Proceedings of the 2007 Mathematics and Statistics in Industry Study Group

Editors: Tim Marchant, Maureen Edwards and Geoff Mercer
Proceedings of the
2007 Mathematics and Statistics in Industry Study Group, MISG2007

Director: A/Professor Tim Marchant
Associate Director: Dr Maureen Edwards
School of Mathematics and Statistics
University of Wollongong

Editors: A/Professor Tim Marchant, Dr Maureen Edwards and Dr Geoff Mercer
Administrator and Assistant Editor: Ms Joell Hall

First published in Australia January 2008
©MISG2007 (Mathematics and Statistics in Industry Study Group, School of Mathematics and Applied Statistics, University of Wollongong)

This book is copyright. Apart from any fair dealings for the purpose of private study, research, criticism or review, as permitted under the Copyright Act, no part may be reproduced by any process without written permission. Enquiries should be made to the publisher.

All manuscripts for the Proceedings of the MISG were written by the moderators in consultation with company representatives. Manuscripts submitted to the Editors, A/Professor Tim Marchant, Dr Maureen Edwards and Dr Geoff Mercer, were subsequently reviewed by expert referees. On the advice of the referees all manuscripts were accepted for publication subject to the recommended revisions, and then formally approved by the editorial committee.

Preface

MISG2007 Opening Ceremony: From left, Dr. Maureen Edwards, MISG Assoc. Director, Prof. Robert McKibben, Invited Speaker, A/Prof. Tim Marchant, Director and Mr. Joe Maisano, Trading Technology Australia

The Mathematics and Statistics in Industry Study Group (MISG2007), was hosted by the University of Wollongong, during the week 5-9th February 2007 at the University of Wollongong. The MISG is a special interest meeting of ANZIAM, the Australian and New Zealand Society for Applied and Industrial Mathematics. The event was directed A/Prof Tim Marchant and Dr Maureen Edwards, from the University of Wollongong. Administrative support was ably supplied by Ms. Joell Hall and Ms. Sue Denny.

Six industry projects were presented; five of these came from Australia and one from New Zealand. Two strong project themes, this year, were financial mathematics and electricity supply and generation. About 110 people attended MISG2007 including 20 postgraduate students.

MISG2007 was opened by Mr. Stephen Lowe, General Manager Trading, Integral Energy and Prof. Margaret Sheil, DVC-Research, University of Wollongong. The invited speaker was Prof. Robert McKibben, from Massey University. Dr. Mike Camden and Dr. Jeff Dewynne gave talks at the student sessions.

The Australian Mathematics Sciences Institute (AMSI) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) both supported postgraduate attendance at MISG2007. Thanks go to these organisations, and to the University of Wollongong, for their
financial support.

Due to the broad range of skills required to tackle modern industrial mathematics problems many high-profile scientists from the Australian and NZ Statistics and Financial Mathematics communities attended MISG2007 as delegates or moderators. If the MISG meeting is to remain relevant and important in the coming years then this multi-disciplinary approach to industrial problem solving needs to continue, with participation at MISG from all the Mathematical Sciences.

Our thanks go to all of these people and organisations, and of course, to the project moderators. The moderators take responsibility for the industry projects and put in an inordinate amount of time and effort. These contributions are critical to the success of MISG.

Tim Marchant, Maureen Edwards, Directors MISG2007
Contents

Preface 3
Sponsors 6
Projects 7
Overheard in passing ... 9
Delegates 10

Bluescope Steel, Strip track-off and buckling between transport rollers 13
  W.B. Fraser, C. Macaskill, M. McGuinness and A. Thornton

DSTO, The maritime surveillance problem 32
  P. Kilby, P. Tobin, R. Luscombe, S. Barry and R. Hickson

Integral Energy, Calibrating mean-reverting jump diffusion models 57
  J. Alcock, J. Goard and T. Vassallo

Integral Energy, A jump diffusion model for spot electricity prices 82
  R. Bhar

Integral Energy, Calculation of a risk measure for the net system load profile 91
  K. Russell and P. Cerone

Trading Technology Australia, Determining the independence of various measures of financial risk 102
  E. Maberly and A. Wilkins

Transpower/EECA NZ, Maximizing the contribution of wind power in an electric power grid 114
  G. Pritchard, W. Sweatman, K. Nan, M. Camden, B. Whiten

Editors note

1The papers “Calibrating mean-reverting jump diffusion models” and “A jump diffusion model for spot electricity prices” result from the same MISG2007 Industry project and represent different modelling perspectives on the one project.
MISG2007 SPONSORS

The organisers of MISG2007 gratefully acknowledge all contributions that were made by the sponsors:

Host of MISG2007
University of Wollongong

Financial Sponsors
Australian Mathematical Sciences Institute
CSIRO

Project Sponsors
Bluescope Steel
Australian Government Department of Defense - Defense Science and Technology Organisation
Energy Efficiency and Conservation Authority NZ
Integral Energy
Trading Technology Australia
Transpower

Educational Sponsors
Ceanet (Maple 11)
Thomson
Projects

Bluescope Steel
Strip track-off and buckling between transport rolls

Industry Contact: Dr Andrew Dixon Bluscope Steel Research
Dr Daniel Yuen Bluscope Steel Research

Moderators: Dr Barrie Fraser University of Sydney
Dr Charlie Macaskill University of Sydney
Dr Mark McGuinness Victoria University of Wellington

Student Moderator: Mr Aaron Thornton University of Wollongong

Defence Science and Technology Organisation
Application of the non-stationary traveling salesman problem to maritime surveillance

Industry Contact: Dr David Marlow DSTO

Moderators: Dr Philip Kilby National ICT Australia
Dr Patrick Tobin Swinburne University

Student Moderator: Ms Ruth Luscombe University of Melbourne

Integral Energy
Calibrating the mean reverting jump diffusion model to Australian spot electricity prices

Industry Contact: Dr Andrew Ziogas Integral Energy

Moderators: A/Prof Ram Bhar UNSW
Dr Jamie Alcock University of QLD

Student Moderator: Mr Tony Vassallo University of Sydney

Integral Energy
Calculation of risk multiplier for net system load profile

Industry Contact: Mr Alan Prictor Integral Energy

Moderators: A/ Prof Ken Russell University of Wollongong
A/Prof Pietro Cerone Victoria University

Student Moderator: Ms Vivien Challis University Queensland
Trading Technology Australia

Determining the independence of various measures of financial risk

Industry Contact: Dr Alex Radchik Trading Technology Australia

Moderators: Dr Andy Wilkins University Queensland
            Prof Ed Maberly University of Canterbury/Monash University

Transpower NZ and the Energy Efficiency and Conservation Authority NZ
Operating and planning and electricity transmission grid to maximize the contribution of wind

Industry Contact: Mr Conrad Edwards Transpower/EECA NZ

Moderators: Dr Geoff Pritchard University of Auckland
            Dr Winston Sweatman Massey University

Student Moderator: Mr Kim Nan University of Auckland
Overheard in Passing...

Ken Russell - *I once wrote a limerick that ends in heterscedasticity.*

Ruth Luscombe - *The problem is soluble, just add water and it has a solution.*

Tony Gibb - *The power function is initially cubic then constant before it drops suddenly to zero - I am using a linear approximation.*

Andrew Dixon - *When I say thin, I mean not thick.*

Ken Russell - *That sounds so logical; we must have done something wrong.*

David Marlow - *So it’s not just a question of a buckle being a fulfilled wrinkle.*

Patrick Tobin - *It’s somewhere between the waylines; let’s call it the half-way line.*

Ruth - *When do you not go backwards? Judy - When there are no ships behind you.*

**And the Grand Winners of MISG2007 overheard in passing...**

Phil Kilby - *The undetected ships are displayed in Cyan, which is apparently a bad colour to use.*

Ken Russell - *That’s why they’re undetected*
Delegates

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>First Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcock</td>
<td>Dr</td>
<td>Jamie</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>Alzoubi</td>
<td>Mr</td>
<td>Loai</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Andersen</td>
<td>Dr</td>
<td>Bob</td>
<td>CSIRO</td>
</tr>
<tr>
<td>Arundell</td>
<td>Dr</td>
<td>Justin</td>
<td>Integral Energy</td>
</tr>
<tr>
<td>Bachtiar</td>
<td>Mr</td>
<td>Alhaj Akbar</td>
<td>Sydney University</td>
</tr>
<tr>
<td>Baloi</td>
<td>Ms</td>
<td>Luminita</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>Baowan</td>
<td>Ms</td>
<td>Duangkamon(Yui)</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Barry</td>
<td>Dr</td>
<td>Steven</td>
<td>UNSW at ADFA</td>
</tr>
<tr>
<td>Bassett</td>
<td>Dr</td>
<td>Mark</td>
<td>UCME</td>
</tr>
<tr>
<td>Beunen</td>
<td>Mr</td>
<td>Jos</td>
<td>Australian Bureau of Statistics</td>
</tr>
<tr>
<td>Bhar</td>
<td>Dr</td>
<td>Ramaprasad</td>
<td>UNSW</td>
</tr>
<tr>
<td>Broadbridge</td>
<td>Prof</td>
<td>Phil</td>
<td>AMSI</td>
</tr>
<tr>
<td>Camden</td>
<td>Mr</td>
<td>Mike</td>
<td>Statistics New Zealand</td>
</tr>
<tr>
<td>Cerone</td>
<td>A/Prof</td>
<td>Pietro</td>
<td>Victoria University</td>
</tr>
<tr>
<td>Challis</td>
<td>Ms</td>
<td>Vivien</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>Chambers</td>
<td>Prof</td>
<td>Ray</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Chen</td>
<td>Mr</td>
<td>Weiming</td>
<td>Massey University</td>
</tr>
<tr>
<td>Clark</td>
<td>Dr</td>
<td>Robert</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Cogill</td>
<td>Dr</td>
<td>John</td>
<td>Retired</td>
</tr>
<tr>
<td>Cox</td>
<td>Dr</td>
<td>Grant</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Cozijnisen</td>
<td>Dr</td>
<td>Maria</td>
<td>BlueScope Steel</td>
</tr>
<tr>
<td>Crook</td>
<td>Mr</td>
<td>Jonathan</td>
<td>Victoria University of Wellington</td>
</tr>
<tr>
<td>Dalton</td>
<td>Mr</td>
<td>Barry</td>
<td>Integral Energy</td>
</tr>
<tr>
<td>Davis</td>
<td>Mr</td>
<td>Walter</td>
<td>Statistics New Zealand</td>
</tr>
<tr>
<td>Denny</td>
<td>Ms</td>
<td>Sue</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Dewyne</td>
<td>Dr</td>
<td>Jeff</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Dixon</td>
<td>Dr</td>
<td>Andrew</td>
<td>BlueScope Steel Research</td>
</tr>
<tr>
<td>Duan</td>
<td>Mr</td>
<td>Wen</td>
<td>University of Auckland</td>
</tr>
<tr>
<td>Duke</td>
<td>Mr</td>
<td>Timothy</td>
<td>Statistics New Zealand</td>
</tr>
<tr>
<td>Edwards</td>
<td>Dr</td>
<td>Maureen</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Edwards</td>
<td>Mr</td>
<td>Conrad</td>
<td>Transpower NZ</td>
</tr>
<tr>
<td>Fraser</td>
<td>Dr</td>
<td>W Barrie</td>
<td>University of Sydney</td>
</tr>
<tr>
<td>Ganjehlou</td>
<td>Mr</td>
<td>Asef Nazari</td>
<td>ITMS, Ballarat University</td>
</tr>
<tr>
<td>Gibb</td>
<td>Mr</td>
<td>Tony</td>
<td>SYDAC</td>
</tr>
<tr>
<td>Goard</td>
<td>Dr</td>
<td>Joanna</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Graybill</td>
<td>Mr</td>
<td>Scott</td>
<td>University of Canterbury</td>
</tr>
<tr>
<td>Griffiths</td>
<td>Professor</td>
<td>David</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Gyoung</td>
<td>Ms</td>
<td>Kim Eun</td>
<td>KAIST</td>
</tr>
<tr>
<td>Ha</td>
<td>Mr</td>
<td>Seog One</td>
<td>KAIST</td>
</tr>
<tr>
<td>Hajek</td>
<td>Dr</td>
<td>Bronwyn</td>
<td>University of Adelaide</td>
</tr>
<tr>
<td>Hall</td>
<td>Ms</td>
<td>Joell</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Harun</td>
<td>Ms</td>
<td>Fatimah</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Hoseini</td>
<td>Mr</td>
<td>Sayed</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
<td>Affiliation</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Jaekyu,</td>
<td>Mr Lee</td>
<td>KAIST</td>
<td></td>
</tr>
<tr>
<td>Jeon,</td>
<td>Ms Da Eun</td>
<td>KAIST</td>
<td></td>
</tr>
<tr>
<td>Kang,</td>
<td>Mr Chulmin</td>
<td>KAIST</td>
<td></td>
</tr>
<tr>
<td>Kilby,</td>
<td>Dr Philip</td>
<td>National ICT Australia (NICTA)</td>
<td></td>
</tr>
<tr>
<td>Kordzakhia,</td>
<td>Dr Nino</td>
<td>Macquarie University</td>
<td></td>
</tr>
<tr>
<td>Kwon,</td>
<td>A/Prof Kil H.</td>
<td>KAIST</td>
<td></td>
</tr>
<tr>
<td>Lee,</td>
<td>A/Prof Chang-Ock</td>
<td>KAIST</td>
<td></td>
</tr>
<tr>
<td>Lin,</td>
<td>A/Prof Yan-Xia</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Looker,</td>
<td>Dr Jason</td>
<td>DSTO</td>
<td></td>
</tr>
<tr>
<td>Lucombe,</td>
<td>Ms Ruth</td>
<td>University of Melbourne</td>
<td></td>
</tr>
<tr>
<td>Maberly,</td>
<td>Prof Edwin D.</td>
<td>Monash University</td>
<td></td>
</tr>
<tr>
<td>Macaskill,</td>
<td>A/Prof Charlie</td>
<td>University of Sydney</td>
<td></td>
</tr>
<tr>
<td>Maisano,</td>
<td>Mr Joe</td>
<td>Trading Technology Australia</td>
<td></td>
</tr>
<tr>
<td>Marchant,</td>
<td>A/Prof Tim</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Marion,</td>
<td>Ms Kaye E.</td>
<td>RMIT University</td>
<td></td>
</tr>
<tr>
<td>Marlow,</td>
<td>Dr David</td>
<td>DSTO</td>
<td></td>
</tr>
<tr>
<td>McCrae,</td>
<td>A/Prof Michael</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>McCue,</td>
<td>Dr Scott</td>
<td>Griffith University</td>
<td></td>
</tr>
<tr>
<td>McGuinness,</td>
<td>Dr Mark</td>
<td>Victoria University of Wellington</td>
<td></td>
</tr>
<tr>
<td>McKibbin,</td>
<td>Prof Robert</td>
<td>Massey University</td>
<td></td>
</tr>
<tr>
<td>Mercer,</td>
<td>Dr Geoff</td>
<td>UNSW at ADFA</td>
<td></td>
</tr>
<tr>
<td>Molefe,</td>
<td>Mr Wilford</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Nan,</td>
<td>Mr Ke</td>
<td>University of Auckland</td>
<td></td>
</tr>
<tr>
<td>Neame,</td>
<td>Dr Philip</td>
<td>University of Technology Sydney</td>
<td></td>
</tr>
<tr>
<td>Nelson,</td>
<td>Dr Mark</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Nilsson,</td>
<td>Dr Rod</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Ponsaing,</td>
<td>Ms Anita</td>
<td>University of Queensland</td>
<td></td>
</tr>
<tr>
<td>Prictor,</td>
<td>Mr Alan</td>
<td>Integral Energy</td>
<td></td>
</tr>
<tr>
<td>Pritchard,</td>
<td>Dr Geoffrey</td>
<td>University of Auckland</td>
<td></td>
</tr>
<tr>
<td>Radchik,</td>
<td>Dr Alex</td>
<td>Trading Technology Australia (TTA)</td>
<td></td>
</tr>
<tr>
<td>Russell,</td>
<td>A/Prof Ken</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Shand,</td>
<td>Mrs Judith</td>
<td>Open Training Education Network of TAFE</td>
<td></td>
</tr>
<tr>
<td>Sidhu,</td>
<td>Dr Harvinder</td>
<td>UNSW at ADFA</td>
<td></td>
</tr>
<tr>
<td>Sozio,</td>
<td>Mr Gerry</td>
<td>St Mary Star of the Sea College</td>
<td></td>
</tr>
<tr>
<td>Steel,</td>
<td>Prof David</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Sweatman,</td>
<td>Dr Winston</td>
<td>Massey University</td>
<td></td>
</tr>
<tr>
<td>Szoldra,</td>
<td>Mr Edward</td>
<td>Australian Bureau of Statistics</td>
<td></td>
</tr>
<tr>
<td>Taylor,</td>
<td>Dr Steve</td>
<td>University of Auckland</td>
<td></td>
</tr>
<tr>
<td>Thamwattana,</td>
<td>Dr Ngamta (Natalie)</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Thompson,</td>
<td>Mrs Gwenda</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Thornton,</td>
<td>Mr Aaron</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Tobin,</td>
<td>Dr Patrick Swinburne</td>
<td>University of Technology</td>
<td></td>
</tr>
<tr>
<td>Tritscher,</td>
<td>Dr Peter</td>
<td>University of Wollongong</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Title</td>
<td>First Name</td>
<td>Institution</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Vassallo</td>
<td>Mr</td>
<td>Anthony</td>
<td>University of Sydney</td>
</tr>
<tr>
<td>Wake</td>
<td>Prof</td>
<td>Graeme</td>
<td>Centre for Mathematics in Industry, Massey University</td>
</tr>
<tr>
<td>Wand</td>
<td>Prof</td>
<td>Matt</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Weber</td>
<td>A/Prof</td>
<td>Rodney</td>
<td>UNSW at ADFA</td>
</tr>
<tr>
<td>Wheway</td>
<td>Dr</td>
<td>Virginia</td>
<td>Boeing Australia Ltd</td>
</tr>
<tr>
<td>White</td>
<td>Mr</td>
<td>Richard</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Whiten</td>
<td>Prof</td>
<td>Bill</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>Wilkins</td>
<td>Dr</td>
<td>Andy</td>
<td>University of Queensland</td>
</tr>
<tr>
<td>Williams</td>
<td>A/Prof</td>
<td>Graham</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Wu</td>
<td>Mr</td>
<td>Bisheng</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>You</td>
<td>Ms</td>
<td>Jing</td>
<td>Statistics New Zealand</td>
</tr>
<tr>
<td>Yuen</td>
<td>Dr</td>
<td>Daniel</td>
<td>BlueScope Steel Research</td>
</tr>
<tr>
<td>Zhang</td>
<td>Mr</td>
<td>Jin</td>
<td>University of Wollongong</td>
</tr>
<tr>
<td>Zhou</td>
<td>Dr</td>
<td>Xian</td>
<td>The Hong Kong Polytechnic University</td>
</tr>
<tr>
<td>Zhu</td>
<td>Mr</td>
<td>Justin</td>
<td>DSTO</td>
</tr>
<tr>
<td>Ziogas</td>
<td>Dr</td>
<td>Andrew</td>
<td>Integral Energy</td>
</tr>
</tbody>
</table>
Strip track-off and buckling between transport rollers

W. Barrie Fraser and Charlie Macaskill
University of Sydney

Mark McGuinness
Victoria University of Wellington

Aaron Thornton
University of Wollongong

1 Introduction

The problem brought to MISG 2007 by BlueScope Steel is the formation of ‘ironed-in’ wrinkles in thin steel-sheet products. This is found to occur in a number of environments, but typically is observed in installations where steel sheet is passing over a large number of rollers, some of which drive the sheet and hence provide a tension in the direction of motion. In general, the defects arise sporadically, and only very infrequently. However, they can have a significant cost both because of the loss of product and the difficulty associated with tracking down the cause of the problem. The cause can sometimes be traced to roller misalignment or to metal sheet of non-uniform thickness, for example.

Several different kinds of wrinkles are observed. One type of permanent wrinkle is similar to the folds that are sometimes observed in newspaper printing, which follow a diagonal pattern arising from the edge of a (usually misaligned) roller. The nature of these wrinkles, and the mechanisms leading to their formation, are relatively well-understood in the literature (see e.g. [7] and references therein). This work concentrates on a second kind of wrinkle that occurs more often in practice in cold steel-sheet rolling and that arises in a fundamentally different way. This second kind of wrinkle usually appears as a single thin ridge close to the centre of the sheet and is more prevalent when convex rollers are used to alleviate mistracking (convex rollers have greater radius near the centreline). The wrinkle may have a longitudinal extent of tens of metres but a width of the order of twenty to thirty millimetres and a depth of less than a millimetre. The ridge typically wanders in the lateral direction around the centre of the sheet, as illustrated in Figure 1.

Simple demonstrations with aluminium kitchen foil, as in Figure 2, show that a thin metal sheet under longitudinal tension and with clamped ends will certainly support small amplitude elastic wrinkles, elongated in the longitudinal direction, giving an appearance like corrugated roofing steel. However, these wrinkles are not permanent and disappear when the applied tension is reduced. Moreover, such wrinkles, discussed in the literature in a number of contexts, [2, 6, 13, 14, 15], clearly require a locally compressive field.

This group was asked to model the formation of longitudinal wrinkles in steel moving under tension between rollers, and in particular:
Figure 1: Example of a longitudinal wrinkle in flat steel, and a cross section illustrating typical changes in thickness across the wrinkle.

1. to understand the mechanisms giving rise to local cross-plate compression, which is a necessary prerequisite for initial wrinkle formation;

2. to develop a theoretical model describing the vertical deflection due to this compression;

3. to determine the mechanisms for the iron-on process as a wrinkle approaches a roller.

We use von Kármán’s weakly nonlinear equations for the moderately large deflection of thin, initially flat, elastic plates to address the problem of how wrinkles are initiated under sufficiently high longitudinal tension applied to the ends of a plate. If, in addition, the assumption is made that the vertical deflections are very small, then the equations decouple and the in-plane stresses in the plate obey a (linear) biharmonic equation, so that the stresses can be determined without reference to any knowledge of the vertical deflections. The remaining problem is then to determine the nature of the boundary conditions. The lateral boundaries of the plate are clearly stress-free. The boundary conditions at the ends (at the rollers) are more problematic. Here we make the simplifying approximation that the plate is clamped against movement in both the longitudinal and lateral directions. The underlying assumption is that the friction in the along-roller direction, due to the pressure imposed by the roller nip, is sufficient to prevent any lateral motion there.

Under these conditions, numerical solutions of the biharmonic equation for the stresses reveal a local cross-plate compressive stress quite close to each of the rollers. Furthermore, if a roller is convex, giving rise to a maximum in both the longitudinal and cross-plate tensions at the centreline at the roller location, then there is a corresponding maximum in cross-plate compression at the centreline some distance from the roller. Thus proper consideration of the boundary conditions provides an explanation for point 1 above.

The above findings explain how a maximum elastic deformation will be found near to the roller location and at the centreline of the plate. However, in order to see a permanent distortion of the plate, the elastic stresses must exceed the plastic yield limit, hence ‘setting in’ the distortion. For this to happen the local radius of curvature of the distortion must be quite small: for thin steel sheet we estimate that the vertical deflection divided by the lateral wavelength must be greater than about 1/20. In order to determine this ratio, the full
nonlinear von Kármán equations must be solved iteratively in order to determine the vertical
deflection over the plate, in particular at the location of maximum compression.

