the quality of the data, and in this case, the meta-data must use features that serve the purpose of differentiating algorithm performance.

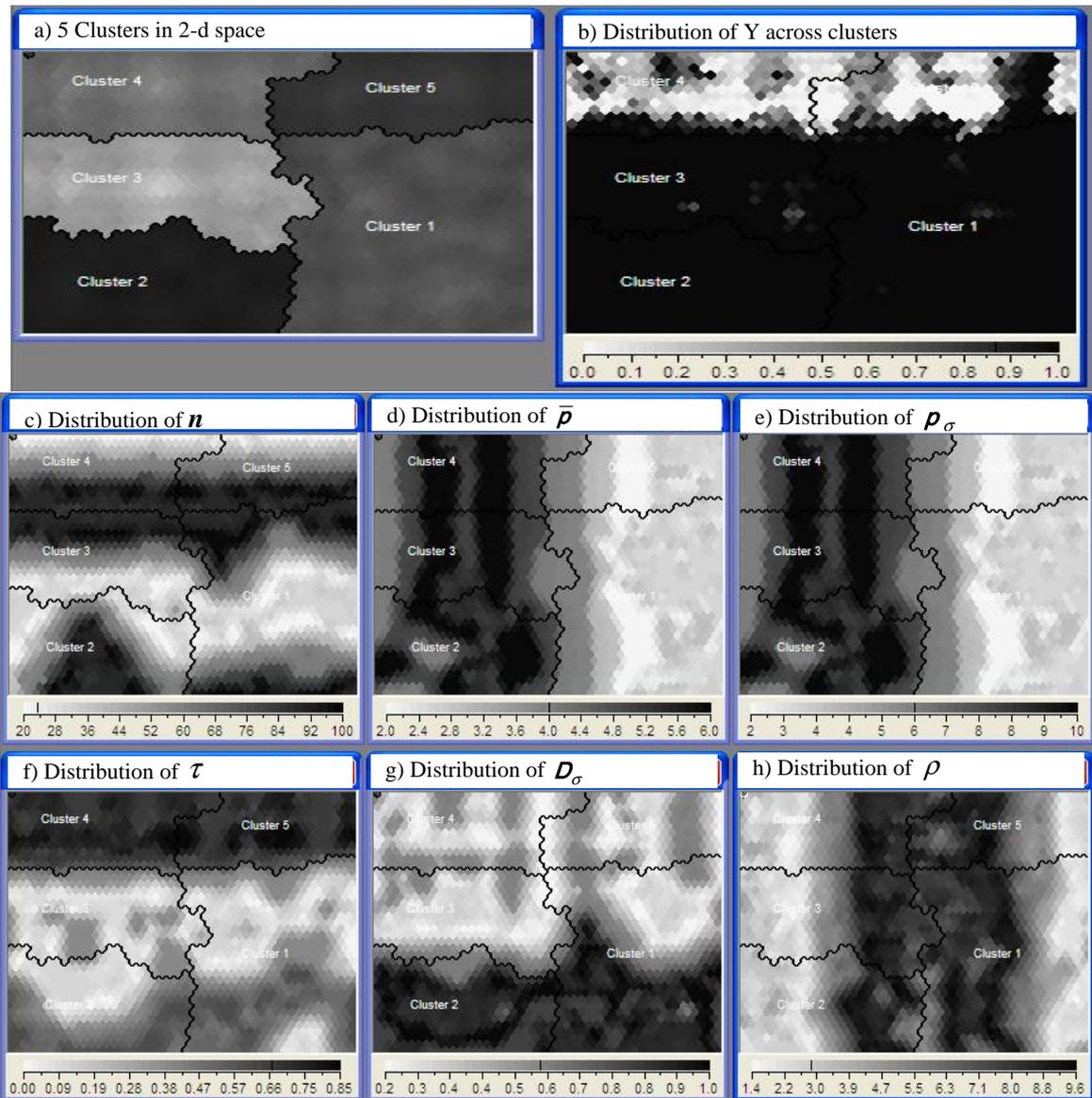


Figure 2. Self-Organizing Map showing 5 clusters (fig. 2a), the heuristic performance variable Y (fig 2b), and the distribution of six features across the clusters (fig 2c – fig 2h). The grey scale shows a minimum value as white, and maximum value as black.

6 References

1. Rice, J. R.: The Algorithm Selection Problem. *Adv. Comp.* 15, 65--118 (1976)
2. Watson, J.P., Barbulescu, L., Howe, A.E., Whitley, L.D.: Algorithm Performance and Problem Structure for Flow-shop Scheduling. In: *Proc. AAAI Conf. on Artificial Intelligence*, pp. 688--694 (1999)
3. Wolpert, D.H., Macready, W.G.: No Free Lunch Theorems for Optimization. *IEEE T. Evolut. Comput.* 1, 67 (1997)
4. Xu, L., Hutter, F., Hoos, H., Leyton-Brown, K.: *Satzilla-07: The Design And Analysis Of An Algorithm Portfolio For SAT*. LNCS, vol. 4741, pp. 712--727 (2007)
5. Leyton-Brown, K., Nudelman, E., Shoham, Y.: *Learning the Empirical Hardness of Optimization Problems: The Case of Combinatorial Auctions*. LNCS, vol. 2470. pp. 556--569. Springer, Heidelberg (2002)

6. Leyton-Brown, K., Nudelman, E., Andrew, G., McFadden, J., Shoham, Y.: A Portfolio Approach to Algorithm Selection. In: Proc. IJCAI. pp. 1542--1543 (2003)
7. Nudelman, E., Leyton-Brown, K., Hoos, H., Devkar, A., Shoham, Y.: Understanding Random SAT: Beyond the Clauses-To-Variables Ratio. LNCS, vol. 3258, pp. 438--452 (2004)
8. Samulowitz, H., Memisevic, R.: Learning to solve QBF. In: Proc. 22nd AAAI Conf. on Artificial Intelligence, pp. 255--260 (2007)
9. Streeter, M., Golovin, D., Smith, S. F.: Combining multiple heuristics online. In: Proc. 22nd AAAI Conf. on Artificial Intelligence, pp. 1197--1203 (2007)
10. Vilalta, R., Drissi, Y.: A Perspective View and Survey of Meta-Learning. *Artif. Intell. Rev.* 18, 77--95 (2002)
11. Knowles, J. D., Corne, D. W.: Towards Landscape Analysis to Inform the Design of a Hybrid Local Search for the Multiobjective Quadratic Assignment Problem. In: Abraham, A., Ruiz-Del-Solar, J., Koppen M. (eds.) *Soft Computing Systems: Design, Management and Applications*, pp. 271--279. IOS Press, Amsterdam (2002)
12. Merz, P.: Advanced Fitness Landscape Analysis and the Performance of Memetic Algorithms. *Evol. Comp.*, 2, 303--325 (2004)
13. Watson, J., Beck, J. C., Howe, A. E., Whitley, L. D.: Problem Difficulty for Tabu Search in Job-Shop Scheduling. *Artif. Intell.* 143, 189--217 (2003)
14. Baker, K.R., Scudder, G.D.: Sequencing With Earliness and Tardiness Penalties: A Review. *Ops. Res.*, 38, 22--36 (1990)
15. James, R. J. W., Buchanan, J. T.: A Neighbourhood Scheme with a Compressed Solution Space for The Early/Tardy Scheduling Problem. *Eur. J. Oper.Res.* 102, 513--527 (1997)
16. Fry T.D., Armstrong R.D., Blackstone J.H.: Minimizing Weighted Absolute Deviation in Single Machine Scheduling. *IIE Transactions*, 19, 445--450 (1987)
17. Schiavinotto, T., Stützle, T.: A review of metrics on permutations for search landscape analysis. *Comput. Oper. Res.* 34, 3143--3153 (2007).
18. Baker K. B., Martin, J. B.: An Experimental Comparison of Solution Algorithms for the Single Machine Tardiness Problem. *Nav. Res. Log.* 21, 187--199 (1974)
19. Smith, K. A.: *Neural Networks for Prediction and Classification*. In: Wang, J.(ed.), *Encyclopaedia of Data Warehousing And Mining*. vol. 2, pp. 86--869, Information Science Publishing, Hershey PA (2006)
20. Quinlan, J. R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco (1993)
21. Kohonen, T.: Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cyber.* 43, 59--69 (1982)

Educating the world about OR with viewer-paced videos on Youtube

Nicola Ward Petty
Department of Management
University of canterbury
New Zealand
nicola.petty@canterbury.ac.nz

Abstract

A series of ten videos designed to teach aspects of Excel, linear programming, and inference receive over 300 views daily. In this presentation I explain how these videos have been effective, give guidelines for developing such clips and discuss the educational and promotional possibilities of viewer-paced short video clips.