A further question to be addressed is that although initiation of the ridge-like distortion
is apparently random (e.g. a roller is slightly out of alignment), once a distortion is initiated
it tends to continue over long lengths of material. One possible explanation for this is that a
local ridge in the plate when reaching the roller will lead to a modified tension at the roller
that in turn influences the compression upstream: this proposed effect can only be explored
with careful solution of the full governing equations with small perturbations in the roller
boundary conditions. This exploration is beyond the scope of the current work.

In sections 2, 3 and 4 we solve von Kármán’s equations for moderately large deflections of
thin elastic plates. Dynamic effects due to the motion of the plate and elastic (bending and
in-plane) waves, and transverse waves due to the tension will all be ignored. This assumption
is discussed more fully in section 6. In section 5 the use of energy methods (see [2]) is discussed
and compared with our results. Finally, in section 7, we describe a mechanism explaining
how these wrinkles may be ironed in due to plastic deformation when a wrinkled section of
the plate moves over an intermediate, single, freely rotating roller.

2 Mathematical formulation

It is observed that when long rectangular plates are subject to large tension in the length-wise
direction that lateral wrinkling of the plate occurs with the wrinkles aligned in the length-wise
direction. Since the velocities of in-plane and transverse elastic waves are large compared with
the transport velocity, we ignore the dynamic effects here.

We model this problem using the weakly nonlinear von Kármán equations. Let the plate
have uniform thickness $h$ and width $2d$, and let the distance between the rollers be $2\ell$. We take
Cartesian axes \((x, y, z)\) so that \(x, y\) lie in the plane of the middle surface of the undeformed plate with \(x\) coincident with the central axis of the strip, and \(z\) in the direction perpendicular to this plane, as illustrated in Figure 3. Let the displacement components of the deformed middle surface in the \(x, y, z\) directions respectively be given by \(u, v, w\). We use the equations

\[
\nabla^4 \phi = -\frac{Eh}{2} L(w, w), \quad D\nabla^4 w = L(\phi, w),
\]

where \(E\) is Young’s modulus, \(\nu\) is Poisson’s ratio and the plate bending stiffness is given by \(D = Eh^3/12(1 - \nu^2)\).

The operators \(\nabla\) and \(L\) are defined as follows:

\[
\nabla^4 \equiv \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2, \quad L(A, B) = A_{xx}B_{yy} - 2A_{xy}B_{xy} + A_{yy}B_{xx},
\]

and the in-plane stress resultants are given in terms of the stress function \(\phi\) by

\[
N_{xx} = \frac{\partial^2 \phi}{\partial y^2} = \phi_{yy}, \quad N_{yy} = \frac{\partial^2 \phi}{\partial x^2} = \phi_{xx}, \quad N_{xy} = -\frac{\partial^2 \phi}{\partial x \partial y} = -\phi_{xy},
\]

where \(N_{xx}, N_{yy}\) and \(N_{xy}\) are forces per unit length. The middle surface strain/displacement equations and in-plane stress resultant constitutive equations are as follows:

\[
\begin{align*}
\varepsilon_{xx} &= \frac{\partial u}{\partial x} + \frac{1}{2} \left(\frac{\partial w}{\partial x}\right)^2 = \frac{1}{Eh}(N_{xx} - \nu N_{yy}), \\
\varepsilon_{yy} &= \frac{\partial v}{\partial y} + \frac{1}{2} \left(\frac{\partial w}{\partial y}\right)^2 = \frac{1}{Eh}(N_{yy} - \nu N_{xx}), \\
\varepsilon_{xy} = \varepsilon_{yx} &= \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial w}{\partial x} \frac{\partial w}{\partial y}\right) = \frac{(1 + \nu)}{Eh} N_{xy},
\end{align*}
\]

where \(\varepsilon_{xx}, \varepsilon_{xy}\) and \(\varepsilon_{yy}\) are the strains.

Figure 3: Sketch of dimensions and axes for modelling strip deformation across rollers.
2.1 Boundary conditions

Equation (1) must be solved subject to boundary conditions, appropriate to the problem, which will now be derived. The lateral edges at $y = \pm d$ are assumed to be free from stress and moment. The derivation of the ‘free edge’ conditions for plate theory, which is due to Kirchhoff, is given in [11] (p. 83). The conditions are that the normal component of the bending moment and the effective shear must be zero on the free edges, and the in-plane stress resultant must also vanish there. In this case the first two boundary conditions are

$$
\begin{align*}
\frac{\partial^2 w}{\partial y^2} + \nu \frac{\partial^2 w}{\partial x^2} &= 0 \\
\frac{\partial^2 w}{\partial y^3} + (2 - \nu) \frac{\partial^2 w}{\partial x^2 \partial y} &= 0
\end{align*}
$$

at $y = \pm d$, \hspace{1cm} (4)

and the zero in-plane tension on the boundary is

$$N_{yy} = \phi_{xx} = 0, \quad \text{and} \quad N_{xy} = -\phi_{xy} = 0 \quad \text{at} \quad y = \pm d. \hspace{1cm} (5)$$

Here we assume that the plate is fully clamped at the rollers at $x = \pm \ell$, which is the simplest possible boundary condition for perfectly aligned rollers. This implies the following conditions on the strains and displacements:

$$
\begin{align*}
w &= \frac{\partial w}{\partial x} = 0, \quad v = \frac{\partial v}{\partial y} = 0 \\
N_{xx} &= \frac{\partial^2 \phi}{\partial y^2} = N_0 f(y)
\end{align*}
$$

on $x = \pm \ell$ \hspace{1cm} (6)

where $f(y)$ is a shape function with unit order of magnitude. To obtain the other boundary condition at the rollers, the above boundary conditions are used to evaluate the strain $e_{yy}$ in equations (3) and we obtain $e_{yy} = 0 = (N_{yy} - \nu N_{xx})/Eh$ on $x = \pm \ell$ so that

$$N_{yy} = \phi_{xx} = \nu N_0 f(y) \quad \text{on} \quad x = \pm \ell, \hspace{1cm} (7)$$

where $N_0$ is the magnitude of the tension applied to the plate between the rollers. Here we take $f(y) = 1 - y^2$, as is appropriate for convex rollers. This completes the formulation of the problem.

2.2 Dimensionless equations

In order to construct a sensible non-dimensionalisation for these equations we note that the order of magnitude of the strain in the $x$-direction caused by the tension $N_0$ is $N_0/Eh$ so that the order of magnitude of the in-plane displacements $u, v$ will have a maximum value of $(N_0/Eh)d$. From the strain/displacement equations (2) we note that the contribution of the lateral displacement $w$ to the middle-surface strain suggests that to be significant the order of magnitude of $w$ will need to be $(\sqrt{N_0/Eh})d$. With this in mind we introduce the following dimensionless (barred) variables:

$$
\begin{align*}
(x, y) &= \frac{(x, y)}{d}, \quad (u, v) = \frac{(u, v)}{(N_0/Eh)d}, \quad w = \frac{w}{(\sqrt{N_0/Eh})d}, \\
\bar{\phi} &= \frac{\phi}{N_0 d^2}
\end{align*}
$$

\hspace{1cm} (8)

$$\bar{\ell} = \frac{\ell}{d}, \quad \bar{N}_{\alpha \beta} = \frac{N_{\alpha \beta}}{N_0}, \quad \bar{\phi} = \frac{\phi}{N_0 d^2}.$$
Note that as all quantities will be dimensionless from now on the barred notation will be dropped.

The dimensionless forms of equation (1) are

\[ \nabla^4 \phi = -\frac{1}{2} L(w, w), \quad \nabla^4 w = \lambda L(\phi, w), \quad (9) \]

where \( \lambda = \frac{12(1 - \nu^2) N_0 d^2}{E h^3} \)

is the non-dimensional tension applied to the plate.

Equations (3) become

\[ e_{xx} = \frac{\partial u}{\partial x} + \frac{1}{2} \left( \frac{\partial w}{\partial x} \right)^2 = \phi_{yy} - \nu \phi_{xx}, \]

\[ e_{yy} = \frac{\partial v}{\partial y} + \frac{1}{2} \left( \frac{\partial w}{\partial y} \right)^2 = \phi_{xx} - \nu \phi_{yy}, \]

\[ e_{xy} = e_{yx} = \frac{1}{2} \left( \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial w}{\partial x} \frac{\partial w}{\partial y} \right) = -(1 + \nu) \phi_{xy}. \quad (10) \]

The dimensionless boundary conditions corresponding to (4) to (7) are obtained by setting \( N_0 = 1 \) in these equations.

3 Solution procedure

For moderate levels of tension \( N_0 \) the plate will remain flat. The question is: at what level of tension do wrinkles appear? Thus we solve the above system as a bifurcation problem for the load parameter \( \lambda \) using an iterative numerical solution as follows.

Step one is to solve the following problem for \( \phi^{(0)} \) when \( w^{(0)} = 0 \):

\[ \nabla^4 \phi^{(0)} = 0, \quad (11) \]

subject to the boundary conditions (5) and (7) (with \( N_0 = 1 \)).

Now solve the following eigenvalue problem to find the first approximation to the critical load parameter \( \lambda^{(1)} \) and the wrinkle pattern \( w^{(1)} \),

\[ \nabla^4 w^{(1)} = \lambda^{(1)} L(\phi^{(0)}, w^{(1)}), \quad (12) \]

subject to boundary conditions (4) and (6) (with \( N_0 = 1 \)). Let the eigenfunction be normalised to some (small) value \( \max|w^{(1)}(x, y)| = \varepsilon \). This solution can now be used to initiate the following iterative procedure.

At the \( n \)th iteration \( \phi^{(n)} \), \( w^{(n+1)} \) satisfy the equations

\[ \nabla^4 \phi^{(n)} = -\frac{1}{2} L(w^{(n)}, w^{(n)}), \quad (13) \]

subject to boundary conditions (5) and (7) and

\[ \nabla^4 w^{(n+1)} = \lambda^{(n+1)} L(\phi^{(n)}, w^{(n+1)}), \quad (14) \]
subject to boundary conditions (4) and (6). This process is repeated for \( n = 1, 2, 3, \ldots \), until convergence is achieved. The iteration is now repeated for a new value of \( \varepsilon \) and in this way the bifurcation diagram showing \( \lambda \) as a function of \( \varepsilon \) can be constructed.

Equations (13) and (14) are solved by discretising using second order finite differences. With \( x_i = i \Delta, i = 1 \ldots m_x, y_j = j \Delta, j = 1 \ldots m_y, \Delta = 2/(m_y - 1) \), (so that \( m_x/m_y = \ell \)) write \( \phi_{i,j}^{(n)} = \phi^{(n)}(x_i, y_j) \) and \( w_{i,j}^{(n)} = w^{(n)}(x_i, y_j) \) and then the discrete form of (13) is

\[
D_\phi \Phi^{(n)} = \frac{1}{2} r^{(n)} .
\] 

Here \( D_\phi \) is the \( m_x m_y \times m_x m_y \) (sparse) banded matrix arising from the finite difference implementation of the biharmonic operator \( \nabla^4 \), \( \Phi^{(n)} \equiv \phi_{i,j}^{(n)} \) is the vector of unknown values of the stress function \( \phi \) at iteration \( n \) and \( r^{(n)} \) is the vector corresponding to the finite difference form of the term \(-L(w^{(n)}, w^{(n)})/2\). (Note that at each iteration \( n \), \( w^{(n)} \) is known, with \( w^{(0)} = 0 \).) The boundary conditions on \( \phi \), equations (5) and (7), are also discretised to second order, with guard points added on each edge of the rectangular domain to allow straightforward treatment of the condition \( \phi_{xx} = 0 \) on \( y = \pm d \). Similarly the discrete form of (14) is

\[
D_w w^{(n+1)} = \lambda B w^{(n+1)},
\]

where \( w^{(n+1)} \equiv w_{i,j}^{(n+1)} \) is the vector of the unknown values of the vertical deformation \( w \) at iteration \( n + 1 \), and where \( B \) is the sparse banded matrix for the discrete form of the operator on the right hand side of (14), i.e.

\[
B \equiv \phi_{xx}^{(n)} \frac{\partial^2}{\partial y^2} - 2 \phi_{xy}^{(n)} \frac{\partial^2}{\partial x \partial y} + \phi_{yy}^{(n)} \frac{\partial^2}{\partial x^2}
\]

with \( \phi^{(n)} \) known for each \( n \) from (15). The sparse banded matrix \( D_w \) represents the biharmonic operator \( \nabla^4 \) but differs from \( D_\phi \) in that the boundary conditions (4) for \( w \) are included rather than those for \( \phi \), again with the use of guard points. Equations (15) and (16) are solved using MATLAB. At each iteration, the linear problem for \( \Phi^{(n)} \), equation (15) is inverted. When \( n = 0 \), the right hand side \( r \) is zero except at boundary points. Once \( \Phi^{(n)} \) is known, the eigenvalue problem (16) must be solved. In practice, it was found that rewriting (16) as

\[
B w^{(n+1)} = \frac{1}{\lambda} D_w w^{(n+1)} = \hat{\lambda} D_w w^{(n+1)},
\]

and solving for \( \hat{\lambda} = 1/\lambda \) was more satisfactory than attempting to solve (16) directly. Equation (18) is a generalized eigenvalue problem and \( B \) is not positive definite symmetric. Therefore the inbuilt MATLAB sparse eigenvalue solver \texttt{eigs.m} cannot be used. However, the routine \texttt{ahbeigs.m} ([1]) solves this more general problem and was used for all results presented here. The four modes with smallest positive eigenvalue \( \lambda \) are those considered in the remainder of this paper; these correspond to the largest positive values of \( \hat{\lambda} \). (Modes with positive values of \( \lambda \) are those that arise when a plate is subject to longitudinal stretching; longitudinal compression gives negative values of \( \lambda \) and buckling modes.) The iteration process must be carried through to convergence for each value of the maximum absolute vertical deformation \( \varepsilon \) and for each of the four eigenvalues corresponding to that value of \( \varepsilon \). The convergence criterion used was that the absolute relative change in the eigenvalue was less than \( 10^{-6} \). No difficulties were obtained with lack of convergence in any of the cases treated here. However,
we note that more sophisticated methods for the determination of the eigenvalues have been used successfully for the related buckling problem, where the plate is subjected to longitudinal compression (see, for example, [3, 5, 9]). In particular, the numerical results of [3] are consistent with the analytical results of [10]. These more advanced numerical techniques should also work for the plate under tension problem considered in the present work. As all calculations use equal increments in $x$ and $y$, cases where the aspect ratio $\alpha = \ell/d$ are large require significantly more computational resources. In order to retain reasonable levels of accuracy in estimating the eigenvalues, only values of $\alpha$ less than or equal to three are considered here. Explicitly using odd/even symmetries for particular modes in both $x$- and $y$-directions and moving to a fully-compiled language will allow the range of computation to be extended.

4 Results

In this section we present results for the force distribution, eigenmodes of deformation and corresponding eigenvalues for two specific cases, the first a square plate where the aspect ratio $\alpha = 1$ and the second an elongated plate, where $\alpha = 3$, with the second case having more relevance to the BlueScope steel problem.

In Figure 4, surface plots of $N_{yy} = \phi_{xx}$ are shown for the case of zero vertical deformation, i.e. corresponding to the solution of $\nabla^4 \phi = 0$. In Figure 4(a), where $\alpha = 1$, we see a significant region in the centre of the plate where $N_{yy}$ is negative, corresponding to a region of compression. In Figure 4(b), where $\alpha$ is much larger, the region of compression has split into two symmetric regions near to the ends $x = \pm \ell$, i.e. close to the rollers. This localisation of the compressive forces presumably remains for larger values of $\alpha$. These results confirm the presence of a significant compressive force just upstream from the downstream roller, as proposed at MISG2007.

Figures 5 and 6 show the first four eigenmodes of deformation at iteration $n = 0$, for $\alpha = 1$ and $\alpha = 3$ respectively. These correspond to the linearised approximation where the amplitude of $w(x,y)$ is not known or equivalently the case of vanishing amplitude, i.e. $\varepsilon = 0$. These modes are those with the four smallest positive values $\lambda$, which by (9) corresponds to those that first arise under longitudinal tension $N_0$ applied at the ends of the plate. In Figure 5, for the square plate all modes are even in $x$, with the first and third modes odd in $y$ and the second and fourth even in $y$. The third and fourth modes have significantly larger eigenvalues (corresponding physically to greater tension $N_0$) but with half the transverse wavelength. Comparing these results with Figure 6, we see corresponding modes for the elongated plate, but with deformation more localised in the vicinity of the rollers, as expected from the stress distribution. Furthermore, modes 3 and 4 are odd in $x$ for the rectangular plate, rather than even. This occurs because the ordering of the higher order modes changes as $\alpha$ is varied between 1 and 3.

We now consider solutions of the full nonlinear system, as discretised in equations (15) and (16). It turns out that the structure of the eigenmodes of deformation is fairly insensitive to the magnitude of deformation $\varepsilon$, at least for the moderate amplitudes considered here, so that the results displayed in Figure 5 and 6 can be viewed as holding across all values of $\varepsilon$. Thus we present here only the variation of the eigenvalues $\lambda$ with $\varepsilon$ for different resolutions as determined by the number of mesh-points $m_x$, $m_y$ in the $x$- and $y$-directions respectively. In all cases no more than 6 iterations were required to obtain convergence for a given value of
\( \varepsilon \). Converged results from a previous value of \( \varepsilon \), e.g. \( \varepsilon = 0.5 \) are used as the starting iteration for the next value of e.g. \( \varepsilon = 0.55 \), in order to speed up the process. This also has the associated benefit of ensuring that a particular mode is followed when the eigenvalues of two modes cross.

In order to improve the accuracy of the predictions presented here an extrapolation procedure is used. As the finite difference approximations to the governing equations and the boundary conditions are second order accurate we write

\[
\lambda_n(\varepsilon) = \lambda_\infty(\varepsilon) + \frac{\lambda_1(\varepsilon)}{m_y^2}
\]

and then use results at two different values of \( m_y \) to find the two corresponding values of \( \lambda_\infty(\varepsilon) \) and \( \lambda_1(\varepsilon) \) for each \( \varepsilon \). Convergence for the first and third modes for the case \( \alpha = 3 \) is shown in Figure 7, using calculations carried out with \( m_x = 240, m_y = 80 \) and \( m_x = 180, m_y = 60 \). The relative errors clearly increase with mode number, but are of the order of 1% or less in the worst case at the highest resolution employed. Results for \( \alpha = 1 \) are somewhat more accurate.
The variation of the eigenvalues with vertical amplitude is quite small, indicating that although large values of applied end-tension are needed to excite deformation, small changes from these values will lead to relatively large changes in amplitude.

Figure 8 shows the extrapolated eigenvalues $\lambda_\infty$ as a function of the vertical amplitude $\varepsilon$ for $\alpha = 1$. These extrapolated results were found using results calculated with $m_x = m_y = 90$ and $m_x = m_y = 120$. The modes clearly separate into two pairs, with the smaller eigenvalues of magnitude around 1200, while the larger eigenvalues are nearly twice as large. As mentioned earlier for each pair there is one mode that is odd in $y$ while the other is even in $y$.

Figure 9, for $\alpha = 3$, shows results for $\lambda_\infty$ similar to those displayed in Figure 8, determined as in Figure 7. The smallest eigenvalues are now significantly larger than those found for the square plate (nearly twice as big), indicating that greater tension is required to cause deformation in the elongated plate case. Although the eigenvalues for the two higher modes are also greater than for the square plate case, the difference between the values for the eigenvalues for the higher and lower modes is significantly reduced. It is possible that as $\alpha \to \infty$ the eigenvalues will continue to grow closer, or even cross, although this remains to be determined.

To see the relevance of these results for the BlueScope steel problem consider the following approximate data (Dr Andrew Dixon, private communication). Typical values for the thick-
Figure 6: The eigenmodes corresponding to the four smallest positive eigenvalues $\lambda$ for a rectangular plate with $\alpha = 3$. The amplitude scale is arbitrary.

ness and depth of the thin sheet steel are $h = 0.5\text{mm}$ and $d = 0.5\text{m}$ respectively. Young's modulus for steel has the value $E = 2 \times 10^{11} \text{N/m}^2$ while typical applied end-stresses are of order of magnitude $N_0/h = 10^7 \text{N/m}^2$. See Table 1 for a list of approximate values of relevant material properties. Using these values in the relation (9) for $\lambda$ gives $\lambda = O(500)$ in practice. This is much less than the critical values of over 2000 we calculate are needed for buckling. Thus significantly greater tensions would have to be applied in a perfectly symmetric situation as considered here to generate any significant distortion, as it appears the practical situation corresponds to large values of $\alpha$. However, if there is any asymmetry in the applied forces, due for example to poor setup of equipment, it seems likely that deformation will arise at smaller applied tension. This implies that when wrinkles are observed in practice, they are likely to be due to roller misalignment, since our symmetric theory says tensions need to be four times the usual value to get wrinkling. However, the problem of asymmetry due to roller misalignment is beyond the scope of the present work and will be taken up in a subsequent paper where the parametric dependence of the eigenvalues on the degree of asymmetry will be explored.

If we assume for the moment that the applied end tension is sufficient to lead to plate deformation then further we can ask whether the amplitude of vertical deformation would be sufficient to cause plastic deformation. On geometric grounds it was proposed at MISG
Figure 7: Convergence of the eigenvalues for $\alpha = 3$. The results for $m_x = 180$, $m_y = 60$ are shown dot-dashed (---), the highest resolution results where $m_x = 240$, $m_y = 80$ with a dashed curve (- - -) and the extrapolated results $\lambda_\infty$ as a solid line (---).

that curvature of the plate would be sufficient to cause plastic deformation if the radius of curvature, say $r$ is less than the critical value $r_c = Eh/2\sigma_{\text{max}} \approx 15\text{cm}$, where $\sigma_{\text{max}}$ is the yield stress for steel (see Table 1). From Figures 5 and 6 we see that the wavelength in the $y$-direction is of order $d/q$ with (approximately) $q = 3/4$ for modes 1 and 2, and $q = 1$ for modes 3 and 4, where $q$ is a non-dimensional wavenumber. Using $q = 3/4$ and approximating the distortion of the plate as approximately sinusoidal, the ratio of deformation to wavelength in the $y$-direction corresponding to the critical radius of curvature $r_c$ is approximately 1/20. Therefore in dimensional variables for plastic deformation we require (recall non-dimensional $w$ is scaled by $d$)

$$\frac{1}{20} \leq \frac{w}{\text{wavelength}} = q\sqrt{N_0/Eh} \varepsilon = \frac{3}{4d} \sqrt{\frac{\lambda}{12(1-\nu^2)}} \varepsilon,$$

and we find that, using the same approximate data as above,

$$\lambda \varepsilon^2 \geq 6 \times 10^4.$$

For example, in Figure 9, where $\lambda \approx 2000$ for the lower modes, we require $\varepsilon \approx 5$ to allow for
Figure 8: The four smallest positive eigenvalues $\lambda_\infty$ for a square plate ($\alpha = 1$) as a function of the maximum vertical deformation $\varepsilon$. The eigenvalues are determined using the extrapolation procedure described in the text.

plastic deformation and so provide permanent distortion. This corresponds to a dimensional amplitude of

$$w = \varepsilon d \sqrt{\frac{N_0}{Eh}} = \varepsilon \sqrt{\frac{10^9}{2 \times 10^{11}}} \approx 20 \text{ mm}.$$  \hspace{1cm} (22)

Once such a transverse buckle has been created near the roller, we hypothesise that it undergoes a process of ironing-in, thus creating a permanent wrinkle at the roller. This process is discussed further in section 7.

5 An energy method

Paper [2] studied the wrinkling of a thin elastic sheet under tension, far from onset, using the principle of inextensibility to provide a lateral compression, and a Lagrangian energy development. They looked for periodic solutions to the Euler-Lagrange equations, and solved
the resulting Sturm-Liouville problem to find the expressions

\[ L = \frac{\sqrt{4\pi \ell h}}{[3(1 - \nu^2)\gamma]^{\frac{1}{4}}}, \quad a = \sqrt{2\nu \ell h} \left[ \frac{16\gamma}{3\pi^2(1 - \nu^2)} \right]^\frac{1}{4}, \quad (23) \]

for the wavelength \( L \) and amplitude \( a \) of the wrinkles, where the other symbols used here are as defined in Table 1. Using \( \ell \approx 10m \) gives \( a \approx 3mm \) and \( L \approx 2m \).

These are quite different results to those of the previous section. The analysis of [2] does not consider the possibility that the steel does not buckle under smaller tensions - it assumes that the lateral compression caused by the applied tension is matched by bending energy across the sheet. That is, no bifurcation phenomenon is considered, and the sheet is assumed to always buckle. The question addressed is how much it buckles, if energy is conserved. Such an analysis might be considered to be more relevant in the presence of asymmetries, like small misalignments of the rollers, which would reduce the critical parameter value at which buckling occurs.
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line speed</td>
<td>3 m/s</td>
</tr>
<tr>
<td>Distance between rollers</td>
<td>2ℓ ≈10–20 m</td>
</tr>
<tr>
<td>Roller radius</td>
<td>R = 25 – 30cm</td>
</tr>
<tr>
<td>Crown height of rollers</td>
<td>0.1 mm</td>
</tr>
<tr>
<td>Roll location accuracy</td>
<td>1 mm</td>
</tr>
<tr>
<td>Angle of misalignment of rollers</td>
<td>up to 0.06°</td>
</tr>
<tr>
<td>Steel strip thickness</td>
<td>h ≈ 0.3mm</td>
</tr>
<tr>
<td>Steel strip width</td>
<td>2d ≈ 0.5m</td>
</tr>
<tr>
<td>Young’s Modulus</td>
<td>E ≈ 2 × 10^11 Pa</td>
</tr>
<tr>
<td>Poisson’s ratio for steel</td>
<td>ν_{steel} ≈ 0.28</td>
</tr>
<tr>
<td>Poisson’s ratio for paper</td>
<td>ν_{paper} ≈ 0.01</td>
</tr>
<tr>
<td>Flexural rigidity of steel</td>
<td>D ≈ Eh^3/(12(1 − ν^2)) ≈ 0.49N.m</td>
</tr>
<tr>
<td>Line tension</td>
<td>N₀/h = σ ≈ 10MPa</td>
</tr>
<tr>
<td>Longitudinal stretching strain</td>
<td>γ ≈ N₀/(Eh)</td>
</tr>
<tr>
<td>Yield stress for steel</td>
<td>σ_{max} ≈ 200MPa</td>
</tr>
<tr>
<td>Coefficient of friction for steel</td>
<td>μ ≈ 0.1</td>
</tr>
<tr>
<td>Areal density of steel</td>
<td>ρ ≈ 2kg/m²</td>
</tr>
</tbody>
</table>

Table 1: Material properties used for rolled steel

6 Waves on steel

The group also considered the speeds of bending waves and compressional waves on and in steel, compared with the typical operational linespeeds of the steel over the rollers. Bending waves are governed by the balance

\[ D \frac{\partial^4 w}{\partial x^4} = -\rho \frac{\partial^2 w}{\partial t^2} \]  

(24)

where ρ is the areal density, about 2 kg/m². Setting \( w = \exp[i(kx - \omega t)] \) where \( \omega = 1/\tau \) and \( \tau \) is the period of the waves, gives

\[ k^4 D = \omega^2 \rho \Rightarrow \tau = L^2 \sqrt{\frac{\rho}{D}} \]  

(25)

and hence the wavespeed is \( c = \frac{k}{\tau} \approx \frac{1}{L} \), since \( D \approx \rho \). Hence for wavespeed to be greater than the linespeed of 3 m/s, the wavelength needs to be shorter than 1/3 m. These shorter bending waves can propagate upstream against the flow of the steel. In particular, so can the waves predicted by the theory of [2] discussed in the previous section.

Tensional wave speed may be estimated by balancing

\[ N₀ \frac{\partial^2 w}{\partial x^2} \sim \rho \frac{\partial^2 w}{\partial t^2}, \]  

(26)

which gives the wavespeed squared as \( c^2 \sim \frac{N₀}{\rho} \), which leads to \( c \approx 40\text{m/s} \), which is much faster than linespeed. Hence tension changes propagate rapidly compared with linespeed, and linespeed may be ignored when calculating stress fields as in the first section.
7 Ironing in the wrinkles

Here we consider whether the pressure on the steel as it goes around a roller is sufficient to iron-in a wrinkle. To iron the wrinkle in, yield stress needs to be exceeded in the steel as it passes over the roller. The coiling pressure gives a measure of the downward pressure that is applied to a wrinkle. This is

\[ P = \frac{\sigma h}{R} \approx 0.01 \text{ MPa}, \] (27)

which is taken to act towards the roller axis as a force per unit length across the roller and which also gives rise to a total uniform loading \( P_b \) towards the roller axis acting on the wrinkle. The numerical value in equation (27) is found using the values given in Table 1.

First we use a very simplified model of the deformation shown in Figure 10(a) in order to approximate the relation between the height \( a \), the width of the wrinkle \( b \) and the contact distance \( \delta \). We consider a strip of plate of unit width in the direction aligned with the roller axis and approximate the deflection \( w(y) \) given by the triangular function shown in Figure 10(b). Assuming symmetry of this triangular shape, the forces acting on the section AB of the plate are shown in Figure 10(c). \( H = \mu P \delta \) is the force due to the friction between the roller and the section of the plate fully in contact with the roller.

![Diagram](image)

Figure 10: (a) Sketch of a wrinkle wrapping around a roller with (b) the approximate deflection \( w(y) \) and (c) the corresponding free body diagram.

Considering the moment equilibrium about the point B we obtain \( H_a = P_b^2/8 \). After rearrangement \( H = \mu P \delta = P_b^2/8a \), so that, with the friction coefficient \( \mu = 0.1, \delta = 5b^2/4a \) which provides an approximate relation between \( a, b \) and \( \delta \). We also have the relation \( 2\delta + b = 2d \).

Finally we require the critical lateral force per unit length \( F_c \) that is required to cause this buckle in a plate. We will approximate this value by using the following result on plate
buckling from [12], p.352,

\[
\frac{m^2 F_c}{\pi^2 b^2 D} = \left( \frac{m^2}{b^2} + \frac{n^2}{4L^2} \right)^2
\]

where \( m \) is the lateral wavenumber, \( n \) is the longitudinal wavenumber, and \( 2L \) is the length of the plate. Taking \( n = 0 \) and \( m = 1 \) gives the critical lateral force per unit length for buckling as \( F_c = \frac{\pi^2 D}{b^2} \).

For friction to provide enough lateral force to sustain the wrinkle of amplitude \( a \) as the sheet passes over the free roller we require \( H > F_c \), which implies

\[
H = \mu P \delta \geq F_c = \frac{\pi^2 D}{b^2}.
\] (28)

Substituting for \( \delta \) we obtain the cubic equation for the value of \( b \) at the critical buckling load

\[
b^2(2d - b) = \frac{2\pi^2 D}{\mu P},
\] (29)

or \( b^2(0.5 - b) \approx 0.01 \), by using the numerical value of \( P \) given by equation (27) and the values of \( \mu, d \) and \( D \) in Table 1. The cubic equation for \( b \) has two positive roots, \( b \approx 0.18m \) and \( b \approx 0.45m \), where the larger value is near to the full strip width. For \( b \) between these two limits, friction can hold against the critical lateral bending force, and the wrinkle is ironed-in as it passes over the roller. Hence if the contact length between roller and steel is in the range 160mm \( \geq \delta \geq 25mm \), there is enough friction to hold the steel edges in place and prevent the wrinkle from disappearing. The smallest wrinkle width that can be supported according to this model is about \( b = 175mm \) across. This is significantly larger than the measurement shown in Figure 1. However, it appears that the pressure induced by the longitudinal stress at the wrinkle as it wraps around the roller gives a much larger value for \( P \) than that used here, so providing a much smaller critical value for the smallest wrinkle width \( b \) that can be supported (see [4]) and correspondingly a value more closely in accord with observations.

The corresponding amplitude of the wrinkle for this range of \( b \) (or equivalently \( \delta \)) can be estimated. Using \( H = \mu P \delta = Pb^2/8a \), we have

\[
a = b^2/(8\mu \delta) = b^2/(4\mu(2d - b)).
\] (30)

Therefore, according to this balance between downward pressure and friction the smallest wrinkle amplitude \( a \) that can be supported is about 250mm when \( b = 180mm \). The amplitude \( a \) increases monotonically as \( b \) is increased. This large value of the amplitude is clearly physically unrealistic, but again is very significantly reduced if the enhanced pressure proposed by [4] is used.

8 Conclusions

The formation of longitudinal wrinkles in thin steel sheet can be identified with the modes of deformation occurring when a thin plate is placed under sufficiently large longitudinal tension. Solving the biharmonic equation describing the stress distribution in a rectangular plate, with free transverse edges and and applied stress at the ends corresponding to that for convex rollers, shows that a local compressive in-plane stress is produced transverse to the plate. In response to this compressive stress wrinkles aligned in the longitudinal direction may arise,
depending on the magnitude of the applied tension. However, the critical stress for wrinkles to initiate is about four times the stress we predict in the BlueScope Steel setup, suggesting that any longitudinal wrinkles observed are due to asymmetries from, say, misalignment of rollers or damaged roller surfaces.

The zeroth order approximation to the full von Kármán plate bending equations gives rise to a linear eigenvalue problem for the applied longitudinal tension, with any particular eigenvalue giving the critical (non-dimensionalised) applied force above which the corresponding eigenmode of vertical deformation will arise. However, this linearised problem does not give any information about the amplitude of vertical deformation. In order to determine the amplitude, the nonlinear von Kármán equations must be iterated. This process results in generally small changes in the stress distribution in the plate and the form of the corresponding modes of vertical deformation as compared to the linear case.

As the eigenvalues are found to vary quite slowly with the vertical amplitude, the physical conclusion is that if the applied tension is sufficient to excite a particular mode, then the amplitude of deformation may be quite significant for small changes in applied tension above critical. The expression (20) indicates that local plastic deformation may be observed if the longitudinal tension is sufficient to cause wrinkling distortions. In particular, this is expected to occur at the point where maximum deformation is observed. This will occur at the mid-line of the plate for the low-order mode even in $y$ (mode 2 in Figures 5 and 6).

According to the model used in section 7, only quite wide, large amplitude wrinkles can be ironed-in at the roller. In order to provide a mechanism for the formation of unwanted permanent distortions of sheet metal as observed in practice in some cases, significantly greater local pressure at the wrinkle is required. One proposed mechanism for this is the enhanced longitudinal stress at the wrinkle as it passes over the roller, as discussed in [4].

Acknowledgements

The moderators would like to thank the MISG participants who contributed so much to this work: Bob Anderssen, Steve Barry, John Cogill, Grant Cox, Jonathan Crook, Bronwyn Hayek, Roslyn Hickson, Da Eun Jeon, Scott McCue, Robert McKibbin, Steve Taylor and Peter Tritzsch. We also thank the BlueScope steel representatives, Andrew Dixon, Daniel Yuen and Maria Cozijnsen for their participation, expertise, and strong support in all phases of the project.

References


The maritime surveillance problem

Philip Kilby
NICTA and Australian National University

Patrick Tobin
Swinburne University of Technology

Ruth Lascombe
University of Melbourne

Steven I. Barry and Roslyn Hickson
University of New South Wales @ ADFA

1 Introduction

The problem examined here is that of using aircraft to find and classify ships at sea. Identifying the position and type of ships within, or close to, Australian sea borders is an important part of Australia’s national security activities. Conducting the surveillance within Australia’s sea borders is a mandated role for the Australian Defence Forces. Currently it is carried out by the Royal Australian Air Force. The Defence Science and Technology Organisation (DSTO) is looking at how these missions can be flown with the maximum efficiency.

There are two important considerations in running these surveillance missions. The first is to reduce as much as possible the number of ships that are “missed” in the survey. The second is to make the flying time as small as possible. In this study we will describe the problem in detail, and give some initial results on how well these two objectives can be met under different scenarios. There are three questions the DSTO are particularly interested in:

• What is the effect of treating the targets as stationary?
• Can flying time be improved?
• Is execution time of methods an issue?

The aim of the study was to find answers to these questions, or at least determine how they might be answered.

In Section 2 we describe the Maritime Surveillance Problem as presented to the MISG, and in Section 3 the problem is formulated in terms of related problems, particularly from the routing literature. In Section 4 the simulation system that was developed during the week is described. This system was able to go a long way towards answering DSTO’s original questions. The results are presented in Section 5. On the way to answering these questions, a number of related questions were also examined. The results are presented here, including

• determining the heading to intercept a moving ship (Section 6),
• estimating tour length (Section 7), and
• finding what proportion of ships are detected? (Section 8).
2 The maritime surveillance problem

The basic objective of maritime surveillance is to monitor the ocean regions around Australia’s coastline. A particular region of ocean may be identified, and some part of that region assigned to a single surveillance aircraft (called a platform). The surveillance problem is therefore partitioned into smaller problems, one for each ‘Area of Interest’ (AI). In this study we consider the problem for a single AI. Typically, an AI is a square or rectangle with sides in the range of 200-300 n mile\(^1\).

Various types of aircraft are used for surveillance. Typically, it is an AP-3C Orion aircraft, but may be a helicopter. The flight speed of the platform is an important variable in determining the flight path. Values range from 100-350 kn.

Each AI has a default flight path, defined by a list of way points, known at the start of the mission. It is a requirement that all way points be visited, and that they be visited in the order specified in the default flight path. The way points are selected so that if the platform flies the default path, every point within the AI comes within surveillance range. An example set of way points is given in Figure 1. If no ship is seen, the platform would simply fly the route defined by default flight path.

\[\text{Figure 1: Way-points - blue dots - defining the default flight path.}\]

As the platform proceeds, ships will come within range of the radar on board. The distance at which a ship can be detected is variable, and depends upon factors such as the weather, the flight altitude, and the type of equipment used on the platform. For the purposes of this study, the detection distance is treated as an input variable. Typical values for detection distance are 20-100 n mile. We make the simplifying assumption that detection is binary – either a

\(^{1}\)The Nautical Mile (n mile) is the standard unit of distance at sea. It is equivalent to one minute of latitude along any line of longitude, and is about 1852m. Its associated speed unit is the knot (kn) = \(\frac{1}{60}\) n mile/hour.
ship is detected or not. In reality, there is a drop-off in the sensitivity near the boundaries of the radar where detection is fuzzy, but for this exercise we are ignoring these effects. We are also assuming that when a ship is detected, we can tell the speed and direction of travel. This is also a simplification, but one that is fairly reasonable, as updates to these estimates can be accommodated as the platform flies closer to the target. Another simplification made here is that ships do not change their speed or direction. This is an oversimplification that a practical system would not be able to assume. However for the questions examined here, it is a reasonable assumption.

When the platform detects a target, it flies to a point that is close enough to be able to classify the ship. “Classifying” a target involves seeing it well enough to determine the type of the ship. Once again, the distance at which this classification can take place is variable depending on factors such as the weather. We use an input variable classification distance to determine what is “close enough”. Typical values are 0-20 n mile.

One restriction that is placed on routes in this study is that the platform is not permitted to leave the Area of Interest. That is, if a ship is detected close to the border of an AI, it may leave the AI before the platform can reach it. If this is the case, the platform is not permitted to “chase” it outside the AI. Similarly, a ship detected outside the AI must enter the AI before the platform can classify it. Note that it is also possible for a ship to pass through an AI without being detected, depending on the timing.

A second restriction is that the flight time is limited. Since all way-points should be visited, this places a hard deadline on the last way-point. It is a hard deadline as it relates to the ability of the platform to return to base without running out of fuel! This means that some targets may not be able to be classified if visiting them means the platform cannot reach the final way point in time.

As the platform is flying toward one target (or way point) other targets may be detected. During the course of a mission, more than 100 targets may be detected. At any one time, we may have tens of ships detected but not yet classified. The problem we address is to find a route which visits all these unclassified targets, plus the remaining way points, in as short a time as possible.

At all times the platform maintains a plan of the flight path. This flight path evolves as new detections are made. The route will visit all remaining way points, plus as many unclassified targets as possible. It will obey all of the constraints discussed above. Calculation of a new route is triggered whenever a new target is detected. It is also triggered when a previously detected target enters the AI. When the current target is classified, the platform selects the next target or way-point in the current route as the next point to be visited. In a practical system, other events such as a ship adopting a new speed or heading, or loss of contact with a target, may also trigger a re-calculation. However these complications are out of the scope of the current study.

2.1 Current solution method

DSTO currently model the maritime surveillance problem within a larger simulation model. This larger model is used in undertaking operational analysis in such areas as tactics development and capability assessment. An algorithm has been developed to solve the maritime surveillance sub-problem within the context of this larger simulation system. The algorithm employed is a simple single-swap crossover genetic algorithm. A population size of two is usually used, with no mutations. A fixed number of generations is used. This method is used
to optimise the route each time a trigger event occurs, such as a new ship being detected. This method makes the assumption that ships are stationary, and so must also be invoked periodically in order to re-calculate a route with updated ship positions.

3 Problem formulation

This problem can be seen as having aspects of many well-studied problems, but drawn together in a unique way. First, it has a strong connection with the Travelling Salesman Problem (TSP) [8]. In the TSP, a salesman wishes to visit a set of cities. He knows the cost of travel between each pair. What is the shortest route that visits each city exactly once, and then returns to the starting city? In the Open TSP, the requirement to return home is removed.

Our problem is much like an Open TSP (with the last city defined by the last way point). However, unlike the standard TSP, our cities – the ships – can move. The moving target TSP has received some attention in the literature [6, 9, 10, 13, 14, 17], but none of these solutions are immediately applicable to the problem at hand.

One important feature of the moving target problem is that, having decided to visit a particular target, an intercept point must be calculated. This point must minimise the flight time from the platform’s current position to a point that intercepts the direction of travel of the target ship. These calculations are discussed in more detail in Section 6.

Another complicating factor is the online nature of the problem. We do not know all of the tasks that are to be performed at the start of the mission – they are revealed only as we proceed. There has been a growing interest in online, or dynamic, routing and scheduling problems recently – [2, 3, 5, 11, 12] to cite just a few.

The crux of the online problem can be expressed in the following question: Do I spend x minutes visiting a target now, or will I perhaps be able to visit two as yet unseen targets in x minutes later in the route? The answer depends on a host of variables, including expected density of targets and remaining flight time.

Each target has a potential time window for visit. In routing terminology, a time window is the period during which a visit may take place. Since we know the position, speed and direction for a target, we know when it will enter the AI (if it has not already) and when it will leave the AI. These two times define the time window for each target.

Finally, we have the requirement that as many targets as possible are visited in the available time. In the routing literature, such problems are sometime called “prize-collecting” problems [1, 4]. The literature also has reference to the “Close-Enough” TSP where, as in our problem, it is not necessary to visit the actual location of the “city”, but it is sufficient to be “close enough” [7]. To draw these all together, we can define the problem as follows:

Maximise the number of classified targets; and within this maximum, minimise the flight time, subject to the following constraints:

- Flight time ≤ maximum flight time.
- All way-points visited in order.
- The route never leaves the AI.

In terms found in the routing literature, it is an Online Prize-collecting, Open, Close-Enough Travelling Salesman Problem with Time Windows and Precedence Constraints.
4 Simulation system

A simulation system called TPP (for Travelling Pilot Problem) was developed during MISG in order to begin answering the specific questions posed by DSTO (Section 1).

![Figure 2: A screen-shot of the TPP simulation animation program.](image)

The system had a graphical interface which allowed simulation runs to be visualised. An example screen-shot from the animation is given in Figure 2:

- The search area is boxed in grey.
- The aircraft performing the search is the red rectangle.
- The circle around the search craft is the radar detection range.
- The orange line gives the currently planned route.
- The grey line gives the default flight path.
- Targets are the triangles. Colours indicate their status:
  - Green: Classified,
  - Blue: Detected,
  - Cyan: Detected, but outside the search area (and hence ignored),
  - Pink: Undetected (within search area),
  - Purple: Undetected (outside search area).
<table>
<thead>
<tr>
<th>Key</th>
<th>Default</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxTimeMin</td>
<td>480</td>
<td>minutes</td>
<td>The maximum flight time.</td>
</tr>
<tr>
<td>areaLenNM</td>
<td>350</td>
<td>n mile</td>
<td>The x-dimension of the Area of Interest.</td>
</tr>
<tr>
<td>areaWidNM</td>
<td>350</td>
<td>n mile</td>
<td>The y-dimension of the Area of Interest.</td>
</tr>
<tr>
<td>platformSpeedKnots</td>
<td>250</td>
<td>kn</td>
<td>The speed of the platform.</td>
</tr>
<tr>
<td>targetSpeedKnots</td>
<td>10</td>
<td>kn</td>
<td>The speed of the targets.</td>
</tr>
<tr>
<td>classifyDistNM</td>
<td>0</td>
<td>n mile</td>
<td>The minimum distance the platform must approach the target, in order to classify it.</td>
</tr>
<tr>
<td>detectDistNM</td>
<td>100</td>
<td>n mile</td>
<td>The radar range of the platform.</td>
</tr>
<tr>
<td>numTargets</td>
<td>30</td>
<td></td>
<td>The total number of targets.</td>
</tr>
</tbody>
</table>

Table 1: Configuration items.

4.1 Configuration

The system has a number of configuration parameters, reflecting the variables of interest to DSTO. These are listed in Table 1.

Values for these configuration parameters appear in a file that controls the execution of the program. Unless varied for a particular test, the value of each parameter is set at the default given in Table 1. All targets are assigned the same speed. This was done to understand the effect of target speed on solutions.

The targets created are distributed uniformly over the Area of Interest (AI) and the eight neighbouring areas of equivalent size – that is, areas of the same size are created to the North, NE, E, SE, S, SW, W and NW of the original AI. Targets are distributed uniform-randomly within the larger region. This allows targets to enter the AI during the simulation from a point outside the original boundaries.

The result of this is that not all of the targets will enter the AI. A report of the number that do is given at the end of the simulation.

4.2 Solution method

Ideas developed during MISG were implemented in the TPP system. The system uses the idea of the “current tour” – that is, the list of targets and way-points in the order in which they will be visited.