As is often the case, there is a wide range of prior knowledge among the students in the first year Quantitative Methods for Business course and in the Management Science course taught at the University of Canterbury (NZ). Many students struggle with mathematical tasks, and are resistant to learning quantitative concepts. There is often also a reluctance to use Excel.

There are certain concepts or skills that many students find difficult, and instructors found themselves repeating very similar explanations many times to individual students. A video was developed, with still shots, narration and screen capture and using an imaginary example from business, to teach about relative and absolute references in Excel. The video was uploaded onto Youtube so that students could gain easy access to it.

The success of the first video led to a series of videos teaching Excel, Linear Programming and statistical concepts. Uploaded to the UCMSCI Youtube account (www.youtube.com/UCMSCI), these have been well received by students in the class, and thousands of others worldwide.

The videos were developed using principles of good multimedia instructional design. Key features that appear to have led to their success is their short length, use of humour, no talking heads, conversational narrative tone and the facility for viewers to control the viewing experience by pausing and repeating the clips.

Operations Research as a discipline suffers from lack of “brand recognition”. These videos, and others to be developed will help to inform and educate the next generation about Operations Research.

Key words: Excel, teaching of OR.

Eating the Elephant!

Applying Theory of Constraints in Assurance of Learning

Victoria J. Mabin
Associate Dean (Teaching and Learning)
Faculty of Commerce and Administration
Victoria University of Wellington
New Zealand
vicky.mabin@vuw.ac.nz

Abstract

Setting up an 'Assurance of Learning' (AoL) system is a non-trivial task, with its combination of technical challenges, academic challenges, budgetary pressures and political feasibility. It's not something that you manage in one bite! So how do you 'eat the AoL elephant?' Having recently set about creating processes and leading our teaching and learning activities, including the Assurance of Learning required for accreditation by AACSB, our Faculty has encountered a steep learning curve and numerous challenges, but have received high praise for our processes. One key philosophy used to guide the process was Theory of Constraints (TOC) - a management philosophy with a set of thinking tools. In this paper, I will share some of the challenges, together with some of the TOC tools that we have found useful, such as the Evaporating Cloud for thinking through dilemmas, the Prerequisite Tree for planning to achieve targets and the Five Focusing Steps for focusing improvements. Using TOC to harness resistance to change is also discussed.

Key words: Theory of Constraints, Thinking Processes, Accreditation, Teaching & Learning.

1. Introduction

Our business school has recently established Assurance of Learning systems and processes in line with requirements for accreditation by an international organisation, The Association to Advance Collegiate Schools of Business, AACSB. The crucial difference between this and other curriculum alignment processes is the focus on direct measurement of student achievement of specific student learning objectives. Such direct measurements provide valuable information on gaps in student learning, which are then rectified by changes to systems, curricula, delivery of material, or assessment.

While this may appear to be merely a matter of following AACSB's standards and recommended procedures for Assurance of Learning, there is ample room for business schools to create/choose their own processes. Indeed, this is necessary. Rather than provide clear guidance, the AACSB mantra is that everything must be related to the organisation's mission. As a consequence, AACSB accreditation is far from a prescriptive 'cookie-cutter' approach. There is ample room for interpretation of the standards and recommended elements contained in the standards. The resulting challenge is to set up an Assurance of Learning system that is right for the

organisation, and meets the standards. There will undoubtedly be many issues to be resolved on the way, as evidenced by the AACSB Listserv discussions.

We have developed a system that received high praise from the accreditation teams visiting our institution. This paper describes the approach we have adopted, some of the challenges we have encountered, and how we have approached those.

“All models are wrong; some are useful” George E.P. Box

2. Assurance of Learning

The Assurance of Learning Process can be depicted in Figure 1 - in effect, it is a combination of a typical Plan-Do-Check-Act cycle (Deming 1982) and a curriculum alignment process. (Wiggins and McTighe, 1998, 2005)

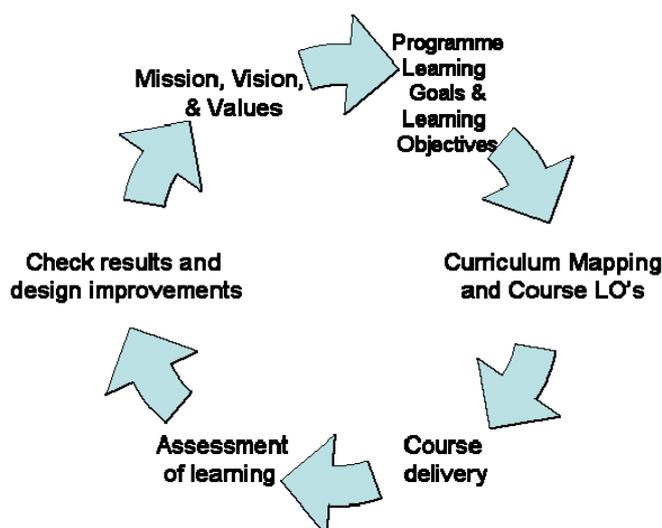


Figure 1: The Assurance of Learning Cycle

The 7-step cycle involves:

1. *Mission, Vision and Values*

This is usually set by the business school after considerable debate and input from all stakeholders.

2. *Programme Learning Goals and learning objectives*

These are determined for all programmes – ideally 4-10 Goals. Objectives are specific activities or components contributing to a goal.

3. *Curriculum Mapping and Course Learning Objectives*

Reviewing the Curriculum Aligning to LG's, mapping, streamlining. Some schools treat this as an administrative exercise, working from stated course learning objectives. In our case, we found we needed to also consult with staff to get the level of detail required. The map developed provides a basis for ongoing dialogue.

4. *Course delivery*

Including assessment of student work for grades Here course delivery needs care and thought to balance coverage of both content and learning goals, with sensible and consistent workloads for both staff and students.

5. *Assessment of student learning of the Learning Goals and Objectives*

For assurance of learning, separate assessment against learning goals/objectives is conducted on a sample of student work to determine the levels of achievement of learning goals and objectives. Measurement is usually performed using a 'rubric'

developed by a group of staff (not just the instructor). Statistics are collected and analysed.

6. Checking results to determine gaps in student learning, prioritise and design improvements.

Results of the assessments are analysed, individually as well as looking for patterns over groups of courses. Changes are selected and implemented.

7. Repeat the cycle ... starting from 1 if appropriate, otherwise 2.

2.1 Institutional background

In 2006, Victoria University of Wellington (VUW) embarked on a review of its undergraduate programmes referred to as Pathways to Success (PTS), which aimed to provide a framework to enable VUW to implement its goal of equipping its graduates with its set of Graduate Attributes. During the same period, the Faculty of Commerce and Administration started on its pursuit of AACSB accreditation. The two activities were distinct but compatible in their aims, creating synergies that were exploited in developing processes for each aim. As part of the PTS review process, we developed various systems diagrams, attempting to capture the various interacting components in the university system which needed to work together to produce the graduate attributes. A summary version of this is shown in Figure 2.

Subsequent processes prompted by AACSB accreditation, including strategic planning exercises, have built on the work from PTS. Thus, by mid 2008, the FCA had established its Mission, Learning Goals and Objectives for its main undergraduate degree - the BCA, restructured its BCA Core offerings, and adopted a change to simplify course points/credit values across the university. The FCA had also started exploring the assessment of achievement of learning goals and performed its first, albeit rudimentary, ‘Assurance of Learning’ exercises.

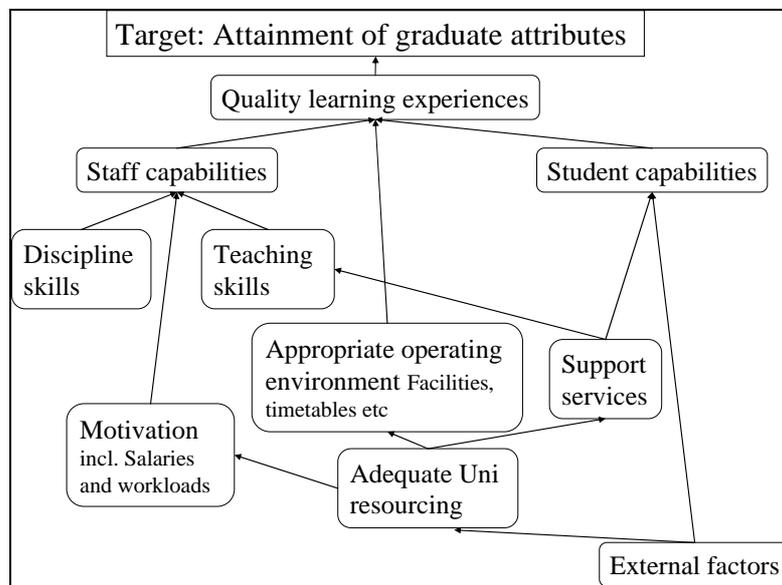


Figure 2 Intermediate Objectives Map developed for 'Pathways to Success'

In the quality movement in the 1980s, some organisations attracted to ISO certification treated it as a compliance exercise and gained little advantage from their efforts. However the more enlightened organisations treated ISO certification as a

lever to improved performance (Deming 1982). Likewise, while we could regard the AoL processes developed for AACSB Accreditation simply as a requirement for accreditation, the real benefit comes from the impetus it provides for improvements throughout our business school, and even beyond into other parts of the university.

2.2 The Philosophy guiding development of our AoL system

Next, we will outline contextual factors that led to the philosophy we adopted in undertaking this task and some of the challenges we have encountered and addressed. Universities are not dissimilar to other business environments – we are constrained by international economic and political pressures and a strengthening exchange rate that erodes our international student fee income, together with internal economic pressures, shrinking budgets, and rising costs. Being in the tertiary education sector, we have increasingly uncertain and generally declining funding from Government, coupled with strict rules which constrain our ability to respond in ways available to free enterprise, which makes it a very difficult time for universities. Ernest Rutherford's comment, "We don't have a lot of money, so we have to think!" seems to apply equally well to us today.

Furthermore we did not have a lot of time – for example, our Annual Progress Report was due less than two months after the date of my secondment into the role of AD (T&L), with responsibility for setting up an AoL system.

3 Theory of Constraints (TOC)

Being so constrained, the Theory of Constraints (TOC) was a natural framework to adopt. Within the TOC methodology are several methods/suites of tools. The Five Focusing Steps method is perhaps the best known, as popularised through the business novel *The Goal*, (Goldratt and Cox 1992). Excellent descriptions may also be found in (Scheinkopf 1999) and (Dettmer 1997), while the results of approximately 100 applications are detailed and summarised in Mabin and Balderstone (2000) and further analysed in Mabin and Balderstone (2003).

A second very popular suite of TOC tools are collectively known as the Thinking Processes (Goldratt 1994; Scheinkopf 1999; Dettmer 2007). They comprise 5 main logic diagrams which guide the process of managing change: through diagnosing what needs to change, what it needs to change to, and how to cause the change. Later variants add Why change?, How to sustain the change?, and How to develop a process of ongoing improvement? The use of these tools has been surveyed in Kim, Mabin et al. (2008). A review of twenty years of TOC is available in Watson, Blackstone et al. (2007).

This paper will describe some of the tools from TOC that have been very helpful when thinking through the strategic and day-to-day decisions encountered in this role. In addition to the Five Focussing Steps, we outline two of the Thinking Process tools, the 'Prerequisite Tree' for achieving an ambitious target; and the 'Evaporating Cloud' for resolving dilemmas. We have also used the Negative Branch Reservation method – a sub-tree of the Future Reality Tree – as described briefly later. In addition, managing resistance to change and the TOC approach to this will be discussed. The project reported on is an on-going one – the shaping of this new role and the changing world of teaching and learning.

3.1 TOC's 5 focusing steps

The Five Focusing Steps can/have been applied on a number of levels; one example will be provided here. First, one defines the system, its purpose/goal and how progress is to be measured. In this case, we will consider the FCA to be the system; the goal to demonstrate the FCA has established an AoL system; and progress to be measured by satisfactory performance in the AACSB accreditation review process.

Step 1. Identify the constraint

While significant progress had been made by my predecessors, in the absence of anyone dedicated to this role, our AoL comprised a few elements that needed to be built into a system. We had scant data to include in our Annual Progress Report (APR), and little time to gather this. Additionally, being new to the role, my knowledge of AoL was exceedingly limited. By the first month, it became clear that there was a need for Learning Goals to be finalised across programmes, assessment plans, and assessment data. Most importantly, we required evidence of 'closing the loop' ie. use of that data to make useful changes. So what was the constraint? In retrospect, the effective resource devoted to AoL, which was governed by the personnel devoted to the task, together with their knowledge of AoL, and time they had available.

Step 2. Exploit the constraint

We made the most of the knowledge we did have in house, and from our mentor. We built on the examples we had developed: learning goals were formalised across all programmes; we made the most of the few assessments we had done, and replicated those in a number of other courses; assessment plans were drawn up; one department was ahead of the others, and they ran some extra assessments; lessons were drawn from the assessments already done allowing indicative remedial 'closing the loop' actions to be identified.

Step 3. Subordinate other activities

There was no time to try to do other tasks – for example, research and teaching duties were abandoned, while this target was met. A trip originally meant for academic research was used to find out more about AoL from one of our 'aspirant' universities. There was an intense focus on achieving the goal.

Step 4. Elevate the constraint

The creation of an Associate Dean (Teaching and Learning) position had been a significant 'elevation', as had the appointment of a new administrator who stepped ably into the role. Load was spread over the relevant committees. We borrowed ideas from other Universities. External assessors provided input.

Step 5 Go back

The next most important constraint we identified was Staff resistance – very few people were driving or supporting the process. This is tackled in the next section.

3.2 Prerequisite Tree

Once the Annual Progress Report was finished, the AoL team met again to discuss how we were going to achieve our target of achieving accreditation. While we could have written a list of tasks, we chose the TOC Prerequisite Tree approach which instead starts out by asking participants to tap into their intuition to identify obstacles that stand in the way. These will often be either more personal or interpersonal issues that need to be addressed, but in reality more often remain unspoken and unaddressed in traditional list-making approaches. Scheinkopf (1999) provides clear guidance on using the PRT approach.

To tackle our target of achieving accreditation, we first stated the target clearly, then listed obstacles and then identified actions or outcomes that we needed to

achieve if we were to achieve our target. Table 1 shows the list of obstacles and Intermediate Objectives (IO's).