When we talk about visiting a target, we always mean choosing a flight path that will intercept the current path of the target as soon as possible. This is a relatively straightforward calculation, although practical considerations such as limits on the turning circle of platforms may influence the calculation. These considerations are discussed in Section 6. The intercept calculation used in the simulation system does not account for turning circles, and always heads directly toward the intercept point regardless of the classification distance.

We use “visit” to refer to both target ships and way-points. Thus a tour is a sequence of visits. The system operates using the procedure described in Figure 3.

If conditions 2a or 2b are met, we will always have a new “next visit”. If condition 2c is
0) At the start of operations, form an initial tour considering all detected targets, plus all way-points. (If there are no detected targets, the tour will consist only of the way-points in order).
1) Move towards the first visit in the current tour.
2) Each minute, check the following conditions:
   2a) If the next visit is a ship, and it is within the classification distance, mark the target as classified and remove it from the tour.
   2b) If the next visit is a way-point and the distance to the point is sufficiently small, mark the way-point as visited and remove it from the tour.
   2c) If any new targets have been detected, attempt to create a new tour incorporating the new targets. If successful, replace the current tour with the new tour.

Figure 3: Basic TPP algorithm.

met, we may go to a new target before classifying/visiting the current target. At steps 0 and 2c, only “eligible” targets are considered for inclusion in the tour. Eligible targets are those that

- have been detected,
- have not yet been classified,
- are currently within the area of interest,
- will not cause the platform to exceed the deadline at the last waypoint (using direct flight to the target and then to the waypoint).

The method of tour construction for step 0 is a traditional construct-and-improve method, described in Sections 4.2.1 and 4.2.2. The method for extending the tour at step 2c is described in Section 4.2.3. We discuss the consequences of moving targets in Section 4.2.4.

4.2.1 Tour construction

Nearest Neighbour: One of the simplest methods for creating a route in the Travelling Salesman Problem is called “Nearest Neighbour”. At each stage, the eligible targets and next way-point are considered, and the next to be visited is simply the one that is currently the closest.

Insert Heuristic: Many insert techniques have been described, for example Solomon’s methods [18]. The version used here is very basic. The route is initialised with the list of way-points in order. Each eligible target is then considered. The cost of inserting the target between each pair of visits in the tour is calculated. The new intercept points of visits after the insert target are not calculated – we essentially assume all ships are stationary. The target is inserted into the position which gives the least increase in cost. If, after insertion, the route becomes infeasible due to a missed deadline at a target or the final way-point, the tour reverts to the previous configuration.
4.2.2 Tour improvement

After construction, and after any other change to the tour, a tour improvement procedure is called. This procedure runs a number of standard TSP improvement operators. Note that the operators are restricted so that changes that would alter the order of way-points are not considered.

The operators used are:

**Move:** Each visit is removed from its current position, and re-inserted into the (legal) position which causes the least increase in distance.

**Two-opt:** The two-opt operator removes two links in the tour, and replaces them with two others, effectively reversing the order of visits between the broken links. It is used often, and described in, for instance, [16].

**Or-opt:** First described by Or [15] and also described in [16], this procedure removes a chain of $k$ consecutive visits, and tries to re-insert it between each pair of visits in the remaining tour. If no improvement can be found, the procedure is repeated with chains of length $k - 1$, and on down to length 2. A length 1 Or-opt is exactly equivalent to the Move operator. In TPP, chains of length $(k)$ up to 5 were considered.

These improvements procedures are called in the order given. Improvements are implemented as they are found (i.e. first-found rather than best-first). When all three complete without improving the solution, improvement is finished. One important consideration is the procedure for 0-cost changes (i.e. changes where the costs before and after are equal). 0-cost changes are important as they can move the solution to a new part of the solution space from which improvements can be found. However, accepting all 0-cost changes results in cycling behaviour. In the method implemented in TPP, 0-cost improvements are accepted with probability 0.5.

4.2.3 Tour extension

At step 2c of the algorithm in Figure 3, newly detected targets are incorporated into the existing tour. In TPP, this is handled as a tour modification procedure in the following way.

1) First, an attempt is made to include the new target in the current tour. This is done by adding the new target, and running the tour improvement procedure until completion. If the resulting route is legal, it is accepted and we exit – we have a new tour with more targets. Otherwise, continue to step 2).

2) The illegal tour is examined. The target (other than the new target) that causes the greatest deviation is removed from the tour. The improvement procedure is re-run. If the resulting tour is legal, the cost of the new tour is compared to the original tour. If the original tour is shorter, then it is kept, and the addition of the new target has failed. Otherwise, we move on to step 3).

3) The new target has been added, and another target deleted, with a shorter route resulting. With the extra time now available, it may be possible to include a target that had previously been skipped. If there are any eligible targets not currently in the tour, the target that is closest to the new tour (i.e. that causes the least deviation) is identified. We return to step 1 with the identified target as the target to be inserted, and the new tour treated as the “original” tour.

The result of this procedure is either the original tour, a new tour with more targets but greater distance, or a new tour with the same number of targets but smaller distance. These outcomes are consistent with the objectives of the problem.
4.2.4 Moving target considerations

Until now, algorithms have been described in terms which largely ignore the movement of targets. We now consider how this target movement affects the algorithms. We have implemented three variants of the algorithms described above.

Stationary Ships
This is the method currently used by DSTO in their maritime surveillance search model. The ships are assumed to be stationary. The current intercept point for all ships is calculated each time a tour is to be created, and this position is used for all calculations.

The change in costs for the improvement operators described in the improvement are easy to calculate in this case. Consider a tour \(a, b, c, d, e\), and \(D(x, y)\) gives the distance between current intercept points for visits \(x\) and \(y\). The cost of moving \(d\) to follow \(a\) is

\[
D(c, e) - D(c, d) - D(d, e) + D(a, d) + D(d, b) - D(a, b).
\]

The same formula holds true for any pair \(a\) and \(b\) on the tour, regardless of whether they are before or after \(d\). In fact the expression \(D(c, e) - D(c, d) - D(d, e)\) can be pre-calculated and used for all pairs on the route.

Thus the test for acceptance of an improvement is \(O(1)\) (i.e. constant-time) operation. If we have \(n\) visits in the tour, then we have \(O(n)\) insert positions, and \(O(n)\) candidate visits to move – hence all Move type improvements can be tested in \(O(n^2)\).

Moving ships
This is a fully-dynamic version of the algorithm. In this variant, each time a tour is modified, new intercepts for all targets are calculated. This has a large effect on the testing of improvement operations. Each improvement – such as the example move discussed above – requires essentially the whole tour to be reconstructed in order to find the new intercept points. Testing of acceptance of a move improvement is now \(O(n)\), and testing all Moves is now \(O(n^3)\).

Jumpy ships
There is a mid-point between these two variants. Jumpy ships works as follows. During a particular iteration, the position of targets is treated as fixed, so improvements can be tested quickly using the same techniques as the “Stationary Ships” variant. However, once a new tour has been calculated, the position of each ship is updated. The new intercept point based on the new route is calculated. But rather than simply updating the ship’s position, it is moved to the mid-point between the old and new points. The improvement phase is re-run using the new fixed positions. Improvement and update is iterated until the position of targets converges sufficiently. If the positions do not converge sufficiently within 20 iterations, the procedure is terminated and the position updated using the order of visits in the last tour.

The advantage of not updating the position of the target in a single step is that occasionally a better route can be found that uses the fact that the target will be in a different area at a different time. It is a way of preventing the solution converging to a given point too quickly, and allows a neighbourhood of that solution to be explored.

4.2.5 Other heuristics

Pro-rata deadlines
It was seen that when there are many potential targets, the algorithms would tend to make the platform spend disproportionate amounts of time in the first part of the AI, and
then have no time to explore the rest of the space. In order to overcome this, an artificial
deadline is assigned to each way-point. The deadline is calculated pro-rata, based on how far
along the default flight path the way-point is located. So, given an 8-hour deadline, if the
way-point is \( \frac{1}{4} \) of the way along the default flight path, it would have a deadline of \( \frac{1}{4} \times 8 = 2 \) hours.

5 Results from the simulation system

5.1 Stationary ships assumption

The first question we looked at was the validity of the “stationary ships assumption” – that
is, what is the effect of target speed on the measures of effectiveness (MOEs).

The most important MOE is the classification rate – how many of the targets that are in
the area are classified? We consider “number in area” to mean the number of targets that are
ever in the area. This can mean targets that are close to the border when the mission starts,
and leave shortly afterwards, plus those that enter the area while the mission proceeds.

This definition unfortunately introduces a bias into the statistics. An efficient tour will
mean the mission is finished early, and hence targets that move into the AI after a short
while later are not counted. Unfortunately all other methods of direct comparison considered
made a larger bias. For instance taking the “number in area” count over any defined period
disadvantages an efficient tour.

An experiment was run using 100 scenarios. Target speeds of 0, 5, 10, 15 and 30 kn were
examined. Speeds up to 15 kn are common in shipping, and the upper limit of 30 kn was also
tested.

In order to look at a wide variety of scenarios, scenarios with 25, 50, 100, 200, 500, and
750 targets were run. As described previously, these counts are for the larger area, and hence
different numbers of targets will enter the AI over the course of the mission. In general, the
“number in area” count also goes up with target speed, as targets that start further away are
able to reach the AI.

The solution method described in Section 4.2 was tested. Each of the 3 methods of
updating tours – Stationary Ships, Jumpy Ships and Moving Ships (described in Section 4.2.4)
– was tested. As the improvement methods used are stochastic (in the acceptance of 0-
cost moves) each was run 5 times to arrive at a reasonable average. Average (arithmetic
mean) performance is reported here. A maximum elapsed time of 5 seconds was placed on
calculations at each time point. This equates to a restriction that a pilot not wait more than
5 seconds for a new tour to be calculated after updated positions were supplied.

The results are shown in Figures 4 to 8. The graphs show the percentage of targets
successfully classified as a function of the number of targets in the area. Figure 4 shows
that if the targets are indeed stationary, all three methods perform almost exactly as well.
The three graphs are coincident for most values. In Figure 5 (target speed 5 kn) we see that
some improvement is possible, with the effect growing with the number of targets. However
the effect is fairly marginal – Moving Ships method producing about 1.4% more classifications
than Stationary Ships (76.5% compared to 75.1%) for scenarios with 50 or more ships in area.

A similar story is told in Figures 5 and 6, with the difference growing. For target speed
10 kn, the comparison is 71.4% vs 68.4% - a 3% improvement. For 15 kn, the improvement is
4%. At the extreme (target speed 30 kn), moving ships performs 5% better on the 50+ scenar-
ios. In each case, the Jumpy Ships method produced an inferior value. For scenarios with less

41
Figure 4: Classification Rate – target speed 0 kn.

Figure 5: Classification Rate – target speed 5 kn.
Figure 6: Classification Rate – target speed 10 kn.

Figure 7: Classification Rate – target speed 15 kn.
than 50 ships, the Moving Ships and Stationary Ships methods are almost indistinguishable.

The conclusions from this are pretty clear. For the simple methods considered thus far, the stationary ships assumption is reasonable for scenarios with fewer than 50 ships. If there are more than 50 ships, however, the stationary ships assumption yields inferior results. Up to 5% more ships might be classified for targets travelling at 15 kn. At the extreme, 12% more classifications were achieved using the Moving Ships method. The Jumpy Ships method should not be used.

5.2 Time to complete mission

Next we examined the time to complete mission, that is, how long it took to reach the final way-point. We looked only at a “typical” case of target speed 10 kn. The results from the experiments reported in Section 5.1 for target speed 10 kn were used. The results are shown in Figure 9. The graph clearly has two parts. In the first part – up to about 40 targets – there is sufficient time to visit all targets. In the second part, the flight time is limited by the 8-hour maximum.

This divergence at 40 targets is also interesting in the context of Section 5.1, as it partly explains why the results there seemed to have two distinct parts. The differences between the algorithms are very small – of the order of only 1 minute in 8 hours for scenarios with less than 40 targets. The conclusion is that there is not much difference between methods on the time-to-complete MOE. Section 7 also gives some results on estimating tour length.
Figure 9: Total flying time – target speed 10 kn.

5.3 Execution time

The final question from DSTO regarded the execution time of the methods. The methods were tested on a Linux system with 2 Intel Pentium III-based CPUs, each running at 2.13 GHz and 2048 Mb cache. Figure 10 shows the average time to calculate a new route when a new target is detected. Note that a 5-second maximum time was in force during execution.

All the methods run fairly quickly, and the 5-second maximum was seldom reached. The full moving ships code runs faster than jumpy ships. This means that the time it takes for a jumpy-ships solution to converge is more than the time saved by the faster check. Hence (again), jumpy ships is not recommended. The full moving-ships code is still quite fast enough - even for the larger problems. An answer is usually returned in less than a second. These results indicate that execution time may not be important in a standalone system. However, the effect of the execution times will need to be tested in the context of repeated runs within the larger DSTO maritime surveillance model.

6 Determining the heading to intercept a moving ship

In this section we consider the problem of which heading a plane should take in order to investigate a moving ship. While easy to solve in its simplest form, the solution becomes complicated when the physical restrictions of a finite turning circle are included and when the plane only needs to get within a certain radius of the ship in order to investigate it.

For clarity we initially consider the simple problem of the straight line intersection of the plane and ship trajectory, since this provides a gentler introduction to the methodology. We then generalise this to when the plane only needs to come within a certain distance $r$ of the ship (the classification distance) and then to the case where the plane is only able to
undertake a finite turning circle with radius \( r_c \). The plane is assumed to have a constant speed \( v \), and initial heading \( \phi_i \), with the ship moving with constant speed \( w \), fixed heading \( \theta \) and initial position \((x_0, y_0)\). The aim is to find the optimal heading of the plane \( \phi \), and the resultant time to intersection, \( t_f \). We assume the plane is at \((0,0)\) when \( t = 0 \), the time when it decides to intercept. This is illustrated in Figure 11. The angles \( \xi_0 \) and \( \xi_1 \) are the parameter angles representing position around the turning circle at the beginning and end of the turn. The angle \( \eta \) is the parameter representing position around the detection circle.

### 6.1 Straight interception

Here we assume the plane has no turning circle and that it needs to intercept the path of the ship exactly. The plane is assumed to initially be at \((0,0)\). The interception point of the plane and ship is when

\[
\begin{align*}
x_0 + vt \cos \theta &= vt \cos \phi \\
y_0 + vt \sin \theta &= vt \sin \phi
\end{align*}
\]

which we have to solve for \( t \) and \( \phi \). Squaring and adding both equations gives

\[
(x_0 + vt \cos \theta)^2 + (y_0 + vt \sin \theta)^2 = v^2 t^2
\]

which, in anticipation of later solutions, is written as

\[
\alpha t^2 + \beta t + \gamma = 0, \quad \text{where} \\
\alpha = w^2 - v^2, \quad \beta = 2x_0 w \cos \theta + 2y_0 w \sin \theta, \quad \gamma = x_0^2 + y_0^2,
\]
Figure 11: Illustration of a plane turning to intercept a ship. The plane is initially on a heading $\phi_0$, moving with velocity $v$, and begins turning at $(0,0)$ at $t = 0$ with a turning circle of radius $r_c$ which has centre $(x_c, y_c)$. It ends the turn at $t = t_1$ at $(x_1, y_1)$ and moves along a heading $\phi$ until it detects the ship at the end of its detection radius $r$. The ship moves with velocity $w$ along a heading $\theta$.

giving the obvious solutions

$$t = \frac{-\beta \pm \sqrt{\beta^2 - 2\alpha \gamma}}{2\alpha}, \quad \phi = \cos^{-1}\left(\frac{x_0 + wt \cos \theta}{tv}\right).$$  

(4)

In equation (4), $\alpha < 0$ since the plane is faster than the ship, and $\gamma \geq 0$, so the equation always has real roots. In addition, $\sqrt{\beta^2 - 2\alpha \gamma} \geq \text{abs}(\beta)$, so there is at exactly one positive solution for $t$.

6.2 Interception including detection zone and turning circle

In order to consider the effect of a non-zero classification distance, a simulation system was written in MATLAB that looked at flying between two way-points, with a variable number of targets to be visited along the way. A MATLAB library routine using a genetic algorithm was used to solve the routing problem.

For around 10 intermediate visits, the route length was of the order of 2.5 to 3 times the length of the straight line path linking the way-points (i.e. the default flight path). Further simulation runs suggested that this path may be reduced by up to 25% in length if the classification range was extended to 20 nautical miles.
This emphasises that a large classification range can have a very beneficial effect on flight distances if it is fully exploited.

Since the plane only has to come within a distance \( r \) of the ship in order to investigate it, the plane path \( x(t), y(t) \) must intersect the circle around the ship defined by

\[
\begin{align*}
x(t) &= x_0 + wt \cos \theta + r \cos \eta \\
y(t) &= y_0 + wt \sin \theta + r \sin \eta
\end{align*}
\]

(5)

where \( \eta \) is the angle parameter defining the detection circle as shown in Figure 11.

The position of the plane is governed by three equations:

- for \( t \leq 0 \), prior to the plane beginning to turn,
  \[
  x = vt \cos \phi, \quad y = vt \sin \phi
  \]
  (6)

- for \( 0 \leq t \leq t_1 \), when the plane is turning,
  \[
  x = x_c + r_c \cos \left( \frac{vt}{r_c} + \xi_0 \right), \quad y = y_c + r_c \sin \left( \frac{vt}{r_c} + \xi_0 \right)
  \]
  (7)

where \( (x_c, y_c) \) is the centre of the turning circle, \( \xi_0 \) is the parameter angle at the start of the turning circle when \( t = 0 \);

- for \( t > t_1 \), when the plane is on a heading to the ship,
  \[
  x = x_1 + v(t - t_1) \cos \phi, \quad y = y_1 + v(t - t_1) \sin \phi.
  \]
  (8)

where \( (x_1, y_1) \) are the coordinates around the circle when the plane has finished turning.

Some of these variables are easily determined. For example, setting \( t = 0 \) in equation (8) gives the centre of the turning circle as

\[
x_c = -r_c \cos \xi_0, \quad y_c = -r_c \sin \xi_0.
\]

(9)

The angle around the turning circle when the plane begins to turn is \( \xi_0 \) and \( \xi_1 \) when it finishes its turn, so by geometry

\[
\xi_0 = \phi_i - \frac{\pi}{2}, \quad \xi_1 = \frac{3\pi}{2} + \phi.
\]

(10)

We note that these equations assume a counter-clockwise turning circle in the geometry given in Figure 11 and code can easily be implemented to account for a variety of geometries although care must be taken to program all geometries carefully. Hence, similar equations can be derived for the clockwise turning situation. The time \( t_1 \) that the plane finishes turning is

\[
t_1 = (2\pi + \phi - \phi_i) \frac{r_c}{v}
\]

(11)

at position \( (x_1, y_1) \) given by equations (7) with \( t = t_1 \). Note again that for some geometries the additional \( 2\pi \) is not necessary.

Thus equating (8) and (5) and using equations (7) and (11) gives

\[
\begin{align*}
x_0 + r \cos \eta + wt \cos \theta &= x_c + r_c \sin \phi + v(t - t_1) \cos \phi \\
y_0 + r \sin \eta + wt \sin \theta &= y_c - r_c \cos \phi + v(t - t_1) \sin \phi
\end{align*}
\]

(12) (13)

48
where our unknowns are the plane heading \( \phi \), the intersection time \( t \), and the detection circle angle \( \eta \) with \( t_1(\phi) \) given in equation (11) and \( x_c, y_c, \theta, r, r_c, v, w \) all known quantities. This system is solved for \( t(\eta), \phi(\eta) \) by squaring and adding the equations, in much the same way as equation (1), and then \( \eta \) is determined by minimising \( t(\eta) \).

Hence, the equation analogous to (2) is,
\[
(x_0 + r \cos \eta - x_c + wt \cos \theta)^2 + (y_0 + r \sin \eta - y_c + wt \sin \theta)^2 = r_c^2 + v^2(t - t_1)^2
\]
which is a quadratic in \( t \), as in equation (3), with
\[
\alpha = w^2 - v^2, \quad \beta = 2(x_0 + r \cos \eta - x_c)w \cos \theta + 2(y_0 + r \sin \eta - y_c)w \sin \theta + 2v^2 t_1
\]
\[
\gamma = (x_0 + r \cos \eta - x_c)^2 + (y_0 + r \sin \eta - y_c)^2 - r_c^2 - t_1^2 v^2.
\]
The resultant solution \( t(\phi) \) is substituted into equation (12) to give an implicit equation for \( \phi \) which must be solved numerically. Because \( \eta \) is still not determined for this most general case, a numerical solution involves iterating over all \( \eta \in [0, 2\pi] \) and solving the resultant \( \phi \) numerically for each \( \eta \) before choosing the value of \( \eta \) which minimises \( t \). Whilst this is relatively easy to program, a simpler approximate solution would be desirable.

### 6.3 Approximate solutions when \( y_0 \gg r, r_c \)

A simpler approximate solution is possible when the distances between the ship and plane are all large relative to the detection radius and the turning circle. In the limit when \( r, r_c \to 0 \) we naturally recover the simple straight intersection solution outlined earlier. Without loss of generality the problem can be rescaled with \( x_0 = 0 \) and \( y_0 = 1 \) hence allowing us to write \( r = \epsilon \) where \( \epsilon \) is a small parameter. Once the perturbation solution is found it is a simple matter to program the re-scaled solution. Thus we assume a regular perturbation system with
\[
\phi = \phi^{*0} + \epsilon \phi^{*1} + \cdots, \quad t = t^{*0} + \epsilon t^{*1} + \cdots,
\]
\[
r_c = r_c^{*}, \quad x_c = x_c^{*} \epsilon, \quad y_c = y_c^{*} \epsilon, \quad t_1 = t_1^{*} \epsilon,
\]

since we expect \( r_c, x_c, y_c, t_1 \) all to be of similar small order.

The zeroth order solution to \( t^{*0} \) and \( \phi^{*0} \) is the straight intersection solution given earlier. The \( O(\epsilon^1) \) solution is
\[
\begin{bmatrix}
w \cos \theta - v \cos \phi^{*0} \\
w \sin \theta - v \sin \phi^{*0}
\end{bmatrix}
\begin{bmatrix}
t^{*0} v \sin \phi^{*0} \\
t^{*0} v \cos \phi^{*0}
\end{bmatrix}
= \begin{bmatrix}
x_c^{*} + r_c^{*} \sin \phi^{*0} - t_1^{*} v \cos \phi^{*0} - \cos \eta \\
y_c^{*} - r_c^{*} \cos \phi^{*0} - t_1^{*} v \sin \phi^{*0} - \sin \eta
\end{bmatrix}
\]
which can be inverted to find \( t^{*1} \) and \( \phi^{*1} \). However, the value of \( \eta \) still needs to be evaluated. This can be done by minimizing \( t \) as outlined earlier, but this has to be done numerically. As we are only concerned with a simple approximation we take \( \eta \approx \pi + \phi^{*0} \) which is the case for slow moving ships.