Table 1: PRT List of Obstacles and Intermediate Objectives (no order):	
Target: Achieve AACSB Accreditation	
Obstacles	Intermediate Objectives
1. Dean must understand and lead the process	Dean to be visible and connected internally and externally
2. Faculty/Staff not engaged; some even antagonistic	More informal interactions - investigate morning teas
3. Meetings not run well, not focused enough	Inter-meeting discussions; save meetings for decision-making and actions
4. "It's not our responsibility"	Develop a collegial culture Quality is everybody's job
5. Assurance of learning -assessment reports, T & L reports	Stocktake - convene meetings per Learning Objective. compile reports
6. Competitive environment (School vs School)	Develop a collegial culture. More informal interaction between schools
7. 12th floor viewed as distant, aloof	Develop a collegial culture drawing faculty together. More informal interaction within faculty. Inclusivity in T & L; seeing benefits to staff & students
8. Accreditation seen as compliance	Get Mary & others to spread glad tidings about accreditation, more people to seminars. Build enthusiasm; identify early adopters; pick easy tasks, quick wins, appreciate effort. Get AoL activities into promotions
9. Resentment about overhead burden	Demonstrate more value out of overheads. Inclusivity in T & L; seeing benefits to staff & students
10. AoL processes in place by end of 2008, start with core then new majors	Prevent unwitting or deliberate sabotage - templates, blueprints
11. AACSB Accounting Standards	Harvest info
12. A lot to do	Time management
13. School focus, Faculty is still an administrative notion	Develop a collegial culture
14. Differing instructions from the top	Unanimous voice, shared responsibility
15. Committee attendance is uneven	Reward Committee members
16. Uni is not supportive/facilitative of consulting under Uni's name	Easy "send us your bill " system. Town-gown

The next step is to arrange the IO's in a tree based on precedence ordering or necessary sequences of events/actions. Quite often, IO's identified will overcome more than one obstacle. We combined similar IO's into one action, and sequenced this smaller set of actions rather than IO's, resulting in the PRT shown in Figure 3. The arrows are read using "We must achieve <lower IO> (in order to get over the relevant obstacle) before we can do <upper IO> logic.

The resulting tree is not strictly a PRT as it incorporates a feedback loop: "We need to Build support for AoL internally *before we can* Improve delivery of courses

before we can Improve T&L outcomes, before we can See more value out of overheads, before we can Build (more) support for AoL internally”. Such a feedback loop would normally be part of a Future Reality Tree, and we have taken a liberty in incorporating this feature here, and have additionally annotated this feedback loop using the blue arrow – labelled R for ‘reinforcing’ – borrowed from Causal Loop Diagramming (Senge 1990). The finished diagram provided a set of guidelines for action, and a way of seeing the steps required to achieve the final aim. Normally the obstacles being overcome would be shown attached to each arrow of the tree, but these have been omitted in this summary diagram for simplicity.

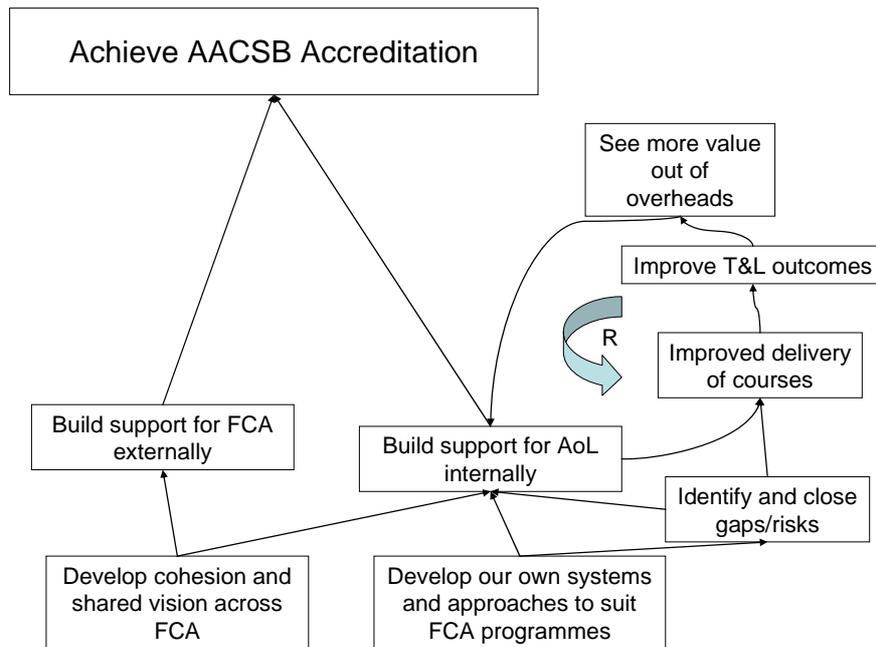


Figure 3: A modified Prerequisite Tree (PRT) for Achieving Accreditation.

PRT’s have also been used elsewhere in our AoL work. One major focus for improvement has been written communication skills, as part of the Graduate Attribute and Learning Goal relating to Communications skills. In crafting the strategy, we identified the need to find out what obstacles students themselves faced. The PRT process was very helpful in eliciting information from students, both on the obstacles preventing students from writing good assignments, as well as suggesting actions or outcomes they thought were needed to overcome these obstacles. The following themes came through very clearly from the student groups:

Table 2: Obstacles emerging from Student Groups
Time-Management
Lack of Motivation
Language skills
Lack of experience, information, feedback, understanding

We had expected language and technical skills to rate highly, and were surprised at the number of comments about time management and motivation. Hence the PRT method provided staff with a clear message concerning additional structures and

processes that need to be put in place in order to achieve the faculty’s target of all students being able to demonstrate good writing skills.

3.3 Evaporating Clouds

A number of dilemmas have arisen which have been worked through using the ‘Evaporating Cloud’ process: for example, some argued we should use external assessors to perform assessments of student work, while others favoured use of internal assessors. Conflict was avoided by use of the EC (see figure 4), which allowed us to examine the dilemma, exposing weak assumptions, in order to devise strategies that achieved both sides’ requirements. Other dilemmas have included: Should course delivery encourage a few pieces of assessed work, or should assessment be spread through the course? Should there be a standard policy on tutorials? University-wide issues have also encroached on our activities – the redesign of degree structures in other parts of the university posed some thorny dilemmas that needed careful thought. This provided an example of where an evaporating cloud plus the associated ‘Negative Branch Reservation’ proved helpful as all the proposed ‘solutions’ posed some negative side effects. Reservations concerning such side effects needed to be resolved/ avoided/ mitigated.

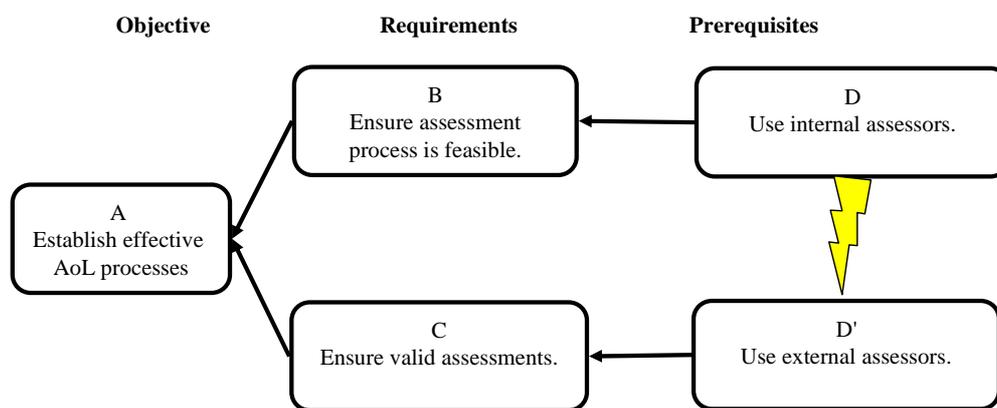


Figure 4: The evaporating cloud for the assessment dilemma

3.4 Layers of Resistance and Steps to Buy-in

As an integral part of managing change, TOC gives considerable emphasis to the notion of harnessing resistance to change (Houle and Burton-Houle 1998; Mabin, Forgeson et al. 2001). Not only is resistance acknowledged, it is actively sought out as a means of tapping into the collective intuition to improve solutions. As can be seen in the PRT for “Achieve accreditation” (Table 1/Fig 2), our AoL team was concerned about the lack of buy-in from staff. Hence the actions that have been taken have sought to involve staff in ways that ‘build support for AoL internally’ to overcome this resistance. One example is the way that assessment has been conducted. We perceived possible sources of resistance from staff in the ‘perceived overhead, and cost’ of AoL, and also the fear of outside assessors. The emergent resistance needed to be reduced, while not compromising the integrity of the AoL process. Use of the EC in Fig. 4 led to an active search for/ design of a streamlined and cost-effective AoL system.

In general, the actions taken so far have sought to “Demonstrate more value out of overheads; inclusivity in T & L; seeing benefits to staff & students”. The “Achieve

Accreditation” PRT also listed as an obstacle, “Accreditation seen as compliance”. Actions decided on to address this were: “Get Mary¹ & others to spread glad tidings about accreditation; get more people to seminars; build enthusiasm; identify early adopters, pick easy tasks, quick wins, appreciate effort; get AoL activities into promotions.”

These have been some of the strategies that have been adopted to date and found effective in building buy-in, though there is obviously much more to do.

4 Reflections

In reflecting on the processes followed, our path has been guided by TOC, and in this paper we have outlined the use of four different approaches drawn from the TOC methodology:

1. The Five Focusing Steps to focus on the most pressing concerns;
2. Prerequisite Tree (PRT) to identify obstacles – and their corresponding stepping stones – to achieving targets;
3. Evaporating Cloud (EC) to clarify and resolve particular dilemmas;
4. Steps to gain Buy-in to overcome resistance in a positive way.

We have also used Negative branches to explore the likely consequences of proposed actions, in order to accentuate the positive and eliminate any negative consequences. Examples of such thinking may be found, for example, in Boyd and Cox (1997); Scheinkopf (1999); Mabin, Davies et al. (2006); Dettmer (2007).

In designing a writing skills improvement programme, a number of these tools have been used in concert. In particular, we used the PRT approach to elicit and collate student views on obstacles and stepping stones, in order to raise the standard of written communication skills of our commerce graduates.

Our aims for teaching and learning in our business school go far beyond meeting the Assurance of Learning requirements for accreditation, with these broader aims including goals for student learning and staff satisfaction. The processes described above have laid a solid framework for achieving these broader aims, providing a coherent, robust and accessible framework for decision-making. A natural next step would be to explore the use of the TOC Strategy and Tactics Tree for teaching and learning.

Reflecting on the role that accreditation has played, it is clear that the pursuit of accreditation has provided the mandate to press ahead with teaching and learning initiatives started by the Pathways to Success project; without accreditation, progress in this area would have been slow. So what real benefits have been achieved?

The key to meeting AACSB’s requirements for assurance of learning is ‘closing the loop’. Results from assessments have revealed reasonable achievement of many learning goals/objectives, though there have been some surprises when aspects of learning were found to be not achieved to the degree expected. The use of learning goals, learning objectives and their corresponding rubrics has been helpful in making the learning objectives of assignments clearer to students, and in providing a clearer

¹ Name changed for confidentiality, Mary represents a bubbly departmental head.

outline of what is required. Tutors and students have responded well. The writing skills programme with accompanying assessment tasks and rubrics has provided a clearer learning focus, and improvements have been discernible. More remains to be done and we are encouraged by some gratifying early successes.

Acknowledgements

The author acknowledges the many people involved and assisting with the initiatives described, throughout the FCA, the wider university (University Teaching and Development Centre, Student Learning Support Services) and other universities.

References

- Biggs, J. 2003. Aligning Teaching and Assessing to Course Objectives. In *Teaching and Learning in Higher Education: New Trends and Innovations*. University of Aveiro. <http://event.ua.pt/iched/main/invcom/p182.pdf>
- Box, G.E.P.1979. Robustness in the strategy of scientific model building, in *Robustness in Statistics*, R.L. Launer and G.N. Wilkinson, Editors.. Academic Press: NY.
- Fink, L. D. 2005. *Integrated course design*. IDEA Paper #42. http://www.idea.ksu.edu/papers/Idea_Paper_42.pdf
- Boyd, L. H. and J. F. I. Cox 1997. "A Cause-And-Effect Approach To Analysing Performance Measures." *Production and Inventory Management Journal* **38**(3 (Third Quarter)): 25 - 32.
- Deming, W. E. 1982. *Out of the Crisis*. Cambridge, MA, MIT Center for Advanced Engineering Study.
- Dettmer, H. W. 1997. *Goldratt's Theory of Constraints: A Systems Approach to Continuous Improvement*. Milwaukee, WI, ASQC Quality Press.
- Dettmer, H. W. 2007. *The Logical Thinking Process: a systems approach to complex problem solving*. Milwaukee, WI ASQ Quality Press.
- Goldratt, E. M. 1994. *It's Not Luck*. Great Barrington, MA, North River Press.
- Goldratt, E. M. and J. Cox 1992. *The Goal: A Process of Ongoing Improvement*. Croton-on-Hudson, North River Press.
- Houle, D. T. and T. Burton-Houle 1998. *Overcoming Resistance to Change the TOC way*. APICS Constraints Management Symposium Proceedings, Seattle, APICS, Falls Church.
- Kim, S., V. J. Mabin, et al. 2008. "The Theory of Constraints Thinking Processes: Retrospect and Prospect." *International Journal of Operations & Production Management* **28**(2): 155-184.
- Mabin, V., J. Davies, et al. 2006. "Using the Theory of Constraints Thinking Processes to Complement Systems Dynamics' Causal Loop Diagrams in Developing Fundamental Solutions." *International Transactions in Operational Research* **13**(1): 33-57.
- Mabin, V. J. and S. J. Balderstone 2000. *The World of the Theory of Constraints: A Review of the International Literature*. Boca Raton, FL, St. Lucie Press.
- Mabin, V. J. and S. J. Balderstone 2003. "The Performance of the Theory of Constraints Methodology: analysis and discussion of successful TOC applications." *International Journal of Operations and Production Management* **23**(6): 568-594.
- Mabin, V. J., S. Forgeson, et al. 2001. "Harnessing resistance: using the theory of constraints to assist change management." *Journal of European Industrial Training* **25**(2/3/4): 168.
- Scheinkopf, L. 1999. *Thinking for a Change: Putting the TOC Thinking Processes to Use*. Boca Raton, FL, St Lucie Press / APICS Series on Constraints Management.
- Senge, P. M. 1990. *The Fifth Discipline*. New York, Doubleday/Currency.
- Watson, K. J., J. H. Blackstone, et al. 2007. "The evolution of a management philosophy: The theory of constraints." *Journal of Operations Management* **25**: 387-402.
- Wiggins, G. and McTighe, J. 1998 *Understanding by Design*, Association for Supervision and Curriculum Development, Alexandria, Va. Expanded 2nd Edn 2005

Standardising spreadsheet LP: Do textbooks make learning LP easier?

Dr Shane Dye

shane.dye@canterbury.ac.nz

Dr Nicola Ward Petty

nicola.petty@canterbury.ac.nz

Department of Management

University of Canterbury

New Zealand

Abstract

In this paper we put forward specifications for a standard for spreadsheet linear programming models when linear programming is taught using spreadsheets to entry-level business students. The primary purpose of the standard should be to make it easy for students to recognise and interpret linear programs formulated as spreadsheet models. The specifications for the standard are supported by the results of educational research, and the experience of teaching linear programming to entry-level business students for over twenty years.

A number of popular textbooks are examined to see how well their standard formats, if any, fit the requirements. While most use a standard format, their standards do not appear to be designed specifically to aid student learning of linear programming.

Key words: Linear programming, Education, Spreadsheet models.

1 Introduction and credentials

The teaching of linear programming has been evolving for the last 50 years. As the discipline of operations research, the theory and practice of teaching, the range of students and the technology have changed so the practice of teaching linear programming has also changed. By and large this evolution has led to improved teaching, courses and textbooks. Consequently we should expect that the current practice of teaching linear programming is satisfactory, and attempts at improvement may not appear radical. The standardisation promoted here is based on educational research, ideas from the literature and reflective teaching experience. There are aspects of it in many textbooks. But no single textbook fully embraces this teaching practice. As such, one of the goals of this paper is to provide suggestions to textbook authors and course designers on how to improve their textbooks and teaching aimed at introducing non-operations research majors to the field.