Figure 12 shows this approximation along with the exact solution for a specific case where the ship initial position is \( (x_0, y_0) = (0.5, 1) \), ship heading is \( \theta = \pi/8 \) with speed \( w = 2 \), the plane velocity is \( v = 4 \) with turning circle \( r_c = 0.1 \), and detection radius \( r = 0.25 \), all in non-dimensional units. Initially the plane is on a heading \( \phi_i = 3\pi/2 + \pi/6 \). Even for this relatively large detection radius the perturbation solution is a good approximation.
Figure 12: Illustration of the exact and perturbation solutions for a plane intersecting a ship with parameters \((x_0, y_0) = (0.5, 1)\), \(\theta = \pi/8\), \(w = 2\), \(v = 4\), \(r_c = 0.1\), \(r = 0.25\) and \(\phi_i = 3\pi/2 + \pi/6\).

7 Estimating tour length

There is a theoretical result [19] that gives bounds on the shortest travelling salesman tour length of a standard TSP on a unit square with \(n\) cities, in the worst case the tour length is

\[
\alpha^* \sqrt{n} + o(\sqrt{n}) \text{ where } 1.075 \leq \alpha^* \leq 1.414
\]

We are motivated by the above result to generate empirical estimates for the tour length under perfect and incomplete information. Under perfect information the problem is simply to generate a tour on the set of known targets. With incomplete information, a tour is generated on the current set of detected targets and is updated each time new targets are detected.

A simulation program was written for this experiment in MATLAB. The program solves a TSP problem using a genetic algorithm. The TSP algorithm was sourced from the MATLAB Central file exchange.

The problem instances for estimating tour length all have an AI with dimension 400 \(\times\) 100 n mile. The coordinates of the two way-points are \(w_1 = (50, 50)\) and \(w_2 = (350, 50)\). These way-points represent the start and end points for a tour.

For the air platform with perfect information, the detect radius is \(\infty\) and classification radius is 0. In the incomplete information case, we have detect radius 50. The classification radius is still set at 0. The vessels in each instance are all stationary and are assigned random
Figure 13: The solid rectangle is the AI and the way-points are shown by filled circles. The large dashed circle denotes the detection radius about the platform at its current position and the vessels are denoted by triangles with arrows representing velocity vectors.

<table>
<thead>
<tr>
<th>N ships</th>
<th>Perfect Info</th>
<th>Incomplete Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_t$</td>
<td>$\sigma_t$</td>
</tr>
<tr>
<td>5</td>
<td>427.03</td>
<td>52.36</td>
</tr>
<tr>
<td>10</td>
<td>557.47</td>
<td>60.85</td>
</tr>
<tr>
<td>15</td>
<td>672.81</td>
<td>70.19</td>
</tr>
<tr>
<td>20</td>
<td>777.72</td>
<td>78.20</td>
</tr>
<tr>
<td>25</td>
<td>879.27</td>
<td>79.07</td>
</tr>
</tbody>
</table>

Table 2: Results for estimating tour length

positions in the AI according to a uniform distribution. The experiment is run with number of vessels $N = 5, 10, 15, 20, 25$. The experiments were repeated for 1000 instances of each parameter combination. The results are presented in Table 2.

The entries in Table 2 show the means and standard deviations of the tour lengths in the perfect and incomplete information cases. The final column of the table shows the average proportion of the targets that are classified. The results show that the average tour length with perfect information is greater than the tour length with incomplete information. This can be attributed to some of the targets being missed in the incomplete information case. Table 2 also shows that the number of targets detected and classified increases with the target density.

8 What proportion of ships do we detect?

To gain a better perspective on how effective a search pattern is in detecting ships we consider how many ships enter the search area versus how many are detected. For example, Figure 14 shows a plane’s circular detection region moving from left to right with a random distribution of ships moving in and out of the search region with a distribution of velocities. In the simulations shown here the plane moves left to right with speed 150 kn, the plane detects ships within a 50 n mile radius, and the ships were scattered over a 300 by 300 n mile square region.

If we take the ship velocities as $v_i(t), i = 1, \ldots, n$; the detection radius is $r$, and $p(t)$ the plane path then the proportion of ships detected is

$$q = \frac{\sum_{i=1}^{N} H \left( \int_{0}^{T} 1 - H(|v_i - p| - r) dt \right)}{\sum_{i=1}^{N} H \left( \int_{0}^{T} 1 - H(|v_i| \in R) dt \right)}$$
Figure 14: A plane moving left to right has a circular detection region within the rectangular search region. Various ships move in and out of the search region hence only a proportion of ships will be detected.

where $R$ is overall search region and $H$ is the Heaviside function, that is $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ if $x < 0$. Hence if ship $i$ enters the detection region then $|\mathbf{v}_i - \mathbf{p}| - r < 0$ so that $1 - H(|\mathbf{v}_i - \mathbf{p}| - r) = 1$ when a ship is detected. By integrating this over all times and applying a second Heaviside function means that the numerator is the number of ships that are detected. The denominator is similar with $H(|\mathbf{v}_i| \in R) = 0$ if a ship is within the search zone. Thus $q$ is a mathematical way of expressing a rather simple numerical calculation of counting the number of ships entering the space and the number that is counted.

In Figure 15 we show the proportion of ships detected, $q$, for 1000 different simulations and a histogram of this proportion. Here the ship velocities were normally distributed with mean 15 kn and variability ± 1 standard deviation. The ships were scattered in a uniform random distribution with random direction.

Figure 16 shows the mean of the proportion detected as a function of the mean ship velocity. Different ship velocity distributions are compared, with the ships having either a uniform distribution of velocities, a constant velocity or a normal distribution with two different variances. As expected when the velocity of the ships increases then less are detected – since more ships move in and out of the region without being spotted. If all the ships have velocity $\mathbf{v}_i = 0$ then all will be detected giving $q = 100\%$. If the ships have velocities with a larger standard deviation, then the proportion falls, since with more ships move out of the region before being detected.

Figure 17 shows how the proportion detected changes with the variation in the ship
Figure 15: Proportion of ships detected for 1000 simulations. The lower plot shows the histogram for all simulations.

velocities. Shown is the mean proportion and upper and lower bounds represented by the 25 and 75% levels on the detection distribution. In all cases the mean ship speed was 15 kn. As the variation in ship velocities increases, so the proportion detected decreases, as discussed in Figure 15. Not obvious however, is why the proportion detected has slope near zero for low variation and then changes to uniform slope for higher variations.

The simple simulations shown above are illustrative only, with the aim of showing the general trends in detection rate with ship velocity. However, similar simulations can be done with realistic flight paths, to estimate how effective a plane’s path is as measured by detection rate. Further work needs to be done on this to obtain mathematical results which allow prediction of these simulation results.

9 Conclusions

The group looked at a number of issues concerned with route planning within the context of maritime surveillance. The routing problem is very complex. The factors to be considered include moving targets, dynamically changing data, time windows, precedence constraints, intercept points, and demand prediction.

A simulation system was developed which went some way to answering the main questions posed by DSTO.

- For the typical scenarios faced by the RAAF, ships can be treated as stationary, but slightly better results (in terms of number of classifications and distance travelled) can
be obtained by more realistic modelling. However, modelling suggests that with only a few fast-moving ships, the detection rate can be badly affected. Further investigation is required to test this effect.

- Flying time can be improved with better algorithms. With standard TSP algorithms, these gains are modest; but we would expect that better treatment of the online aspects of the problem would yield better results.
- The computation times for all the methods tried were quiet modest (for the typical problem sizes used), but the effect needs to be tested in the context of the larger model within which the search model sits.

We also examined some additional areas of interest. In particular, the flight paths which account for turning circles and non-zero classification distances. The effect of large classification distances is particularly interesting, as large savings seem to be possible.

This study indicates that, while some questions have been answered, more questions have been raised. These suggest that further study in a number of areas would help produce better answers to the problems seen in maritime surveillance - particularly better handling of the online aspects of the problem, exploiting large classification distances, and investigating the effect of fast-moving ships.

**Acknowledgements**

Thanks very much to the industry representatives, David Marlow and Jason Looker from the Defence Science and Technology Organisation, Fishermens Bend, Victoria. This report captures the input from a large group of people who contributed during the week. The organisers and industry representatives would like to thank those who participated – including
Figure 17: Proportion of ships detected as a function of variation in ship speeds. The ships have a normal distribution of speeds with mean 15 kn and variation \( \sigma \) shown on the \( x \) axis.

References


Calibrating mean-reverting jump diffusion models: an application to the NSW electricity market

*Jamie Alcock*

*University of Queensland*

*Joanne Goard*

*University of Wollongong*

*Tony Vassallo*

*The University of Sydney*

1 Introduction

In recent years the Australian National Electricity Market (the NEM) has been undergoing the transition to a fully deregulated marketplace. In December 1998, for the first time, the wholesale price of electricity was subject to market forces. The NEM includes the Queensland, New South Wales, Victorian, South Australian, Tasmanian and Australian Capital Territory electricity markets. Almost 8 million end users are supplied by the world's longest interconnected power system, and the NEM trades up to $7 billion of electricity annually [20].

NEMMCO Ltd (The National Electricity Market Management Company) was established in 1996 to manage the NEM, a role which carries with it the responsibility for setting the spot price. The spot price is determined via a sellers’ dutch auction. Each day, each generator submits a complex bid of prices and volumes. The demand fluctuates throughout the day, and every 5 minutes short-term supply and demand are realigned by NEMMCO. The 5 minute dispatch price for all bidders is set to the winning bid of the marginal supplier, and six sequential dispatch prices are averaged to determine the half-hour spot price. All successfully bidding generators receive the spot price for their product. Currently spot prices are artificially bound by NEMMCO to remain between −$1000 and $10,000 per MWhr.

The introduction of market forces to the NEM has provided a plethora of challenges for mathematicians, economists and financial economists\(^1\) amongst others. As electricity is not storable the spot price process is extremely volatile. However the trend underlying the spot price process is highly predictable and highly periodic. The supervolatility of spot price stems from a combination of unexpected events. Unexpectedly high demand (possibly because of an unexpected temperature change) or unexpectedly low supply (because of an unscheduled generator outage or distributional failures /constraints) results in a rapid change in price. There is a strong positive relationship between the spot price spikes and large changes in demand (load).

Retail customers typically pay a fixed price for electricity. Hence the electricity retailer manages a portfolio of floating-for-fixed swap instruments. The retailer receives a fixed pay-

\(^1\)For example see recent and significant works by [5, 7, 21]
Figure 1: Time series of NSW Electricity spot prices. Figure (a) provides a 5-year sample of spot prices; (b) provides a daily sample series.

Management for electricity and pays a floating rate to the generator. Managing financial risk in such a volatile market is a formidable challenge. The magnitude of this challenge is highlighted when one considers that the retailers financial risk is the product of load and spot price. Insofar as there is a positive correlation between load and price, high demand implies a doubly heavy loss for the electricity retailer. Understanding the dynamics of the electricity spot-price/demand process is a fundamental step in developing adequate risk-management strategies.

Much of the existing literature on modelling commodity prices deals with the assumption of geometric Brownian motion (GBM) and methods by which the assumption of Gaussian returns can be improved. As such, many papers deal with stochastic volatility [14, 15] and fat tail distributions [11] and related issues. With the emergence of electricity as a traded commodity in the 1990’s came the challenge to model the spot price behaviour. As electricity is largely non-storable, market supply and demand must be kept in constant, continuous balance. This results in electricity prices not following a ‘smooth’ process as other commodities do. As a result, a GBM alone cannot successfully describe the evolution of its price.

Clearly an equilibrium spot price model should include information on supply, or potential supply, in addition to demand. A rapid shift in available supply, via an unscheduled outage, is likely to have a similar impact on price as a rapid shift in demand. However data availability remains an issue. While load data is readily available, potential supply data is more difficult to obtain.

The aim of this article is to develop and calibrate a model for the spot price of electricity as a function of market demand (load), and to calibrate this model using historical load and spot prices from the NSW market. In particular, this model will exploit two, quite apparent, features of historical spot price data: (i) mean-reversion, and (ii) sudden price spikes, or jumps. Hence we shall explore the suitability of fitting a mean-reversion jump-diffusion model to NSW spot price and load data. The justification for, and development of, these models is given in Section 2. The exposition of a calibration methodology is given in Section 3.
model is fitted to the given data using this calibration technology in Section 4. We conclude in Section 5.

2 Model development

Two notable characteristics of electricity spot prices are mean-reversion and price-spikes (see Figure 1 for a sample time series of the NSW spot price). Electricity prices can also be highly volatile, with warmer seasons exhibiting significantly higher volatility than the colder seasons. This indicates a higher stability of the mean price for cold seasons as compared to the warmer seasons. In order to capture the characteristics of the electricity spot prices, [16] indicates the need to introduce jumps and stochastic volatility into their models. In [3] tested the effectiveness of Brownian motion (BM), Mean Reversion, GBM and Geometric Mean Reversion (GMR) and found that the GMR Model outperformed the others, and that including jumps into the models further improved their performance. Other papers related to electricity pricing include those by [4, 9, 10, 17]. As noted by many of these authors, general diffusion models of the form

$$dS_t = a(S_t, t)dt + b(S_t, t)dZ_t,$$

where $dZ_t \sim N(0, dt)$ is an increment in a Wiener process, fail to capture the large, non-negligible observed spikes in the market.

An obvious extension to the models is to include a jump component modelled by a Poisson distribution. This was first introduced by [18] and then extended by [7, 8] to incorporate the mean-reversion feature. An example of a simple mean reversion jump diffusion model (MRJD) is

$$dS_t = \alpha(S^* - \ln S_t)S_t dt + S_t\sigma dZ_t + S_tK dq_t,$$

where $\alpha$ is the mean-reversion rate, $S^*$ is the mean-reversion level, $\sigma$ the volatility of the spot price, $K$ is the jump size which may, for example, be taken to follow a normal distribution $N(\mu, \nu^2)$ and $dq_t$ is a Poisson process so that

$$dq_t = 1 \text{ with probability } \lambda dt, \quad = 0 \text{ with probability } 1 - \lambda dt,$$

(where $\lambda$ is the average number of jumps per year). Many variations of (2) exist that might include multiple jump processes, doubly-stochastic jump processes and stochastic volatility (see e.g. [9]). In this article we examine two classes of MRJD and within both classes examine the performance with various forms of volatility functions.

Model 1

As mentioned, spot price is a function of both demand (state load) and available supply. To partially verify this a scatter plot of spot price vs state load is given in Figure 2 (b). Clearly a monotonic relationship between spot price and demand exists. To capture this relationship, we assume that the log price $Q_t(= \ln p_t)$, where $p_t$ is the current spot price of electricity, depends on the load via

$$Q_t = f(L_t) + X_t,$$

where $dX_t = -\alpha X_t + \sigma dZ_t + K dq_t$

and $L_t$ is the load at time $t$. 

59
From (9), \( X_t \) follows a mean-reverting process (to a mean of zero), where \( \alpha \) is the reversion rate, \( K \) is the jump size, \( dq_t \) is as in (3) and \( dZ_t \sim N(0, dt) \) denotes throughout this paper an increment in a Wiener process. We made no prior assumptions about the distribution of the jump size \( K \).

The dependence of \( Q_t \) on load in (4), will be reflected in the stochastic process followed by \( Q_t \). The following forms for the volatility \( \sigma_t \) will be considered:

\[
\begin{align*}
\sigma_t &= \sigma_0, \\
\sigma_t &= \sigma_0 L_t, \\
\sigma_t &= \sigma_0 (dL_t), \\
\sigma_t &= \sigma_0 e^{Q_t^2}, \\
\sigma_t &= \sigma_0 e^{nQ_t}, \\
\sigma_t &= \sigma_0 + \sigma_1 \cos(2\pi t) + \sigma_2 \cos(4\pi t),
\end{align*}
\]

where \( \sigma_0, \sigma_1, \sigma_2 \) are constants and \( dL_t \) denotes the change in the load, \( L_t \), in one time-step.

In volatility model (6a) we assume a constant volatility throughout the period, while in volatility models (6b) and (6c) we assume that volatility depends on the load or the change in the load respectively. As will be seen shortly, models (6d) and (6e) imply a volatility for the stochastic differential equation (SDE) for \( p_t \) of the form \( \sigma_p^{3/2} \) and \( \sigma_p^{n+1} \) respectively. These were chosen for comparison with mean-reversion models of the type \( dr = (\alpha + \beta r)dt + \sigma r^\gamma dZ \), for which \( \gamma = 1.5 \) is often the unconstrained estimate [6]. A similar dependence could possibly occur with electricity prices. The volatility model (6f) was chosen in an effort to capture the yearly and half-yearly cycles in volatility.
From (4) and (5) the process followed by $Q_t$ is
\[
dQ_t = \alpha \left\{ \frac{1}{\alpha} f'(L_t) L'_t + f(L_t) - Q_t \right\} dt + \sigma_t\, dZ_t + K\, dq_t.
\] (7)

Equation (7) represents Model 1. From it we can also derive the process followed by $p_t$, namely
\[
dp_t = \alpha p_t \left\{ \frac{1}{\alpha} f'(L_t) L'_t + \frac{\sigma_t^2}{2\alpha} + f(L_t) - \ln p_t \right\} dt + \sigma_t p_t\, dZ_t + p_t(e^K - 1)\, dq_t.
\] (8)

Note then that the volatility terms (6d) and (6e) imply the volatility forms $\sigma_0 p_t^{3/2} dZ$ and $\sigma_0 p_t^{n+1} dZ$ in (8) respectively.

Model 2

For Model 2 we simply assume that $Q_t$ follows a seasonal pattern in time with
\[
Q_t = \ln p_t = h(t) + X_t
\] (9)

where $dX_t$ follows the process as in (5) with the volatility as in (6a) - (6f). This then implies the process for $Q_t$ as
\[
dQ_t = \alpha \left\{ \frac{1}{\alpha} h'(t) + h(t) - Q_t \right\} dt + \sigma_t\, dZ_t + K\, dq_t
\] (10)

which we call our Model 2. From (10) the process followed by the price $p_t$ itself is
\[
dp_t = \alpha p_t \left\{ \frac{1}{\alpha} h'(t) + \frac{\sigma_t^2}{2\alpha} + h(t) - \ln p_t \right\} dt + \sigma_t p_t\, dZ_t + (e^K - 1)p_t\, dq_t.
\] (11)

3 Data

Integral Energy Australia provided half-hourly data for the electricity spot prices and load for the years 2002-2006. We generated a new time series for the price by calculating the arithmetic average of the 48 data values for each day and then taking logs of those values. This is done rather than taking the average of the 48 log prices for each day (which would have been equivalent to a geometric average) since we needed to analyse the jumps in the data rather than attempt to mask them. This series is plotted in Figure 3. The daily series is initially formed in order to test the methodology on a smaller dataset.

At first glance the price spikes in Figure 3 are prominent and seem relatively frequent. Statistics for the $Q_t = \ln p_t$ series as well as for the return $d\ln p_t = \ln p_t - \ln p_{t-1}$ are provided with the figure. Notice in particular the high standard deviations in the two series and the very high kurtosis coefficients. The kurtosis estimates are significantly different from 3 (the kurtosis for a normal distribution) under the null hypothesis of normality and imply leptokurtic distributions with high peaks, thin midrange and fat tails. This is supported by the histogram of the returns in Figure 4.

In Figure 5, the average daily load is plotted from which a yearly seasonal pattern is obvious as well as an increasing trend. Statistics for the average daily load and for $d\, \text{load}_t = \text{load}_t - \text{load}_{t-1}$ are provided with the figure.
Jumps

The discontinuous jump component in (7) and in (10) controls the nature of the jumps - their frequency and size. In order to estimate the parameters for the jumps (see Section 4.4) we first extract them from the data.

There are various approaches by which this can be achieved depending on what is defined as a ‘jump’. For example price returns beyond three standard deviations of the mean could be classified as a jump. We choose a particular threshold jump size $j_s$ and on finding a return $\ln p_{t+1} - \ln p_t$ larger than $j_s$, remove that return and subsequent ones until $\ln p_{t+k} - \ln p_t < j_s$. The jump size $j_s$ is chosen so that upon filtering in such a way, the remainder is a series that can be approximated by a normal distribution. A higher $j_s$ is preferable, as it would imply fewer ‘random’ jumps.

From an examination of percentage returns within 1, 2 and 3 standard deviations of their mean in the filtered data we chose $j_s = 0.7$ and $j_s = 0.6$ (see Table 1). Kurtosis estimates also given in Table 1, are higher than 3, but are significant improvements on the kurtosis estimate of the unfiltered data namely 25.57. We note that a double exponential distribution or Levy process would seem a better choice to model the filtered data. However, here our aim is to model the data using a MRJD model where a jump component is added to a GBM.

Using $j_s = 0.7$ and 0.6 respectively, the filtered data series and the histogram of their returns are plotted in Figures 6 and 7.
4 Calibration

We estimate the functions and the parameters in Models 1 and 2 (equations (7) and (10)) respectively. We do this by first estimating the values in the ‘non-jump’ part of the process (i.e. the diffusion process) using the filtered data, and then analysing the jump parameters from the jump data set that was extracted from the raw data.

4.1 First approximations for the seasonal function and mean reversion rate

From a plot of the log price versus the load (see Figure 8), a linear relationship is indicated so that we take \( f(L_t) = a_0 + a_1 L_t \) in equation (4).

The least squares fit approximations of \( a_0 = 1.7565 \) and \( a_1 = 0.000171 \) for \( js = 0.7 \) and \( a_0 = 1.7946, a_1 = 0.000166 \) for \( js = 0.6 \) are used as first approximations for the final fit.

For Model 2, we approximate \( h(t) \) with a truncated Fourier series

\[
h(t) = \sum_{n=0}^{12} (b_n \sin n \pi t + c_n \cos n \pi t)
\]

and again find least squares estimates for the coefficients. These are listed in Table 2. We
can also estimate the mean reversion rate \( \alpha \) in the filtered data using a least squares fit of \((X_{t+1} - X_t, X_t)\) data. For Model 1 we get \( \alpha \approx 75.58 \) for \( js = 0.7 \) and \( \alpha \approx 69.9 \) for \( js = 0.6 \). For Model 2, \( \alpha \approx 119.7 \) for \( js = 0.7 \) and \( \alpha \approx 115.34 \) for \( js = 0.6 \).

4.2 The estimation technique for the diffusion process

Paper [13] Generalised Method of Moments (GMM) is used to estimate the parameters in the diffusion processes corresponding to Models 1 and 2. The GMM approach has the advantage that it does not require that the distribution of log price changes be normal. As well the GMM estimators and their standard errors are consistent even if their disturbances \( \epsilon_{t+1} \) are conditionally heteroscedastic. GMM is used by [6, 1, 12], amongst others, to empirically test interest rate models.

We use a discrete time specification. For Model 1:

\[
Q_{t+1} - Q_t = \alpha \left( \frac{1}{\alpha} a_1 \left( \frac{L_t}{dt} - L_{t+1} \right) \right) + (a_0 + a_1 L_t - Q_t) \ dt_{t+1} + \epsilon_{t+1}
\]

\( E(\epsilon_{t+1}) = 0, \quad E(\epsilon_{t+1}^2) = \sigma_t^2 dt \),

where we use a backward difference (BD) approximation rather than a forward difference (FD) approximation for the derivative \( L'(t) \), as the load in the next time-step is assumed not to be known.