Though we begin with the premise that the teaching of linear programming using spreadsheets is satisfactory, it is possible that we can do much better. Wild (1994), while commenting on the teaching of statistics, stated “If we have not thought through very carefully what we are trying to achieve, we are in no position to assess the quality of what we provide. Unfortunately, so much of what we do is not thought through from a careful consideration of customers, aims, and objectives; it just grows in an ad hoc

way over the years, building on what has been done before.” We suggest that the same could be said of the teaching of OR, and particularly the teaching of LP using spreadsheets. It may not have been thought through carefully, particularly bearing in mind the changing clientele and the possibilities that the spreadsheet provides.

A recognised approach in the scholarship of teaching involves critical reflection on practice, and analysis of the difficulties students encounter. This can then be measured up against educational and pedagogical theory from related disciplines. Weimar (2006) provides a classification scheme for the Scholarship of Teaching Literature. This paper would be classified as a recommended-practices report, within the category, ‘The Wisdom of Practice’. Powell (2006), in his review of Weimar’s book, endorsed her opinion that journals should apply the standards that it ‘relates to a meaningful aspect of instruction, offers ‘good’ advice [and] communicates constructively’. In addition Weimar (p71) expressed concern that it is not always clear what are ‘the qualifications of the person offering the advice’ or ‘the justification for the advice given.’ We provide therefore, our credentials: The authors have taught introductory level management science for a total of over twenty five years, with increasingly high student ratings. One of the authors has also previously trained and worked as a high school mathematics and computing teacher and was awarded a university teaching award in 2002. In addition, the ideas in this paper are supported by research in cognitive psychology.

Liebman (1998) addressed the teaching of OR, looking at what cognitive psychology could offer, especially in the area of active learning. We also look at the contribution that research in cognitive psychology can make, but focus specifically on the teaching of LP using spreadsheets. The areas of research that we draw on are the constructivist theory of knowledge, the novice/expert comparisons, and transfer of learning.

We believe that the requirements promoted here for a standard structure for spreadsheet linear programs make a difference in the way that students make sense of linear program models. We offer them as a suggestion to improve teaching practice on courses providing an introduction to operations research to non-OR majors.

As textbooks provide a window into the way material is presented, guidance for novice instructors and, often, the foundation for a course, we evaluated how well textbooks captured the essence of our standard. To this end we have examined 13 textbooks that use Excel and Solver as a basis for teaching LP. The examined textbooks include all the better known introductory texts and many others. Details of the textbooks are given in the appendix. This analysis is inspired by Chelst (1998), who suggested improvements to the teaching of decision analysis over that presented in many textbooks, and Cobb (1987) in his seminal paper on Introductory Statistics textbooks.

2 The rationale for teaching LP to non-OR business students

In this paper we are concerned with teaching business students, mainly at undergraduate level. These students generally have only limited mathematics proficiency, but more to the point, they do not think like mathematicians. Boas (1981), then editor of the American Mathematical Monthly, observed even in 1981 that very few students think like mathematicians, nor wish to, and the teaching of mathematics should be undertaken with that in mind. Many of our students have quite poor mathematical skills, and are quite resistant to mathematical terminology. Bell (2005) calls his non-quantitatively

skilled students ‘poets’, in contrast with the engineers, implying that their strengths lie elsewhere.

The statistics education research literature accepts as given that there are challenges in teaching quantitative courses. Ben-Zvi and Garfield (2004) state four main challenges to success in teaching and learning statistics, which must resonate with many OR instructors. These can be paraphrased and adapted as: It can be hard to motivate students to do hard work. Many students have difficulty with the underlying mathematics, and that interferes with learning the related statistics (or stats content). The context can mislead students who rely on experience and intuition, and students expect the focus to be on numbers, computations, formulas and one right answer.

Given this lack of mathematical inclination in our clientele, it is an important exercise to determine why we choose to teach linear programming, and what exactly we wish the students to gain from it. In contrast, it is obvious that LP should form an important part of a course for operations research majors, who we hope would have strong quantitative skills. It is clearer that we wish them to be able to develop an LP model and that they will need the underlying mathematical understanding for analysis and more advanced topics such as integer programming, duality and network models. Most of our students will not become OR majors, but there is still much that business students can gain from the study of linear programming. Bell (2005) expounds the importance of linear programming (which he calls simultaneous decision situations) and suggests that ‘Those who understand the simultaneous framework and can recognize the kinds of problems where a sequential decision approach fails, and who know a better approach, have a competitive advantage in the marketplace.’

Different instructors will have different ideas as to the what and why of linear programming. However we do suggest that some instructors may not have given much time to contemplating exactly why they teach LP, and what it is that the students should get out of it. “In the absence of pedagogical content knowledge, teachers often rely on textbook publishers for decisions about how to best organize subjects for students.” (Bransford (2000) p45) This may be a satisfactory survival tactic for a novice teacher, but for an expert teacher, and for a textbook author, it is befitting to have a clear idea of the purpose and nature of the introduction to LP.

There are many benefits for non-OR-majors, the focus of this paper, in learning about linear programming. We have identified three: the concept of a model, the power of optimisation, and a recognition of the benefit of OR. First, as linear programming is often the introductory topic in an OR course, it can be students’ first taste of a mathematical model, though they may have some familiarity with statistical models. They learn more about the role of the model in decision-making and the tension between the reality and the model. Second, students can be introduced to the power and the limitations of optimisation methods, and gain an appreciation of how this can lead to better decision-making.

Third, Wild (1994) in discussing the marketing of statistics gives the example of a firm that has need for statistically designed experiments at most twice a year. “What they need on site may not be someone who can design experiments, but several people who can recognise those two situations when and where they arise and call in an outside consultant.” (p165). Similarly what many firms need is people on site who can recognise situations where linear programming and/or operations research modelling would aid in decision-making, and possibly lead to money saving and/or competitive advantage, and call in an outside consultant.

These are the benefits or reasons for learning about LP that we have identified. However, this does not define what exactly students should be able to do during the course. We suggest that students should be able to recognise a situation where a linear program will be useful, interpret simple LP models given in Excel, explain simple OR concepts as they apply to simple LP examples, use the Excel Solver to optimise an LP, interpret the output from Excel Solver and identify a solution to implement, determine which assumptions of linear programming are valid and reasonable for a given simple situation, explain the role of sensitivity analysis, apply simple sensitivity analysis to an Excel LP model and interpret the shadow price of a simple constraint.

We believe that these learning objects help to provide the benefits identified above.

3 Standardisation purpose and requirements

As evidenced by the range of spreadsheet formulations of linear programs in operations research textbooks, there is no obvious standard structure for a linear program in a spreadsheet. When students are introduced to linear programming through spreadsheet models, being exposed to different model structures can make understanding linear programs more difficult. When students are also unfamiliar with spreadsheets, this difficulty is compounded. Since the flexibility allowed by spreadsheets can be introduced once students have a stronger grounding in linear programming, we suggest that there is no need to use this flexibility when students are first introduced to linear programming spreadsheets.

It may be stating the obvious but for a course where the main goal in teaching LP using spreadsheets is to teach LP not spreadsheets, the pedagogical practices in the course should be geared toward making it easier to learn LP and not illustrating the power and flexibility of spreadsheets. A standard structure for spreadsheet LP models can support this. However, the standard structure will provide the most value if it has been designed to meet this purpose. To this end, the structure should make it easy for students to recognize the spreadsheet model as a linear program and to find and interpret the LP model in the spreadsheet. Formatting conventions for general spreadsheet models do not meet these requirements.

3.1 Requirements of the standard layout

The standard structure of a spreadsheet LP should differentiate between decision variables, parameter values, left-hand side constraint calculations, and the objective function. These are the important components of a linear program which we wish students to be able to recognise and understand. Before they understand them sufficiently, students can't be expected to find them easily within a spreadsheet model in a previously unseen format. To aid this differentiation, it is good practice to include a 'style key' and to avoid additional formatting, as the additional 'noise' can hide the information provided by the standard.

The differentiation of these components can be provided by using a consistent format and/or location. The most benefit will be provided if the format, location and shape of the regions these components form, remain consistent within the standard. This reduces the unimportant aspects of the spreadsheets which changes, focussing attention on to the important aspects such as the number of variables and constraints, whether the objective function is maximised or minimised, and the problem context. This means being consistent about whether rows or columns are used for constraints, for instance.