For Model 2:

\[
Q_{t+1} - Q_t = \alpha \left[ \frac{1}{\alpha} \sum_{n=1}^{12} (b_n n \pi \cos n \pi t - c_n n \pi \sin n \pi t) + \sum_{n=0}^{12} (b_n \sin n \pi t + c_n \cos n \pi t) - Q_t \right] \ dt_{t+1} + \epsilon_{t+1}
\]
Figure 6: Using $j_s = 0.7$:  
(a) The filtered $\ln p$ series;  
(b) Histogram of filtered returns.
Figure 7: Using $j_s = 0.6$:  (a) The filtered $\ln p$ series;  (b) Histogram of filtered returns.
Figure 8: Log price vs. load.

\[ E(\epsilon_{t+1}) = 0, \quad E(\epsilon_{t+1}^2) = \sigma_t^2 dt. \]  

(16)

For a given model we let \( \theta_1 \) and \( \theta_2 \) be the parameter vectors for the drift and diffusion components respectively. For example for Model 1, \( \theta_1 = (\alpha, a_0, a_1)^T \), and for volatility (6f), \( \theta_2 = (\sigma_0, \sigma_1, \sigma_2)^T \). Define the vectors

\[ (f_1)_t(\theta_1) = \left( \epsilon_{t+1} \otimes [I_1, \ldots, I_{m_1}]^T \right) \]  

(17)

and

\[ (f_2)_t(\theta_2) = \left( (\epsilon_{t+1}^2 - \sigma^2 dt) \otimes [\bar{I}_1, \ldots, \bar{I}_{m_2}]^T \right) \]  

(18)

where \( I_i \) and \( \bar{I}_i, i = 1 \ldots M \) denote instrumental variables (i.e. predetermined variables that are independent of the errors) and \( M = m_1 \) or \( m_2 \) denote the number of parameters in the corresponding parameter vector.

Under the null hypothesis that (13) and (14) or ((15) and (16)) are true, the orthogonality conditions \( E(f_i(\theta)) = 0 \) hold where \( f = f_1 \) or \( f_2 \). The GMM technique replaces \( E(f_i(\theta)) \) with its sample counterpart \( g_T(\theta) \), using \( T \) observations where

\[ g_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} f_t(\theta), \]

and then estimates the parameters in the vector \( \theta \) that minimises the quadratic form

\[ J_T(\theta) = g_T(\theta)^T W g_T(\theta), \]

where \( W \) is a positive definite, symmetric, weighting matrix with the sample estimate adjusted for serial correlation and heteroscedasticity using the method of [19] with Bartlett weights.
Table 2: Estimates of Fourier series coefficients in equation (12).

Paper [13] shows that setting

\[ W = \left( E(f_t(\theta)f_t(\theta)^T) \right)^{-1}, \]

delivers an estimate for the vector \( \theta \) with the smallest asymptotic covariance matrix for the GMM estimates of \( \theta \). With the number of unknowns exactly equal to the number of orthogonality conditions, the model is exactly identified and so \( J_T(\theta) = 0 \).

4.3 Results

Results for Model 1 using \( js = 0.7 \) are listed in Table 3. For Model 2 using \( js = 0.7 \) the Fourier coefficients in Table 2 are used as no significant improvements in SSE were found with GMM. The other parameters estimated are listed in Table 4. Tables 5 and 6 list the results for Models 1 and 2 respectively when the data is filtered using \( js = 0.6 \).

For both threshold jump sizes \( js = 0.6 \) and 0.7, Model 2 outperforms Model 1. The performance of the volatility is also consistent. From an examination of the SSE (sums of squares of errors) and the MSE (mean square errors), the volatility \( \sigma_t = \sigma_0 e^{\alpha t} \) performs best in all models, followed by \( \sigma_t = \sigma_0 e^{\alpha t/2} \) then \( \sigma_t = \sigma_0 + \sigma_1 \cos 2\pi t + \sigma_2 \cos 4\pi t \). Volatility with a dependence on load does not perform significantly better than a constant volatility \( \sigma_t = \sigma_0 \). The volatility with a dependence on the change in the load performs least well, perhaps not surprisingly as the change in load related to the previous time-step.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>$t$ Value</th>
<th>$p$ Value</th>
<th>SSE (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>67.51146</td>
<td>10.28</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$a_0$</td>
<td>3.593127</td>
<td>24.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$a_1$</td>
<td>-0.00005</td>
<td>-2.73</td>
<td>0.064</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0$</td>
<td>$\sigma_0$</td>
<td>3.791098</td>
<td>31.28</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 L_t$</td>
<td>$\sigma_0$</td>
<td>0.000445</td>
<td>31.28</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 dL_t$</td>
<td>$\sigma_0$</td>
<td>0.003843</td>
<td>4.08</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{Q_t/2}$</td>
<td>$\sigma_0$</td>
<td>0.745511</td>
<td>30.23</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{RQ_t}$</td>
<td>$\sigma_0$</td>
<td>0.376201</td>
<td>3.88</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{RQ_t} n$</td>
<td>$\sigma_0$, $n$</td>
<td>0.696733</td>
<td>8.89</td>
</tr>
<tr>
<td>$\sigma_t = c_0$</td>
<td>$\sigma_0$</td>
<td>3.672192</td>
<td>34.15</td>
</tr>
<tr>
<td>$\sigma_t = c_0 2\pi t$</td>
<td>$\sigma_1$</td>
<td>0.183404</td>
<td>1.11</td>
</tr>
<tr>
<td>$\sigma_t = c_0 4\pi t$</td>
<td>$\sigma_2$</td>
<td>0.669647</td>
<td>4.05</td>
</tr>
</tbody>
</table>

Table 3: Parameter estimates for Model 1 with $js = 0.7$.

In Model 1 we notice that $a_1$, the coefficient of $L$ in the function $f(L_t)$ is negative for both thresholds. As the log price is an increasing function of the load, this indicates that the BD approximation for the derivative $L'(t)$ is not a good estimate for the slope. Re-estimating the parameters for Model 1 using a FD approximation for $L'(t)$ using volatilities (6a), (6d) and (6e) yield the values in Tables 7 and 8.

The estimates for $a_1$ are positive. As well, comparing SSE in Tables 4 and 7 and again in Tables 6 and 8, Model 1 outperforms Model 2. However, for the purpose of forecasting, a FD approximation could not be used as it is anticipatory and assumes knowledge of load in the next time-step.

As an alternative measure to gauge relative performance of the models, we test their forecast powers for (filtered) price changes and squared (filtered) price changes using the coefficient of determination $R^2$ metric. The $R^2$ metrics of the drift and diffusion measure the proportion of the total variation in the ex-post price changes ($R^2_t$) and squared price changes ($R^2_c$) that can be explained by the conditional expected price changes and conditional volatility measures. The metric provides information about how well each model is able to forecast the future level and volatility of the filtered price. The order of performance in the $R^2$ statistics from best to worst is naturally the same as the order of performance of the models by examination of the SSE. Model 1 using the FD approximation for the derivative $L'(t)$ provides the best measures with values of $R^2_t = 21.36\%$ and $31.5\%$ and $R^2_c = 44.2\%$ and $47\%$ for $js = 0.7$ and 0.6 respectively. The forecasting ability of this SDE is very good compared to other stochastic models.

The best values for Model 2 were $R^2_t = 15.21\%$ with $js = 0.7$ and $R^2_t = 15.1\%$ with $js = 0.6$ and $R^2_c = 35.7\%$ and $33\%$ with $js = 0.6$ and $js = 0.7$ respectively and volatility
Table 4: Parameter estimates for Model 2 with $js = 0.7$.

(6c). Model 1 with a BD approximation for $L'(t)$ does not perform as well.

4.4 Estimating the jump parameters

As described in Section 2, the jumps are filtered from the data using a threshold jump size of $js = 0.7$ and also using $js = 0.6$. From the extracted jump data we estimate the parameters in the jump component in Models 1 and 2 (equations (7) and (10) respectively).

We assume that the jumps were instantaneous to the first approximation. In effect we collapse jumps that may have had a duration of 2 days (which occurred 34.78% of the time) into an instantaneous jump on a single day. Hence in the filtering process, when a return is found above the threshold, the jump size is then taken as the sum of this return and if applicable, the subsequent positive return after this. This is necessary for the model to attain extreme jumps. The mean-reversion characteristic of the models allow the prices to return quickly to the pre-jump level without the need to impose any conditions on the jumps.

Statistics for the jumps are provided in Table 9 and their histograms given in Figure 9.

With the distribution of the jumps difficult to determine with so few data points, and the frequency of jumps inconsistent across the months, we use for the term $K dq_j$ in (7) and (10)

$$\sum_{j=1}^{12} K_j \ dq_j \ M_{ij}$$

where $M_{ij} = 1$ if the date $t$ belongs to the $j$th calendar month and 0 otherwise, while $dq_j = 1$ with probability $\lambda_j$, and $dq_j = 0$ with probability $1 - \lambda_j$. Each $K_j$ is a random number from a discrete set $\psi_j$ of equally possible elements. These are listed in Table 10.

We investigate the relationship between jumps and load. For the 5 years of data provided, jumps occurred when the load in the previous time period was above the average load for
<table>
<thead>
<tr>
<th>parameter</th>
<th>t value</th>
<th>p value</th>
<th>SSE (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eq (1)</td>
<td>Eq (2)</td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>64.88192</td>
<td>10.38</td>
<td>58.1593 (0.0338)</td>
</tr>
<tr>
<td>$a_0$</td>
<td>3.541309</td>
<td>26.61</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$a_1$</td>
<td>-0.00004</td>
<td>-2.62</td>
<td>.0089</td>
</tr>
</tbody>
</table>

$\sigma_t = \sigma_0$

$\sigma_t = \sigma_0 L_t$

$\sigma_t = \sigma_0 dL_t$

$\sigma_t = \sigma_0 e^{Q_t/2}$

$\sigma_t = \sigma_0 e^{n Q_t}$

$\sigma_t = \sigma_0 + \sigma_1 \cos 2\pi t + \sigma_2 \cos 4\pi t$

| $\sigma_0$ | 3.477373 | 34.96   | <.0001    |
| $\sigma_0$ | 0.000408 | 35.22   | <.0001    |
| $\sigma_0$ | 0.002767 | 2.73    | .0064     |
| $\sigma_0$ | 0.683717 | 33.52   | <.0001    |
| $\sigma_0$ | 0.400374 | 3.78    | .0002     |
| $\sigma_0$ | 0.656174 | 8.03    | <.0001    |
| $\sigma_0$ | 3.38278  | 39.48   | <.0001    |
| $\sigma_0$ | 0.172996 | 1.19    | 0.2337    |
| $\sigma_0$ | 0.623552 | 4.41    | <.0001    |

Table 5: Parameter estimates for Model 1 with $j_s = 0.6$.

<table>
<thead>
<tr>
<th>parameter</th>
<th>t value</th>
<th>p value</th>
<th>SSE (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eq (1)</td>
<td>Eq (2)</td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>87.81233</td>
<td>0.88</td>
<td>52.9991 (0.0307)</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0$</td>
<td>3.264612</td>
<td>36.04</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 L_t$</td>
<td>0.000384</td>
<td>37.01</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 dL_t$</td>
<td>0.00274</td>
<td>3.33</td>
<td>0.0009</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{Q_t/2}$</td>
<td>0.644729</td>
<td>38.67</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{n Q_t}$</td>
<td>0.373755</td>
<td>4.16</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 + \sigma_1 \cos 2\pi t + \sigma_2 \cos 4\pi t$</td>
<td>3.226232</td>
<td>39.18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{n Q_t}$</td>
<td>0.662956</td>
<td>8.9</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Table 6: Parameter estimates for Model 2 with $j_s = 0.6$. 71
<table>
<thead>
<tr>
<th>parameter</th>
<th>parameter value</th>
<th>t value</th>
<th>p value</th>
<th>SSE (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>67.68492</td>
<td>11.52</td>
<td>&lt;0.0001</td>
<td>57.075</td>
</tr>
<tr>
<td>$a_0$</td>
<td>0.223646</td>
<td>0.26</td>
<td>0.7963</td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>0.000353</td>
<td>3.46</td>
<td>0.0006</td>
<td></td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0$</td>
<td>$\sigma_0$</td>
<td>3.380626</td>
<td>34.27</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{Q_t/2}$</td>
<td>$\sigma_0$</td>
<td>0.683908</td>
<td>27.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{nQ_t}$</td>
<td>$\sigma_0$</td>
<td>0.304197</td>
<td>3.28</td>
<td>0.0011</td>
</tr>
<tr>
<td>n</td>
<td>0.732653</td>
<td>7.87</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Parameter estimates for Model 1 using $js = 0.7$ and FD for $L'(t)$.

<table>
<thead>
<tr>
<th>parameter</th>
<th>parameter value</th>
<th>t value</th>
<th>p value</th>
<th>SSE (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>62.50735</td>
<td>12.06</td>
<td>&lt;0.0001</td>
<td>42.7666 (.00248)</td>
</tr>
<tr>
<td>$a_0$</td>
<td>0.935759</td>
<td>1.55</td>
<td>.1210</td>
<td></td>
</tr>
<tr>
<td>$a_1$</td>
<td>0.000268</td>
<td>3.77</td>
<td>.0002</td>
<td></td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0$</td>
<td>$\sigma_0$</td>
<td>2.925856</td>
<td>34.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{Q_t/2}$</td>
<td>$\sigma_0$</td>
<td>0.580077</td>
<td>35.91</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_t = \sigma_0 e^{nQ_t}$</td>
<td>$\sigma_0$</td>
<td>0.288463</td>
<td>3.31</td>
<td>0.001</td>
</tr>
<tr>
<td>n</td>
<td>0.708476</td>
<td>7.62</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Parameter estimates for Model 1 using $js = 0.6$ and FD for $L'(t)$.
Figure 9: Histogram of jumps using (a) $js = 0.7$ and (b) $js = 0.6$. 
<table>
<thead>
<tr>
<th></th>
<th>$js = 0.7$</th>
<th>$js = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency of jumps</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>mean of jumps</td>
<td>1.8621</td>
<td>1.7255</td>
</tr>
<tr>
<td>std of jumps</td>
<td>0.8818</td>
<td>0.9418</td>
</tr>
<tr>
<td>max jump</td>
<td>4.2484</td>
<td>4.2484</td>
</tr>
<tr>
<td>min jump</td>
<td>0.7074</td>
<td>0.6015</td>
</tr>
</tbody>
</table>

Table 9: Statistics for ‘jumps’.

that month in the year. This occurred 9 out of 12 times in the first year, 4 out of 7 times in the second year, 6 out of 9 times in the third year, 7 out of 11 times in the fourth year and 3 out of 7 times in the fifth year. 85% of jumps occurred when the current load was above the average load for its month/year and all jumps occurred when either the load in the previous time period or the load in the current time period to the jump were above the average load for the month/year when the jump occurs.

This information is only be useful for predicting jumps if we can predict the monthly averages for the loads and the load for the next time period. Regression estimates for the average monthly loads can be used to predict future monthly averages. Using the five year data provided, approximate estimates are

$$ y_{\text{jan}} = 7.1029 + 1.0271t - 0.3343t^2 + 0.0392t^3 $$
$$ y_{\text{nov}} = 8.0474 - 0.1060t + 0.047t^2 $$

where $t$ represents the time (in years). With more data these estimates could be improved.

4.5 Simulations

Simulations of Model 2 with volatilities of the form (6a) and (6e) and using parameter values as in Tables 4 and 6 are provided in Figures 10 – 13 with jump size thresholds $js = 0.7$ and 0.6 respectively.

As the volatility (6e) depends on the price level, in the simulation, if a jump occurred at one time, $t$, then $dZ_t$, the random term in the volatility in the next time step $t + dt$ is set to its expected value of 0. Also, jumps were not allowed if the price was above 4.2. From a comparison of the trajectories and histogram of returns in Figures 10 and 12 with the
<table>
<thead>
<tr>
<th>Calendar Month</th>
<th>$j^s = 0.7$</th>
<th>$j^s = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\lambda_1 = 6/155$</td>
<td>$\lambda_1 = 7/155$</td>
</tr>
<tr>
<td></td>
<td>$\psi_1 = {0.7, 1, 1, 1, 1, 1, 1, 7, 2.5, 2.9}$</td>
<td>$\psi_1 = {0.72, 1, 1, 1, 1, 1, 7, 2.5, 2.9}$</td>
</tr>
<tr>
<td>2</td>
<td>$\lambda_2 = 6/141$</td>
<td>$\lambda_2 = 7/141$</td>
</tr>
<tr>
<td></td>
<td>$\psi_2 = {1, 1, 1.5, 1, 6, 2.5, 4.1}$</td>
<td>$\psi_2 = {0.62, 1.1, 1.5, 1, 6, 2.5, 4.1}$</td>
</tr>
<tr>
<td>3</td>
<td>$\lambda_3 = 1/155$</td>
<td>$\lambda_3 = 1/155$</td>
</tr>
<tr>
<td></td>
<td>$\psi_3 = {4.25}$</td>
<td>$\psi_3 = {4.25}$</td>
</tr>
<tr>
<td>4</td>
<td>$\lambda_4 = 1/150$</td>
<td>$\lambda_4 = 1/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_4 = {0.71}$</td>
<td>$\psi_4 = {0.71}$</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_5 = 5/155$</td>
<td>$\lambda_5 = 5/155$</td>
</tr>
<tr>
<td></td>
<td>$\psi_5 = {1.1, 1.3, 1.6, 2.4, 2.3}$</td>
<td>$\psi_5 = {1.1, 1.3, 1.6, 2.4, 2.3}$</td>
</tr>
<tr>
<td>6</td>
<td>$\lambda_6 = 6/150$</td>
<td>$\lambda_6 = 6/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_6 = {1.4, 1.5, 1.5, 1.6, 1.8, 2.4}$</td>
<td>$\psi_6 = {1.4, 1.5, 1.5, 1.6, 1.8, 2.4}$</td>
</tr>
<tr>
<td>7</td>
<td>$\lambda_7 = 6/155$</td>
<td>$\lambda_7 = 6/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_7 = {1.3, 1.77, 2.2, 2.2, 2.2}$</td>
<td>$\psi_7 = {1.77, 1.9, 2, 2, 2, 2}$</td>
</tr>
<tr>
<td>8</td>
<td>$\lambda_8 = 1/155$</td>
<td>$\lambda_8 = 2/155$</td>
</tr>
<tr>
<td></td>
<td>$\psi_8 = {0.93}$</td>
<td>$\psi_8 = {0.66, 0.93}$</td>
</tr>
<tr>
<td>9</td>
<td>$\lambda_9 = 2/150$</td>
<td>$\lambda_9 = 4/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_9 = {1.3, 1.7}$</td>
<td>$\psi_9 = {0.66, 0.69, 1.3, 1.7}$</td>
</tr>
<tr>
<td>10</td>
<td>$\lambda_{10} = 2/155$</td>
<td>$\lambda_{10} = 2/155$</td>
</tr>
<tr>
<td></td>
<td>$\psi_{10} = {3.2, 3.7}$</td>
<td>$\psi_{10} = {3.7, 3.8}$</td>
</tr>
<tr>
<td>11</td>
<td>$\lambda_{11} = 6/150$</td>
<td>$\lambda_{11} = 7/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_{11} = {1, 1.5, 1.8, 2.5, 3.5}$</td>
<td>$\psi_{11} = {0.69, 1.5, 1.8, 2.5, 3.5}$</td>
</tr>
<tr>
<td>12</td>
<td>$\lambda_{12} = 4/150$</td>
<td>$\lambda_{12} = 5/150$</td>
</tr>
<tr>
<td></td>
<td>$\psi_{12} = {0.81, 1, 1.3, 1.3}$</td>
<td>$\psi_{12} = {0.6, 0.81, 1, 1.3, 1.3}$</td>
</tr>
</tbody>
</table>

Table 10: Data values for the jump component $k dq$. 
trajectory in Figures 3 and 4, our models seem to capture the critical features observed in the spot electricity market, and reproduce the price levels observed in the data. Further, the volatility of the form $\sigma_t = \sigma_0 e^{\mu Q_t}$ in particular, matches the observed magnitude of price fluctuations.

Acknowledgements

Thanks to Philip Broadbridge, Pam Davy, David Griffiths and an anonymous referee for their helpful comments and suggestions.
Figure 10: Simulated time series of prices and corresponding histograms of returns, simulated using Model 2 and parameter estimates in Table 4 when $\sigma_t = \sigma_0$. 

77
Figure 11: Simulated time series of prices and corresponding histograms of returns, simulated using Model 2 and parameter estimates in Table 4 when $\sigma_t = \sigma_0 e^{rQ_t}$. 
Figure 12: Simulated time series of prices and corresponding histograms of returns, simulated using Model 2 and parameter estimates in Table 6 when (a) $\sigma_t = \sigma_0$. 
Figure 13: Simulated time series of prices and corresponding histograms of returns, simulated using Model 2 and parameter estimates in Table 6 when $\sigma_t = \sigma_0 e^{nQ_t}$. 
References


A jump diffusion model for spot electricity prices

Ramaprasad Bhar
The University of New South Wales

1 Introduction

At the outset, to appreciate the background of this project, we quote [3]:

"Over the last 10 years, major countries have been experiencing deregulation in generation and supply activities. One of the important consequences of this restructuring is that prices now determined according to the fundamental rule of supply and demand: there is a 'market pool' in which bids are placed by generators to sell electricity for the next day are compared to purchase orders".

Prior to that, the regulators used to set the price based on the cost of generation, transmission and distribution and the price to the consumer was essentially fixed for long period of time. A large fraction of the literature on electricity today belongs to the economics of deregulated electricity market from the perspective of the regulators. In the market mechanism now operating today, the price will be determined by the interaction of the purchase orders placed by the retailers against the pool prices.

The deregulation of the electricity market has also led to increased trading activities in both spot and related derivatives like forwards and options. The risk of spot-price has forced retailers to manage the risk of the spot price through various hedging mechanisms. Many retailers provide incentives to the consumers to enter into long term contract with predetermined price structure, but that still leaves the risk of buying price. It is in this context modelling the stochastic behaviour of the spot price of electricity has become important.

One feature of the electricity market that is unique to this commodity is that electricity is not storable, although, it may be argued that the concept of storability applies to hydro electricity generation. Since, in general, it cannot be stored the spot price is likely to be determined by the spot concerns, e.g., spot demand and supply constraints. The ability to store any commodity has the effect of smoothing the evolution of the spot price to some extent. As a result of its absence, price spikes are a regular feature of the electricity spot prices in most countries that have deregulated this market. Price spikes are possibly due to disruption in transmission, unscheduled outages, extreme weather changes or a combination of all these events. Additional details about the characteristics of this market may be found in [3].

We will now review some of the salient characteristics of the electricity prices in the deregulated market. In standard commodity-futures markets the concept of convenience yield plays a key role in the relationship between the spot and the forward prices. The convenience yield is a way of expressing the fact that an investor is sure of available supply when the demand for using that commodity arises at a future date. The non-storability of electricity makes the concept of convenience yield difficult to apply. This implies that the spot price itself should contain all the characteristics of the price process that would be necessary to impute prices of derivatives contracts written on electricity prices.
Next, we outline the important temporal characteristics of spot electricity prices observed in most markets. A detailed description of these characteristics may be found in [3]. Mean reversion is an important feature of spot electricity prices. The prices tend to fluctuate around values determined by cost of production and the level of demand. The mean reversion level may be constant or periodic with a trend. Seasonality is another obvious characteristic. The prices change by time of day, week, month and the year in response to cyclical fluctuations in demand. Another feature already mentioned before is that of price jumps or spikes. A point to note is that technically price does not jump to a new level (to stay there) but spikes and quickly reverts to their previous levels. This price spike has been the most difficult aspect to model appropriately.

It is, therefore, clear from the above discussions that a pure diffusion process would not adequately capture the characteristics for electricity price series. A pure diffusion process approach has worked well in stock price modelling. For the electricity market, however, we need to incorporate a jump component with an appropriate intensity function to capture the spikes. Many of the traditional modelling approaches applied to financial market data e.g. equity, foreign exchange, and interest rates etc do not work well with spot electricity prices. This has been the experience for most researchers in this area as discussed in [3]. With respect to the equity market though, the work by [5] is an important contribution to detect jumps (as opposed to spikes). Their focus has been whether jump risks in stock returns are diversifiable.

In this paper we attempt to combine the ideas expressed in the cited literatures and explore a jump diffusion model for spot electricity prices in NSW. We allow both a deterministic time dependent factor as well as a latent factor combined with Poisson jumps to capture the observed characteristics of the spot electricity price series. We show how to calibrate such a model to the market data and describe the appropriate algorithm for that. The algorithm we employ generates, in a natural way, one period ahead forecast of spot electricity price. This, in turn, helps us determine the “goodness of fit” of such a model.

2 A Model for Spot Electricity Prices

In modelling commodity prices the approach of [7] has become quite popular. Their analyses depend upon both short dated and long dated futures contracts of the commodity. It also relates to the convenience yield as normally applied to futures contracts. Since the electricity, as a commodity, is different in this respect due to non-storability of the commodity for possible future consumption, the short-term, long-term concept introduced by [7] may not strictly apply to this market. Nevertheless, the ideas contained in [7] have important bearing in dealing with the electricity market.

It is clear from the earlier discussions that price jumps or price spikes are a natural characteristic of the electricity market and have to be built into the model. It is also useful as we may be able to adopt the models we develop here for pricing derivatives contracts on the electricity spot prices. To reliably model contingent claims prices we have to incorporate jumps in addition to the usual diffusion assumptions in the price process, which makes it far more complex compared to pricing derivatives on equities. In this context we need to be mindful of the theorem by [2] that leads to closed form solution, in most cases, of the contingent claims when the underlying security follows an affine\(^1\) jump-diffusion process (AJD). Although, we are not strictly focusing on electricity derivatives contract in this paper, we will strive to stay

\(^1\)Affine structure implies linear dependence on state variables
close to the AJD process so that our approach can be easily adapted for contingent claims pricing later.

Many researchers traditionally model log of the spot price of the commodity as in [7]. The existence of a significant jump component in the electricity prices it is worthwhile to re-consider whether a logarithmic transformation is useful. The logarithmic transformation affects the estimation of the jump component due its effect on the skewness of the distribution of the series. Since the derivatives contracts are written on spot price level and not on its log transformation, developing models of log transformation of spot price will not be useful. [6] find that the model of price level fits the forward contract prices better than the log-price level. In this paper we will, therefore, model the price level and not its log transformation. That way the models we develop will be better suited to pricing derivatives contracts on spot electricity prices.

In the original approach of [7] the log of the commodity price is modelled via two factors, both unobserved. The first factor captures the short-term variations and is modelled by an Ornstein-Uhlenbeck (OU) process whereas the second factor (the long-term variations) is modelled by an Arithmetic Brownian process (ABM). The commodity examined in Schwartz and Smith is crude oil and it displays non-stationarity. Hence the inclusion of the ABM process in their analysis is not only meaningful but is also a necessity since the OU process alone would not be able to capture the dynamics. Since our spot electricity price series is stationary (found by Augmented Dickey-Fuller tests) we need only include the OU process to capture the dynamics without the jumps. To capture the jump characteristic we include a jump component in the OU process.

Another difference from the structure in [7] for electricity spot prices is the inclusion of a time-dependent, deterministic function to capture the observed seasonality in the series. This arises mainly due to the nature of household consumptions of electricity depending on the season we are in. This also indicates that the intensity process for the Poisson component capturing the jumps in the series may not be constant, and is more likely to depend on seasonal factors.

With this background we are in a position to specify the spot price process \( P_t \) mathematically in terms of a deterministic, time-dependent function \( f(t) \), and a state variable \( X_t \). Although many researchers specify their models in continuous time setting and for implementation purposes use Euler discretisation, we prefer to stay in the discrete framework from the start. We set daily average electricity spot price, measured in dollars per megawatt-hour,

\[
P_t = f(t) + X_t.
\]

The deterministic, time-dependent part is described as a sinusoidal function along with a weekday dummy variable (wkdl). This specification is similar to that used in [6]. This last variable is to help distinguish between the price on a weekday and a weekend. We use the following specification for \( f(t) \):

\[
f(t) = \beta_0 + \beta_1 t + \beta_2 \sin \left( (t + \beta_3) \frac{2\pi}{365} \right) + \beta_4 \sin \left( (t + \beta_5) \frac{4\pi}{365} \right) + \beta_6 \text{wkdl},
\]

where \( \text{wkdl} = 1 \) if the day is a weekday otherwise it is zero.

The OU component \( X_t \) describes the unobserved component that captures the short-term dynamics, allows occasional jumps and is characterised by volatility clustering that is common to most financial time series. The notion of unobserved component is in the sense of state
space models and filtering theory. This volatility clustering is modelled as a GARCH (1,1) process. Thus we specify:

$$X_t = \phi X_{t-1} + h_0 \varepsilon_{1,t} + J(\mu_J, \sigma_J) \Delta \Pi(\lambda_t),$$

(3)

where $\varepsilon_{1,t} \sim N(0,1)$, the time varying variance $h_t = \alpha_0 + \alpha_1 \varepsilon_{1,t-1}^2 + \alpha_2 h_{t-1}$. The jump component is controlled by a Poisson-distributed variable with time dependent intensity function $\lambda_t$, and the jump amplitude is governed by a normally distributed variable with mean $\mu_J$ and variance $\sigma_J$. To capture seasonal effects in the jump component, we set

$$\lambda_t = \gamma_1 \text{winter}_t + \gamma_2 \text{autumn}_t + \gamma_3 \text{spring}_t + \gamma_4 \text{summer}_t$$

(4)

where the seasonal dummy variables indicate whether a particular date is in one of these seasons.

The specifications in equations (2) and (4) are just an assumption of the form of these functions for the dataset being analysed. These are based upon basic analysis of the data and similar decisions taken in the literature already referred to above. In this article we have not explored whether other forms of these functions have better properties.

The set of equations (1) and (3) describe our modelling approach to electricity spot prices and is already in state space form. In this state space representation of our problem equation (1) is the measurement equation and equation (3) is the state transition equation. The number of unknown parameters in this model is seventeen and these are estimated by the maximum likelihood method. Due to the presence of the unobserved component, $X_1$, we resort to Kalman filter to develop the likelihood function recursively. This process is described in detail, along with the modification needed due to the jump process, in the appendix. The parameter set is conveniently given by the vector,

$$\Theta \equiv \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \alpha_0, \alpha_1, \alpha_2, \mu_J, \sigma_J, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \phi\}.$$  

(5)

There are seventeen parameters to be estimated and it is not an easy optimisation task.

Although we have described in the appendix in detail the steps of the filtering algorithm including the modification needed to accommodate Poisson jumps, additional insights may be gained from Chapter 6 in [4]. In order to allow GARCH variance in the state dynamics, we need further non-trivial modification to the standard Kalman filter [4], and in particular Chapter 6 is an excellent reference source for this topic. Thus to conserve space, we refer the reader to that source material.

We implemented this algorithm in Gauss and used numerical optimisation to estimate the parameters as well as the standard errors from the information matrix. The filter algorithm produces the one-step ahead prediction of the state vector, which in this case is $X_t$.

3 Data and empirical results

Data used for this study was provided by Integral Energy. This represents every half hour NSW pool price covering the period 2002 through to 2006. We, however, use the daily average price for our modelling. We also estimate the model for two different periods. The first sample covers 2002 - 2003 containing 730 observations and the second sample covers 2004-2006 containing 1096 observations. Daily average prices are expressed in Australian dollars.
As a first step we check the stationarity of the spot price series using an ADF test. The main focus here is to statistically reject the unit root hypothesis in the electricity spot price series for entire sample period examined. The ADF test is based upon the following equation, where $P_t$ is the daily average electricity spot price and $\Delta$ is the difference operator:

$$\Delta P_t = \epsilon + \gamma P_{t-1} + \sum_{i=1}^{n} \delta_i \Delta P_{t-i} + \eta_t. \tag{6}$$

The hypothesis being tested is $H_0 : \gamma = 0$ as opposed to $H_1 : \gamma < 0$. The quantity $n$ in the above test is decided by a sample specific check so that the residual series is uncorrelated. The $t$-statistic for this test for the sample of 2002-2003 is -13.52 and that for the sample 2004-2006 is -27.04. The critical values for this test are obtained from the econometric software Eviews and the existence of unit root is rejected for both samples.

This unit root test convincingly supports the view that we need only the OU component in our model as opposed to [7] where both an OU and an ABM component were needed. For our analysis this OU component is given by the equation (3).

Focusing on the parameter estimates in Table 1, we note that the time trend component in the deterministic part (equation (2)) is insignificant in the second sample (2004-2006), whereas in the first component it is highly significant and displays downward bias. Although we cannot draw any firm economic conclusion from this, it is worthwhile to keep in mind that this market is still evolving and maturing.

Both samples display heteroscedasticity as seen from the quantity $(\alpha_1 + \alpha_2)$. The persistence in volatility in the short-term component (the OU part) has, however, gone down in the second sample. It is also interesting to observe that the autoregressive parameter ($\phi$) of the unobserved component in the second sample is about half the size of the earlier sample. It may stem from more efficient pricing by the participants in this market and is probably the result of better understanding of this commodity in this evolving market.

We now focus on the jump component, i.e., the parameters $(\mu_J, \sigma_J)$. Both these parameters are highly significant in both samples. Although the mean jump amplitude is higher in the second sample, its volatility is an order of magnitude higher in the second sample. That it might be so is also apparent from Figures 1 and 3 which are plots of the electricity spot prices over the whole periods. This may result from the supply concerns in the second sample period or the reflection of uncertainty in the regulatory environment governing this market. In this paper we are not in a position to shed further light on this aspect of the results. The appropriateness of the time varying jump intensity as captured by equation (4) is supported by the statistical significances of the estimated parameters in both sample periods.

Finally, we check on the predictive power of the model in both sample periods. Since the filtering algorithm recursively produces one step ahead projection of the state variable, we have shown in Figures 2 and 4 the possible price paths, i.e., the expected electricity spot prices. Using this information and the subsequently realised prices we can make comparative judgement about the usefulness of the model. In the traditional statistical sense the forecast ability of a model is judged by some measure of association between the forecasts and the realizations. However, there is an alternative to R-square measures and this is given by Theil's Inequality Coefficient. It was originally proposed in 1961 and has been employed by several researchers since then. We have applied Theil's Inequality Coefficient (TIC) to test this performance.

The TIC (Theil’s Inequality Coefficient) is given by the following expression. We assume
that the variable of interest is $z_t$ for $t = 1, 2, \ldots, T$, and its estimated value is given by $\hat{z}_t$:

$$
TIC = \frac{\sqrt{1/T \sum_{t=1}^{T} (z_t - \hat{z}_t)^2}}{\sqrt{1/T \sum_{t=1}^{T} z_t^2} + \sqrt{1/T \sum_{t=1}^{T} \hat{z}_t^2}}
$$

(7)

This coefficient always lies between zero and one, where the smaller the coefficient the better the estimate. For additional application of this measure the readers may refer to [1].

The computed value of TIC is smaller in the first sample compared to that in the second sample. Both values, however, indicate reasonable success in capturing the price path one day ahead. The lower value of the autoregressive parameter ($\phi$) in the second sample may have contributed to lower predictive power in the second sample.

4 Summary and conclusion

We have explored the modelling$^2$ of the electricity spot prices in New South Wales through a jump-diffusion process mixed with time varying deterministic trend component. The unobserved factor is driven by a diffusion process with time varying variance and a Poisson distributed jump component. We have outlined the algorithm needed to extract this latent factor from the observed price series and find that its one-step ahead prediction does contribute to the forecast of a complex electricity spot price series.

The very nature of the likelihood function of the Poisson mixture of Gaussian distribution requires some approximation for implementation purposes. To keep the computation burden low we have kept the upper limit of the infinite series summation to a small value. This is consistent with published articles in this area of research. One way to extend this study would be to experiment with this upper limit to check whether it improves predictive accuracy. Although we have allowed the jump amplitude to be a normally distributed variable, there may be other distributions that could prove useful in improving the predictive accuracy. This remains another possible extension of this study.

Standard errors are in parentheses below the parameters. TIC represents Theil’s inequality coefficient, a measure of model’s ability to predict the observed data. This is described in the text.

Appendix

State Space Model (SSM) with Poisson Jumps and Kalman Filter

The SSM in its basic form retains a VAR (1) structure for the state equation,

$$
y_t = \Gamma y_{t-1} + w_t + \zeta_t
$$

(8)

where the state equation determines the rule for generation of the states $y_t$, $p \times 1$ vector, from the past states $y_{t-1}$, for all time points $t = 1, 2, \ldots, T$. For completeness we assume that $w_t$ are $p \times 1$ independent and identically distributed zero-mean normal vectors with covariance $Q_t$. We assume that $Q_t$ is diagonal and may be constant. When we allow GARCH effect

$^2$In the published research in electricity spot price modelling, jump-diffusion model has emerged as the main technique. In the absence any other viable approaches it is not practical to compare the forecast ability of our model with a competing model as yet.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>11.88435</td>
<td>1.02417</td>
</tr>
<tr>
<td></td>
<td>(3.4316)</td>
<td>(0.0070)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-7.19647</td>
<td>0.00953</td>
</tr>
<tr>
<td></td>
<td>(1.4750)</td>
<td>(0.0078)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.91000</td>
<td>1.01030</td>
</tr>
<tr>
<td></td>
<td>(0.0260)</td>
<td>(0.0052)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1.03455</td>
<td>1.01031</td>
</tr>
<tr>
<td></td>
<td>(0.0152)</td>
<td>(0.0052)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.81608</td>
<td>1.01055</td>
</tr>
<tr>
<td></td>
<td>(0.0423)</td>
<td>(0.0052)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>1.06108</td>
<td>1.01057</td>
</tr>
<tr>
<td></td>
<td>(0.0159)</td>
<td>(0.0052)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>1.30011</td>
<td>1.04473</td>
</tr>
<tr>
<td></td>
<td>(1.5468)</td>
<td>(0.0122)</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>0.11923</td>
<td>1.05235</td>
</tr>
<tr>
<td></td>
<td>(0.0559)</td>
<td>(0.0064)</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.55549</td>
<td>0.33955</td>
</tr>
<tr>
<td></td>
<td>(0.0075)</td>
<td>(0.0013)</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.25873</td>
<td>0.33867</td>
</tr>
<tr>
<td></td>
<td>(0.0040)</td>
<td>(0.0013)</td>
</tr>
<tr>
<td>$\mu_J$</td>
<td>11.12589</td>
<td>16.18870</td>
</tr>
<tr>
<td></td>
<td>(1.2141)</td>
<td>(0.9561)</td>
</tr>
<tr>
<td>$\sigma_J$</td>
<td>240.36013</td>
<td>1333.91930</td>
</tr>
<tr>
<td></td>
<td>(13.0915)</td>
<td>(57.0935)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>1.97147</td>
<td>2.14528</td>
</tr>
<tr>
<td></td>
<td>(0.1567)</td>
<td>(0.1322)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>1.92343</td>
<td>1.88024</td>
</tr>
<tr>
<td></td>
<td>(0.1460)</td>
<td>(0.1174)</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>2.48438</td>
<td>1.94322</td>
</tr>
<tr>
<td></td>
<td>(0.1686)</td>
<td>(0.0306)</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>1.84487</td>
<td>2.07050</td>
</tr>
<tr>
<td></td>
<td>(0.1436)</td>
<td>(0.0381)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.16446</td>
<td>0.08159</td>
</tr>
<tr>
<td></td>
<td>(0.0265)</td>
<td>(0.0218)</td>
</tr>
<tr>
<td>TIC</td>
<td>0.3839</td>
<td>0.5489</td>
</tr>
</tbody>
</table>

Table 1: Maximum Likelihood Estimates of the Parameters of the Jump Diffusion Model
in some of the elements of the state vector, the corresponding element in $Q_t$ would then be time varying. The noise term $\zeta$ introduces the jump in the process and is assumed to be represented by

$$\zeta_t \sim N(j \cdot \mu_j, j \cdot \sigma_j^2), j = 1, 2, ..., \infty. \quad (9)$$

In equation (2) $j$ is a Poisson distributed random variable during a small interval $\Delta t$ characterised by a single parameter $\lambda_t \Delta t$. The state process is assumed to have started with the initial value given by the vector, $y_0$, taken from normally distributed variables with mean vector $\mu_0$ and the $p \times p$ covariance matrix, $\Sigma_0$.

The state vector itself is not observed but some transformation of these is observed but in a linearly added noisy environment. Thus, the measurement equation is given by,

$$z_t = d_t + A_t y_t + v_t. \quad (10)$$

In this sense, the $q \times 1$ vector $z_t$ is observed through the $q \times p$ measurement matrix $A_t$ together with the $q \times 1$ Gaussian white noise $v_t$, with the covariance matrix, $R$. In equation (3) $d_t$ is a purely deterministic time dependent variable. We also assume that the two noise sources in the state and the measurement equations are uncorrelated.

The next step is to make use of the Gaussian assumptions and the independence of Poisson distributed events and across times and produce estimates of the underlying unobserved state vector given the measurements up to a particular point in time. In other words, we would like to find out, $E(y_t|\{z_{t-1}, z_{t-2}, \ldots, z_1\})$ and the covariance matrix, $P_{t|t-1} = E((y_t - y_{t|t-1})(y_t - y_{t|t-1})')$. This is achieved by using Kalman filter and the basic system of equations is described below.

Given the initial conditions $y_{0|0} = \mu_0$, and $P_{0|0} = \Sigma_0$, for observations made at time $1, 2, 3...T$,

$$y_{t|t-1}^{(j)} = \Gamma y_{t-1|t-1}^{(j)} \quad (11)$$

$$P_{t|t-1}^{(j)} = \Gamma P_{t-1|t-1}^{(j)} \Gamma' + Q_t \quad (12)$$

$$y_{t|t}^{(j)} = y_{t|t-1}^{(j)} + K_t^{(j)} (z_t - A_t y_{t|t-1}^{(j)}), \quad (13)$$

where the Kalman gain matrix

$$K_t^{(j)} = P_{t|t-1}^{(j)} A'_t [A_t P_{t|t-1}^{(j)} A'_t + R]^{-1} \quad (14)$$

and the covariance matrix $P_{t|t}$ after the $t$th measurement has been made is

$$P_{t|t}^{(j)} = [I - K_t^{(j)} A_t] P_{t|t-1}^{(j)} \quad (15)$$

Equation (4) forecasts the state vector for the next period given the current state vector and the Poisson jump. Using this one step ahead forecast of the state vector it is possible to define the innovation vector as

$$v_t^{(j)} = z_t - A_t y_{t|t-1}^{(j)} \quad (16)$$

and its covariance as

$$\Sigma_t^{(j)} = A_t P_{t|t-1}^{(j)} A'_t + R \quad (17)$$

The description of the above filtering algorithms assumes that the parameters are known. In fact, we want to determine these parameters and this is achieved by maximizing the
innovation form of the likelihood function. The one step ahead innovation and its covariance matrix are defined by the equations (9) and (10) and since these are assumed to be independent and conditionally Gaussian, the log likelihood of the Poisson mixture of normal distribution is given by

$$
\ln(L) = \sum_{t=1}^{T} \ln \left[ \sum_{j=0}^{\infty} \omega(\lambda_t, j) \left\{ \left( 2\pi \right)^{-d/2} \cdot |\Sigma_t(\Theta)|^{-1/2} \exp \left( -\frac{1}{2} \nu_t^{(j)}(\Theta) \Sigma_t^{(j)-1}(\Theta) \nu_t^{(j)}(\Theta) \right) \right\} \right].
$$

(18)

In this expression $\Theta$ (collection of all the unknown parameters) is specifically used to emphasize the dependence of the log likelihood function on the parameters of the model. Once the function is maximized with respect to the parameters of the model, the inferred state vector is also available.

In practice the infinite sum in the above log likelihood function has to be approximated by something more appropriate for computation. The published papers in this area using equity market data normally use 10 as an upper limit for the summation term. For example, Kim, Oh and Brooks [5] use 4 for their study of jump risks in equity return. In this paper we use 6. Higher value will of course give better approximation but at the expense of rapidly increased computation time.

There are different numerical approaches that may be taken to carry out the maximization of the log likelihood function. The computational complexity and other numerical issues are beyond the scope of this paper.

References


Calculation of a risk measure for the net system load profile

Ken Russell
University of Wollongong

Pietro Cerone
Victoria University

Vivien Chalkis
The University of Queensland

1 Introduction

Integral Energy is one of three franchises which provide retail electricity in New South Wales (NSW). Integral Energy purchases wholesale electricity from the National Electricity Market and sells this to retail customers. The electricity market is unusual because the price at which electricity is sold to retail customers is fixed, but the price that the electricity retailer must pay for electricity from the National Electricity Market varies. Thus price changes incurred by Integral Energy are not passed on to their customers, introducing electricity price risk. Integral Energy uses the electricity hedge market, customer contract management, and the Electricity Tariff Equalisation Fund, along with market forecasting, to reduce its exposure to electricity price risk. These contracts are used to hedge Integral Energy against risk for electricity demand within the 95% confidence intervals of their long range electricity demand forecasts.

This project is in part concerned with comparing Integral Energy’s electricity price risk with that of the other two NSW electricity franchises: Energy Australia and Country Energy. Integral Energy want a measure of the relative volatility of their demand compared with that of the other two franchises. Integral Energy is most at risk during very high load periods, such as very hot days in summer and very cold days in winter. Integral Energy’s customers are largely based in western Sydney, where it gets hotter in summer and cooler in winter than on the coast. Houses in this area also tend to be less energy efficient, and Integral Energy believe these factors combine to produce higher demand during peak periods. Energy Australia’s customers are largely based on the coast around Sydney. Coastal temperatures are more stable than inland temperatures, which Integral Energy suggest results in less volatile demand. Country Energy largely provides electricity for the rest of the state and, from historical data, appears to have the least volatile demand.

Integral Energy provided data collected every 30 minutes from 1 January, 2002 to 31 December, 2006. The actual values for demand were given, as well as Integral Energy’s long term forecast values and their 95% confidence intervals. The supplied data included the temperature at Bankstown (in western Sydney), the National Electricity Market electricity price, and various load measurements. For this project, Integral Energy wanted us to consider the Net System Load Profiles (NSLPs) for each franchise. The NSLP is a measure of the load
over which the retailer has no control; it is discretionary electricity consumption controlled by the customers. The NSW State Load was also found to be of use in the modelling.