Constraints and the objective function are each constructed from a number of elements. By ensuring that all of those elements associated solely with one of these components are visually connected we make it easier for students to make to connection. This means keeping the coefficients and calculations for each constraint and the objective function in the same row or column.

Some types of constraints must be expressed in a non-intuitive way for the standard structure (e.g. product mix constraints). These should be avoided initially and introduced once the students have a better understanding of LP. Similarly, for some linear programs an alternative formulation may make the model easier to understand and analyze, (such as a transportation problem). Eventually the alternative formulation for such models can be introduced and should be contrasted with the standard layout for the same model. This illustrates how the model is still an LP, but that the alternative layout better suits the purposes of the modeller. This parallels the introduction of the transportation simplex tableau in an algorithmic-based course.

Since all parameters and constraints need to be explicitly identified in the spreadsheet, it is important that the values appear clearly on the spreadsheet. The use of values directly in formulas or Solver dialogue boxes should be avoided, and formulas should never be used in Solver.

In addition, students should not have to guess what the standard is. They should be told and shown specifically what it is and directed to use it. Do not provide choice as novices are not well placed to make good judgements.

The idea of standardising spreadsheets is not new. Leong and Cheong (2008) talk about standardising their spreadsheets, but this is for spreadsheets in general, not LP. “We encourage our students to prepare their spreadsheet models in a standardised manner, particularly by separating data and model and color-coding.” As we explain later, most textbooks follow this practice. The standards they use are suggested “best-practice” for general spreadsheet models, but are not specifically designed to aid the learning of LP.

3.2 An example of a standard layout

In our course we use coloured shading and borders to identify the different roles. Light blue represents decision variables, pale yellow for parameter values, pink for constraint (lhs) calculations, and tan with a thick border for the objective function.

Our standard layout uses one row each for: the decision variable labels, the objective function, the decision variable values and each constraint. Each decision variable has a corresponding column. The calculated part of constraints (usually a formula using sumproduct) and the objective function appear in an additional column followed by a column holding labels indicating the type of constraint ($\leq, \geq, =$) then a column for the constraint right-hand side values. All objective function coefficients and constraint coefficients and right-hand-side values appear as parameters directly (not calculated from other data). Figures 1 and 2 show examples. (They are reproduced here in greyscale rather than colour. Colour versions may be accessed from http://www.minandmax.org.nz/teachers/LP_Examples.xls). The layout suggested also helps to avoid non-linear objective functions and constraints. Two examples are given to illustrate how the standard layout accentuates commonalities and differences between different linear program models.

	A	B	C	D	E	F	G	H	I	J	
1	Moana's Desk Linear Programming Model							Parameter			
2	Linear program							Decision variable			
3	Desk type	Basic	Trendy	Modern	Elite	Deluxe		Constraint formula			
4	Quantity	4	4	3	1	2		Objective function			
5	Objective function						Total				
6	Profit	\$35.00	\$50.00	\$40.00	\$50.00	\$70.00	\$650.00				
7	Constraints						Total	Time avail.			
8	Forming	1.5	2	2	4	3	30	≤	30	hours	
9	Detailing	1	3	2	3	5	35	≤	40	hours	
10	Assembly	2	2	2	2	2	28	≤	30	hours	
11	Finishing	1	2	3	4	4	33	≤	40	hours	

Figure 1: A production LP, maximising profit with five variables and four constraints

	A	B	C	D	E	F	G	H	I	J	K	L
1	Food Boxes for Refugees											
2	Decision variables											
3	Food item	Rice	Tomato soup	Peaches	Sardines	Milk powder	Corned beef	Baked beans		Parameter		
4	Quantity	6	0	7	3	7	14	2	items	Decision variable		
5	Item units	packet	tin	tin	tin	sachet	tin	tin		Constraint formula		
6	Objective function											
7	Cost	\$1.95	\$2.15	\$1.68	\$2.10	\$4.50	\$2.20	\$0.99	\$94.04	Total		
8	Constraints											
9	Weight	1.00	0.50	0.49	0.22	0.45	0.40	0.50	19.84	≤	20	kg
10	Energy	14700	1600	1050	1250	6740	2790	1620	188780	≥	226000	kJ
11	Protein	76	11.6	2.1	22.5	154	80	19	2774.2	≥	1200	g
12	Calcium	300	92	12	33	5410	41	101	40629	≥	27000	mg
13	Iron	36	2.8	1.8	1.6	1.5	8.3	6.3	372.7	≥	340	mg
14	Vitamin B1	4.9	0.2	0.04	0.9	2	0.08	0.5	48.5	≥	31	mg
15	Vitamin B2	0.4	0.15	0.12	0.1	8.6	0.6	0.4	72.94	≥	36	mg

Figure 2: A 'diet' LP, minimising cost with seven variables and seven constraints

All LP spreadsheet models in the course are structured in this way, including in the assessment. The students are also instructed explicitly as to the standard layout.

3.3 Benefits of a standard LP structure

Research on the differences between novices and experts in how they process information gives insights into successful learning. (Bransford (2004) p31) A key principle is that “experts notice features and meaningful patterns of information that are not noticed by novices.” This statement has two important corollaries. Firstly, most of the people who teach linear programming (and hopefully all textbook writers) could be classified as experts in the field of LP. Thus for them it is easy to notice features and meaningful patterns in problems; these aspects are probably not apparent to novices and students. Secondly, “research on expertise suggests the importance of providing students with learning experiences that specifically enhance their abilities to recognize meaningful patterns of information.” (Bransford (2004) p36). We suggest that using a standard format for LP when introducing spreadsheets, combined with repetition will enhance the opportunities for students to recognise and internalise the important meaningful patterns.

Having a standard appearance provides a level of abstraction from the specific. It would be possible, using the standard structure, to draw a general LP spreadsheet model using the structure but with no numbers. This can be seen as replacing the standard algebraic representation of an LP. Research has been performed on instruction that helps students to transfer knowledge from one example to another (see Bransford

(2004) p63). It was found that students who were given abstract training as well as specific examples were better able to transfer that knowledge. The use of specific layout and formats (including coloured shading) reinforces the different components of a linear program, and its distinctive nature. We propose that a standard structure will more easily enable students to apply lessons learned to new examples and instances.

Another benefit of using a standard structure for LP spreadsheets is that it provides a model of good practice for the students. Generally we have found that students don't want to know the many different ways that a spreadsheet can be organised and presented. In general they are much happier to be told exactly what to put where. It is unreasonable to expect them to "reinvent the wheel" with regard to good spreadsheet design and then grade them down when their spreadsheet does not seem logical to us. If we teach one sound, robust method, with reduced opportunity for choice, we are helping the students to avoid unnecessary work that detracts from the learning. The standard structure that we use is readily scalable to any size model, and valid for any LP.

4 Analysis of introductory textbooks

We examined the linear programming sections in a range of textbooks to see whether they use a standard structure for spreadsheet LP models and whether the standard used meets underlying pedagogical requirements.

The textbooks we examined had to be useful for an introductory course on operations research which used spreadsheets as the main medium for linear program (and other) models. We did not include any textbooks which did not use spreadsheet linear programs at all or provided a spreadsheet linear program only as an example of a method to optimise a linear program. We only included the latest edition of textbooks published during or before 2008. Details of the textbooks examined, and those excluded, are given in the appendix.

The textbooks provide a range of practices. Most of the textbooks (11 of 13) provided rules or guidelines regarding the structure and layout of spreadsheet models and LP models in particular. Of these seven provided this as guidelines while the other four provided rules, although none were definitive about following the rules exactly.

Two of the books made a clear distinction between spreadsheet style to be used once the student was familiar with spreadsheet models and the structure used to teach and learn linear programming spreadsheet models. The most explicit of these was Balakrishnan, Render and Stair (2007): "[This] consistent approach is more suited to the beginning student of LP. As you gain experience with spreadsheet modeling of LP problems, we encourage you to try alternate layouts."

The textbook by Powell and Baker (2007) differs from the others. The first topics cover spreadsheet modelling, in general, with an extensive chapter on spreadsheet engineering. Linear programming is a much later topic, following non-linear programming. The standard used for LP examples is explained in terms of the structure of the algebraic model.

All but one of the texts followed best practice for general spreadsheet models. Only two textbooks (Balakrishnan, Render and Stair (2007), and Hillier and Hillier (2008)) used an implied standard which clearly highlighted the structure of linear programs to aid student learning, as advocated above. The rest failed to meet the advocated pedagogical requirements in one or more of the following ways. Seven included additional, unnecessary calculations within the spreadsheet model. Seven did not sufficiently differentiate the parameter cells from cells containing calculations. Six did

not use a consistently shaped area for constraints. Five did not include all information relevant to constraints in a visually connected area. Three did not use a consistent location for various components of the LP.

All of the textbooks examined seem to rely on algebraic models to help students better understand linear programs. We believe that one reason students do better when an algebraic model is provided is because all of the textbooks use a clear standard layout for their algebraic models. It is a standard that serves to provide the underlying pedagogical requirements advocated above. We suggest that removing the algebra removes a barrier for many students, but that the use of algebra should be replaced with a specified standard format designed to help students to generalise from examples.

5 Conclusion

The standard structure proposed for spreadsheet LP aims to make it easy for students to recognise and interpret linear programs formulated as spreadsheet models. The requirements for the standard are supported by the results of educational research, and the experience of teaching linear programming to entry-level business students for over twenty years.

Thirteen popular textbooks were examined and their approach to the teaching of LP analysed to see how well their standard formats fit the requirements. While most use a standard format, their standards do not appear to be designed specifically to aid student learning of linear programming.

6 References

- Bell, Peter C. (2005) "Operations Research for everyone (including poets)" *OR/MS Today* August 2005: pp22-27
- Ben-Zvi, D., & Garfield, J. (2004). "Statistical literacy, reasoning, and thinking: goals, definitions, and challenges" in D. Ben-Zvi & J.Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* pp3-15. Kluwer Academic Publishers (Springer)
- Boas, R. P. (1981) "Can We Make Mathematics Intelligible?" *The American Mathematical Monthly* vol 88 #10: pp727-731
- Bransford, J. D. and A. L. Brown, Eds. (2000). *How people learn: brain, mind, experience and school*. Washington D.C., National Academy of Sciences
- Chelst, Kenneth (1998) "Can't see the forest because of the decision trees: a critique of decision analysis in survey texts." *Interfaces* 1998 vol 28 #2: pp80-98
- Cobb, G. W. (1987) "Introductory textbooks – a framework for evaluation." *Journal of the American Statistical Association* 1987 vol 82: pp321-339
- Leong, T.-Y. and M. L. F. Cheong. (2008). "Teaching Business Modeling Using Spreadsheets" *Informations Transactions on Education*. 9: pp20-34
- Liebman, J. (1998). "Teaching operations research: lessons from cognitive psychology." *Interfaces* 28(2): pp104-110
- Powell (2006) "Review of Maryellen Weimer: Enhancing Scholarly Work on Teaching and Learning (Jossey-Bass, 2006)" *Informations Transactions on Education* 2006 vol 6 #3 <http://ite.pubs.informs.org/Vol6No3/Powell/>

Weimar, Maryellen (2006) *Enhancing scholarly work on teaching and learning: professional literature that makes a difference*. Jossey Bass, 2006.

Wild, C. J. (1994) “Embracing the ‘wider view’ of statistics.” *The American Statistician* 1994 vol 48 #2: pp165-171.

Appendix

Table 1 summarises the results by textbook. The ‘Rules’ column indicates how the textbook communicates any guidelines or rules regarding structure of LP spreadsheet models. The ‘Style’ column indicates where (if at all) the style failed to meet the underlying pedagogical requirements advocated in this paper. For this information, the first three spreadsheet examples in the running text were used to triangulate an implied standard. Those aspects common to the first three examples were deemed to form the standard, regardless of the guidelines given in the textbook. In all textbooks this implied standard was the same, or stricter than the rules or guidelines given in the text.

Table 1: Standardisation of LP spreadsheet models in textbooks

Name	Authors	Edition	Rules	Style
Data Analysis & Decision Making	Albright, Winston, Zappe	3 (2006)	G	c x
Spreadsheet Modelling and Applications: Essentials of Practical Management Science	Albright, Winston	1 (2005)	G	c x
Optimization Modeling with Spreadsheets	Baker	1 (2006)	R	p
Managerial Decision Modeling with Spreadsheets	Balakrishnan, Render, Stair	2 (2007)	R	
Introduction to Management Science - A Modeling and Case Studies Approach with Spreadsheets	Hillier, Hillier	3 (2008)	G	
Applied Management Science - Modeling, Spreadsheet Analysis, and Communication for Decision Making	Lawrence, Pasternack	2 (2002)	G	c p h
Quantitative Business Modeling	Meredith, Shafer, Turban	1 (2002)	G	c p x
Decision modeling with Microsoft Excel	Moore, Weatherford, Eppen, Gould, Schmidt	6 (2001)	G	l p
Management Science - The Art of Modeling with Spreadsheets	Powell, Baker	2 (2007)	G	p x
Spreadsheet Modeling & Decision Analysis - A Practical Introduction to Management Science	Ragsdale	5 (2007)	R	l x
Introduction to Management Science with spreadsheets	Stevenson, Ozgur	1 (2007)	N	p x
Introduction to Management Science	Taylor	9 (2007)	N	l p
Practical Management Science	Winston, Albright	3 (2007)	G	c x

The following codes are used. For the Rules column: N – no rules or guidelines, G – guidelines for structure only, R – clear, but non-definitive rules. For the Style column: c – information for constraints not visually connected, h – elements hidden in Solver dialogue boxes, l – layout of the LP’s are not consistent, p – parameters and calculations are not visually discriminated, x – extra calculations are included with the LP model.

Table 2 lists those textbooks initially considered for the study but excluded. The reason for exclusion of each is given. Mostly these textbooks had two or fewer spreadsheet linear programming models provided in the main text. These are marked as too little or no Excel. Barlow (2005) *Excel Models for Business and Operations Management* was excluded because it did not include a chapter devoted to spreadsheet LP models.

Table 2: Textbooks excluded from study

Name	Author	Edition	Reason
An Introduction to Management Science - Quantitative Approaches to Decision Making	Anderson, Sweeney, Williams	12 (2008)	Too little Excel
Excel Models for Business and Operations Management	Barlow	2 (2005)	No LP chapter
Management Science - Decision Making through Systems Thinking	Daellenbach, McNickle	1 (2005)	Too little Excel
Statistics Data Analysis, & Decision Modeling	Evans	3 (2007)	Too little Excel
Introduction to Operations Research	Hillier, Lieberman	8 (2005)	Too little Excel
Spreadsheet Modeling for Business Decisions	Kros	1 (2008)	Too little Excel
Decision Technology - Modeling, Software, and Applications	Liberatore, Nydick	1 (2003)	No Excel
Essential Quantitative Methods for Business, Management and Finance	Oakshott	3 (2006)	No Excel
Optimization in Operations Research	Rardin	1 (1998)	No Excel
Operations Research - An Introduction	Taha	8 (2007)	Too little Excel
A practical introduction to management science	Waters	2 (1998)	Too little Excel
Quantitive methods for business	Waters	4 (2008)	Too little Excel
Operations Research - Applications and Algorithms	Winston	4 (2004)	Too little Excel
Quantitative Methods for Decision Makers	Wisniewski	4 (2006)	Too little Excel

Note about book variants

Two of the major textbooks Hillier and Hillier (2008) *Introduction to Operations Research* and Winston and Albright (2007) *Practical Management Science* have variants aimed at different courses and markets. We considered all of the different variants for the study. Hillier and Liberman (2005) *Introduction to Operations Research* and Winston (2004) *Operations Research: Applications and Algorithms* do not include sufficiently many spreadsheet LP models and were excluded from the study. The other variants of Winston and Albright (2007) *Practical Management Science* were included. These books contain almost identical introductory LP chapters.