The brief from Integral Energy was to find a model for their NSLP given the temperature in Bankstown, which is used as a proxy in lieu of detailed temperature variation for the franchise, to estimate discretionary Load. This model would, in effect, be used to calculate expected load based around temperature in the short term. The second major task was to quantify the difference in volatility between the three franchise NSLP profiles.

2 Determination of the model

The data that Integral Energy provided came as an Excel spreadsheet of 87,648 rows and 61 columns. The rows each represent one observation on the characteristics of the electricity system. These observations are taken every 30 minutes throughout the day (at 00:30, ..., 24:00), for each day of the year. Data were provided for the calendar years 2002, 2003, ..., 2006.

The decision was taken to find a model that explained the Historical NSLP for Integral Energy over the years 2002 to 2005, and then to use the variables that were selected for that model to predict NSLP during 2006. A resulting good fit would suggest that the basic form of the model is satisfactory. It is important to note that this will not give exactly the same model for 2006 as for 2002 to 2005. For example, suppose that the best-fitting model for 2002 to 2005 is Predicted NSLP = a + b temp + c temp². Then the model that is fitted for 2006 would be of the same form, but new coefficients a₁, b₁ and c₁ would be calculated.

All statistical analyses were done using version 5.1 of the statistical package JMP [3]. As expected, the NSLP is strongly influenced by the temperature at Bankstown airport. It was quickly apparent that the type of day (working or non-working, the latter including weekends and public holidays), time of day, month of the year and year also had an influence, over and above the effect of temperature. Moreover, there was an interaction between the effects of type of day and month of year, meaning that the difference between the effects of two different months could vary according to whether it was a working day or non-working day.

It was considered that the load on the previous day at a given time might also be a good predictor of load on the current day at the same time. Unfortunately, Integral Energy does not learn what its load was for the previous day until several weeks subsequently. However, the load for the whole State can be found for the previous day, so this load was used as a proxy for the Integral Energy load of the previous day. As this reading from the previous day was being tried in a model, it was decided also to include the temperature at the same time of the day previously.

Based on the above considerations, a basic model of the following form is indicated:

\[
\text{NSLP for IE at time } t = \beta_0 + \beta_1 \times \text{temperature at time } t + \beta_2 \times \text{previous day's State load at time } t + \beta_3 \times \text{year} + \beta_4 \times \text{temperature on previous day at time } t + \sum_{i \geq 3} \text{various parameters } \beta_i \text{ multiplied by “indicator variables” to take account of the particular type of day, month of the year, year, and the interaction between type of day and month of year} + \text{random error.}
\] (1)
Note that it is not necessary for the statistician to construct the indicator variables in the above model. Modern statistical computer packages can do this automatically, as long as a qualitative variable is provided that specifies the value of the predictor (e.g., the month) at each observation.

The “random error” term is included in the model to account for the fact that, if observations are made at two times with exactly the same values of each of the predictor variables, the NSLPs will almost certainly be different. Random error does not mean that any mistakes have been made; the term is simply used to account for the variation that cannot be explained by any (combinations of) variables used in the model. If hypotheses about the parameters need to be tested, it is usual to assume that the random errors are independent observations from a $N(0, \sigma^2)$ population; that is, they come from a bell-shaped distribution, and no random error is inherently likely to be more variable than another. However, if no hypotheses will be tested, and the chief aim is to estimate the parameters $\beta_0, \beta_1, \ldots$, it suffices to assume that the random errors are uncorrelated and have a constant variance, $\sigma^2$. In this case, the parameters are estimated by the method of Ordinary Least Squares. This method gives unbiased estimators of the parameters even when the data are correlated, which is an important point to consider, as we subsequently show that the data are not uncorrelated.

In assessing how well a model fits the data, it is common to use the coefficient of determination, $R^2$. This measures what proportion of the variability in the values of NSLP can be predicted by the model being used. The higher that $R^2$ is, the smaller on average are the squared discrepancies between the actual values of NSLP and those predicted by the model. It is important to note that a high $R^2$ does not necessarily mean that we have the correct model; it simply indicates that the model is good at predicting the values of NSLP. For example, if there were only a slight quadratic trend to the NSLP over a limited range of time, and a straight line was fitted to the data, there might be a very high $R^2$, but this would simply indicate the good predictive ability of the model, and not its appropriateness. A referee cautions that there will be some instances where a high value of $R^2$ will not be associated with good predictive ability. We believe that its use in this investigation to select a model has not misled us, as evidenced by the satisfactory manner in which a model based on 2002–2005 data is then applied to 2006 data.

When the model in (1) was fitted to the 2002 to 2005 data, the resulting value of $R^2$ was 0.8343, which is quite reasonable for an initial model. The standard deviation, $\sigma$, of the unexplained variation is estimated by the Root Mean Square Error (RMSE); for this model, the RMSE was 0.1831.

Subsequent investigations found that there was a benefit from including in the model the interaction between the effects of time of day and type of day, and also the interaction between the effects of time of day and month of the year. However, with these terms in the model, there was no significant value in including the temperature at the same time on the previous day, so it was deleted from the model.

A comparison of the values of NSLP predicted by this model with the actual values of NSLP indicated that the model was quite satisfactory for moderate values of temperature, but was failing to predict the NSLP adequately when the temperature was high. The investigators then considered the possibility of allowing an increment to the predicted value of NSLP when the temperature exceeded a certain figure (a ‘breakpoint’). They also considered the possibility of including the square of temperature in the model, to provide a quadratic expression in temperature.

On the last day of the MISG meeting, Professor David Griffiths suggested that we consider
Figure 1: An illustration of a ‘jump’, and a different curve, on either side of a breakpoint at 25°.

the possibility of giving different weights to different observations when fitting the model. He subsequently also suggested that we consider the possibility of applying not just an increment when the temperature exceeded a breakpoint, but instead allowed a different model of the dependence on temperature. An illustration of this is provided in Figure 1, where there is both a ‘jump’, and a different quadratic curve, when the temperature exceeds 25°. [Please note that this Figure is for illustrative purposes only. The values on the y-axis are not indicative of any actual model.] However, not everyone agrees with this approach. A referee says ‘My gut reaction is of distaste for a curve which has a jump at a particular point.’

Breakpoints at 18°, 19°, …, 32° were investigated, and it was eventually determined that 20° yielded the best results. However, there is little to pick between any of these breakpoints. All resulted in $R^2$ values over 0.89, and the value for 20° was 0.9044, with an RMSE of 0.1391. If a differential weighting of various observations is ignored for the moment, the resulting optimal model is of the form:

$$
\text{NSLP for IE at time } t = \beta_0 + \beta_1 \times \text{temperature at time } t + \beta_2 \times (\text{temperature at time } t)^2 + I(\text{temp. } > 20)(\beta_0^* + \beta_1^* \times \text{temp. at time } t + \beta_2^* \times (\text{temp. at time } t)^2) + \beta_3 \times \text{previous day’s State load at time } t + \beta_4 \times \text{year} + \sum_{i \geq 5} \text{various parameters } \beta_i \text{ multiplied by “indicator variables” to take account of the particular type of day, month of the year, year, interaction between type of day and month of year, interaction between type of day and time of day and time of day, interaction between month of year and time of day} + \text{random error},
$$

(2)
where \( I(x) \) is the indicator function, taking the value 1 when \( x \) is true, and 0 otherwise.

The possibility was considered of giving more weight to observations associated with temperatures exceeding some value. Weighted regression is usually applied to data where it is thought that the variances of the observations are unequal (e.g., where the observations are the means of differing numbers of observations), with the aim of making the weighted observations have equal variance. In a situation such as the present one, where the data are regarded as having equal variances initially, the use of a weighted regression is a nonstandard attempt to make the model fit the NSLP values at higher temperatures. However, it is difficult to know what weights to use. Furthermore, the resulting \( R^2 \) values are not directly comparable. The value of \( R^2 \) is the ratio of the Model Sum of Squares (SS) to the Total SS. For unweighted regression, only the numerator changes when different models are fitted, so one can directly compare the values of \( R^2 \) for competing models. However, the effect of fitting a weighted regression is to alter both the numerator and denominator of the \( R^2 \) statistic, making a direct comparison of \( R^2 \) values less intuitive. This is illustrated in Table 1 by the basic ANOVA tables for a model with a breakpoint at 20°, for three sets of weightings: (a) equal weight for all observations, (b) double the weight for observations associated with temperatures greater than 28°, and (c) double the weight for observations associated with temperatures greater than 32°.

The maximum residual (discrepancy between observed and predicted values of NSLP) is minimized when the weight is doubled for observations associated with temperatures above 32°, but this does not have the greatest (weighted) \( R^2 \) or least (weighted) RMSE. This makes it difficult to find a consistent set of criteria with which to select a weighting scheme. Does one select the model that gives the least “weighted \( R^2 \)”, or the least RMSE, or the least maximum residual? The authors’ suspicion is that the selection of the appropriate breakpoint at which

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>641</td>
<td>12719.143</td>
<td>19.8427</td>
<td>1025.172</td>
</tr>
<tr>
<td>Error</td>
<td>69438</td>
<td>1344.003</td>
<td>0.0194</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>70079</td>
<td>14063.146</td>
<td></td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 1: Basic Analysis of Variance tables when the same model is fitted to the 2002 to 2005 data, and (a) all observations are equally weighted, (b) those observations corresponding to temperatures greater than 28° have twice the weight of all other observations, and (c) those observations corresponding to temperatures greater than 32° have twice the weight of all other observations.
Figure 2: Residuals (on the Y-axis) vs Predicted Values for the 2002–2005 data under model (2).

An assessment of model (2), when used with equally-weighted data, includes an examination of whether the assumptions of the analysis are met. For Ordinary Least Squares estimation, the main assumption is that the observations have equal variance. The usual way to examine this is to plot the residuals against the corresponding predicted values, and to see if there is any systematic pattern in the vertical spread of the residuals. (The predicted values are the values which the model predicts for each observation, and the residuals are the differences between the actual observations and the predicted values. The residuals are the best estimates of the error terms.) The lack of a systematic pattern will suggest that the variance is relatively constant.

Figure 2 displays the plot of the residuals against the predicted values for the model given in (2).

In order for tests of hypotheses about the terms in the model to be valid, the residuals should also appear to come from a single Normal distribution. However, the residuals from model (2) failed the Shapiro-Wilk test of Normality. A histogram of the residuals in Figure 3 indicates that there are an abnormally large number of residuals in each tail of the distribution.
Figure 3: Histogram of the Residuals for the 2002–2005 data under model (2).

(particularly the upper tail). One must therefore view any tests of hypotheses with caution. Notwithstanding this, the Analysis of Variance for the terms in model (2) gave $p$-values of "<0.0001" for every term in the model (counting an interaction, e.g., type of day by time of day, as one term), which suggests strongly that each term does belong in the model.

An additional assumption of the tests of hypotheses is that the error terms are uncorrelated. This assumption is examined by seeing whether the residuals are essentially uncorrelated. As the data take the form of successive observations over time (a time series), this is best considered by examining the autocorrelation between observations one, two, ... time units apart. When the temperature at the same time on the previous day was included in the model, there was negligible autocorrelation between residuals. However, when this term was deleted, the autocorrelation became pronounced. Figure 4 shows the values of the autocorrelation function, and partial autocorrelation function, of the residuals for lags of 0, 1, ..., 48 units of 30 minutes. The autocorrelation decreases, and then steadily increases as the lag approaches 48 (i.e., where we consider the correlation between a reading and the reading taken at the same time on the previous day), as would be expected. The partial autocorrelation, which shows the remaining autocorrelation at lag $(x + 1)$ after accounting for the autocorrelation at lags 1, ..., $x$, is small after a lag of 1. This suggests that the times series pattern in the residuals might be modelled by an autoregressive function with a lag of 1: AR(1). The existence of this autocorrelation is another reason to view any tests of hypotheses with
<table>
<thead>
<tr>
<th>Supplier</th>
<th>$R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral Energy</td>
<td>0.8934</td>
<td>0.1666</td>
</tr>
<tr>
<td>Energy Australia</td>
<td>0.9171</td>
<td>0.1299</td>
</tr>
<tr>
<td>Country Energy</td>
<td>0.9379</td>
<td>0.0832</td>
</tr>
</tbody>
</table>

Table 2: Results of applying (2) to the 2006 data of each energy supplier individually.

cautions.

3 Validation of the model

The terms in model (2) were then used to fit a model to the 2006 data, with all observations being equally weighted. Note that the year effect is omitted, as it is constant for all observations in the data set for 2006; any effect due to 2006 is automatically included in the initial constant in the model.

It was found that the same model as in (2) gave an $R^2$ value of 0.8934, and an RMSE of 0.1666. Of course, the actual values of the coefficients of the various terms will be different from those for the 2002 to 2005 data, but the same form of model was used. The plot of Residuals vs Predicted Values had the same basic shape as in Figure 2. The residuals again failed a test of Normality, and their histogram was similar to the one in Figure 3. All terms in the Analysis of Variance table, which tests the statistical significance of these terms, had $p$-values of “$< 0.0001$”. As before, the lack of Normality means that these $p$-values can only be taken as indicative.

We conclude from this validation that the model in (2) will be suitable for the combined 2002 to 2006 data, and should also be suitable for any additional data, provided that no systematic change occurs in the nature of the electrical operations or weather being measured.

4 Comparison of relative volatilities

The same form of model (2) (i.e., the same terms were in the model, but the coefficients were calculated from the data for the particular supplier) was applied to the 2006 data of the other two suppliers. All observations were equally weighted. Table 2 gives the results for all three electricity suppliers.

Thus, although the model had been determined from Integral Energy data, it provided a better fit (as measured by $R^2$) for the other two suppliers, and the measure of the variability of the unexplained variation was smaller for the other two suppliers than for Integral Energy. This suggests that the relative volatility in demand is greater for Integral Energy than for the other two suppliers. However, as the NSLP data has been scaled for each supplier so as to have a mean of 1, it will be necessary for Integral Energy to rescale it in order to get a true figure for the RMSE. [Note: if the raw data are multiplied by a factor of $x$ (say), this will also multiply the RMSE by a factor of $x$.]

As the temperature was measured in Bankstown, it is likely to be more applicable to the Integral Energy data than to the data for each of the other suppliers. Despite this, the NSLPs of the other two suppliers were better forecast by the common form of the model.
Figure 4: The autocorrelation function (higher figure), and (b) partial autocorrelation function, of the Residuals for the 2002–2005 data under model (2).
This demonstrates further that the Integral Energy data have greater volatility than the data from Energy Australia or Country Energy.

5 Discussion

The general form of the model given in (2) provides a quite satisfactory fit to the data, and we confidently recommend the use of this model to predict values of the NSLP. We have not specified the exact equation for the model, as its nature depends upon how any particular statistical software package deals with qualitative variables. For example, some software packages would treat the first month as a 'baseline' month, and measure the effect of month \( i \) \((i = 2, \ldots, 12)\) upon the NSLP by the departure of its effect from the effect of month 1; another package might treat the last month as the baseline. Moreover, as there are over 640 individual terms in the model, specifying the estimated value of each individual parameter would not be practical. A prospective user should set up a model of the form in (2), and then apply it to all the data (e.g., from 2002 to the present) that will be used to construct the model.

From consideration of the problem and the data, other predictors may prove useful to Integral Energy in modelling the NSLP load. In particular, using temperature measurements from around the franchise may provide more insight than just using the Bankstown temperature. Different observations may also be weighted (given more importance) on the basis of customer or usage density. Humidity may be a key predictor, as suggested in [1]. There may be information from the psychology literature, or from studies in physiology, to suggest at what temperature band people tend to change their behaviour, or to indicate whether there is an interaction between the effects of temperature and humidity. Also, other socioeconomic factors and demographic factors may be useful in predicting NSLP more accurately.

We have restricted the analysis to a conventional multiple regression that could be done in any of the standard commercially-available statistical software packages. However, a recent statistical innovation known as semiparametric regression is likely to be of use in trying to model the NSLP. An early reference on this is [2]. The R software [4] could be used to carry out any analyses that were planned. The software is free, readily downloaded, and widely used by statisticians. However, it is a package that is not easy for statistical novices to use. A referee strongly argues for the use of semiparametric regression and the R package in this problem, as they will provide a more robust analysis, not dependent on the normality and independence of the error terms.

Acknowledgment

We acknowledge the suggestions and statistical advice of Ray Chambers, Walter Davis, Kaye Marion, David Steel, and Jing You in this investigation. We especially thank a referee for very useful and helpful comments and suggestions.

References


Determining the independence of various measures of financial risk

Ed Maberly
University of Canterbury

Andy Wilkins
University of Queensland

1 Introduction

Trading Technology Australia (TTA) was founded in 1996 and has a background and client base in the financial markets. In 2002, TTA delivered its first Energy Markets project in the form of a risk management report to the CEO of a large electricity retailer. Since then, TTA has been working to further understand energy market issues such as the relevance of particular models used in trading and risk.

Common measures of financial risk include [1, 2, 5] Value-at-Risk (VaR) and Earnings-at-Risk (EaR). VaR measures the probable change of a portfolio’s position due to market movements within a given confidence interval. For instance, suppose a portfolio has a 1-day VaR of $40000 at the 99% confidence level. This means that, with a probability of 99%, the value of the portfolio will decrease by at most $40000 during 1 day. Companies are required to report their VaR to the regulators of the market.

EaR measures the probable loss in earnings due to market or volumetric movements within a given confidence level. For instance, suppose a company calculates its 30-day EaR to be −$20000 at the 99% confidence level. This means that, with a probability of 99%, the company’s earnings will be at least −$20000 on the 30th day in the future. Presently, companies are not required to report their EaR externally, but it is a useful measure to guide financial strategy. EaR is an equivalent measure to “profit-at-risk” and “relative VaR”, which are terms used in the literature.

The question posed by TTA to MISG2007 was: “How independent are VaR and EaR, specifically in the electricity market?” These quantities are computationally expensive to calculate, and if a map from one to the other could be found, significant savings would be generated for energy companies by making their risk reporting framework more streamlined, transparent and risk-compliant.

Attempting to answer this question, the MISG2007 team focused on simple example portfolios. These portfolios are particularly relevant for an electricity retailer.

In this paper we first give a brief introduction to the electricity market in Australia. Then we present the electricity-retailer portfolio that most succinctly captures the key properties and illustrates our key conclusions. Thirdly, formulae for VaR and EaR are derived for special cases where the forward price is equal to the spot price, in addition to spot price being related to demand. Finally, we state our conclusions.
The main conclusion is that for the electricity retailer, VaR is not strongly related to EaR in a way that can obviously reduce computational time. Of course, both measures will increase with market volatility, but accurately quantifying EaR given VaR does not seem to be possible. This is because of two factors. Firstly, VaR depends on forward prices, while EaR on spot prices, and these two prices may be unrelated. Secondly, VaR is sensitive to drops in prices, while EaR is sensitive to increases in prices and demands. Mathematically, VaR is looking at left tails of probability distribution functions, while EaR is looking at right tails. Since, in general, the distributions are non-symmetric there will be no obvious maps between VaR and EaR.

Throughout this paper we assume that the “risk free rate” (the rate of interest for money deposited in a bank) is zero. This simplifies formulae and is unlikely to change the main conclusions. We also calculate EaR and VaR at the 99% confidence level; other levels may be obtained by simply substituting another numerical value for the “0.01” that appears in all the formulae. A small amount of financial knowledge is assumed. A text that is immediately readable by applied mathematicians is [4], while a more advanced classic text is [3].

2 The electricity market in Australia

Generators, retailers and consumers form the electricity market in Australia. Generators generate electricity, consumers consume it, and retailers are the middlemen buying electricity from the generators and selling to the consumers. This paper concentrates on the electricity retailers. Retailers may also be generators, and generators may also be consumers.

The wholesale electricity market is administered and operated by the National Electricity Market Management Company Limited whose website, www.nemco.com.au, contains a lot of detailed and current information concerning the market. For our purposes, we can think of the generators placing bids into a “pool”: the bids describe the quantity and price of electricity that they are willing to supply. The demand is fixed in real time by the consumers, and electricity gets drawn from the pool to meet this demand, starting from the lowest priced electricity.

So, if the demand is low then electricity is only drawn from the generators who put in bids selling electricity at a low price, while if demand is high then electricity is also drawn (bought) from generators selling at a higher price. For instance, if generator A puts a bid of $1/unit into the pool, and is prepared to generate 100 units at this price, while generator B’s bid is $2/unit, then if the demand, $D \leq 100$ units, the price will be $1/unit. However, if the demand is greater than this the total price will be $100+2(D-100)$. Thus, total price (more strictly, “spot price”) is a piecewise linear function of demand.

The demand changes continually, and the generators’ bids into the pool may change frequently. The generators’ bids dictate the spot price of electricity. Some generators may choose to place low bids, guaranteeing their electricity will be used almost continually, while some generators may prefer to place exceedingly highly priced bids hoping that the demand will spike one day and their electricity will be bought.

The retailer sits in the middle. The important point is that, they sell electricity to the consumers at a fixed price, while, in the simplest situation, they buy electricity from the pool at the spot price. This means that the energy retailers face substantial financial risk. The spot price is set by the generators, and because the demand fluctuates, the spot price may be significantly greater than (over ten times) the fixed price paid by the consumers. Therefore
energy retailers enter into forward contracts with the generators.

3 The example portfolio for an electricity retailer

Recall that the risk-free rate is zero by assumption. Our canonical energy retailer portfolio is set up as follows. Define the following variables.

- \( t \): the time variable. The current time is set to be \( t = 0 \).
- \( P_t \): the price the customers pay the retailer per unit of electricity. Assume this is fixed.
- \( P(t) \): the spot price per unit of electricity at time \( t \). This is a stochastic variable with expected value at time \( t \) of \( P(t) \). Without any forward contracts \( P(t) \) is the price the retailer would have to pay per unit of electricity at time \( t \).
- \( Q(t) \): the demand by customers for electricity at time \( t \). This is a stochastic variable with expected value at time \( t \) of \( Q(t) \).

For the purposes of calculating VaR and EaR, the retailer uses models, historical data, or some other technique to find the probability distribution functions for \( P(t) \) and \( Q(t) \), and hence their expected values for future times \( t > 0 \).

Now suppose the retailer enters into the following long forward contract with the generator.

- The retailer will buy from the generator \( Q(T) \) units of electricity at price \( P_f \) at time \( t = T \).

\( P_f \) is known as the forward price.

Evidently, this is a very simple portfolio, consisting of just \( Q(T) \) forward contracts with the same price expiring at the same time \( t = T \). A real-life situation would consist of many forwards and other derivatives expiring at different times. It is difficult to see how the extra sophistication would change our conclusions, however.

Finally, define the forward price in the future:

- \( F(t,T) \) is the price specified in a forward contract set up at time \( t \) which will expire at time \( T \). It is a stochastic variable, but again assume the retailer can calculate its probability distribution.

Clearly, \( P_f = F(0,T) \).

3.1 VaR

The VaR can be calculated for any time, \( t \), in the future. Suppose at time \( t \leq T \) the retailer would like to sell the \( Q(T) \) forward contracts to another party. It would sell them at price \( F(t,T) \), so the payoff would be

\[
\text{payoff}(t) = Q(T)(F(t,T) - P_f).
\]  

The \( t \)-day VaR is calculated from this payoff by constructing the probability distribution function for the payoff, \( \Pr(\text{payoff}) \) (at time \( t \)) as illustrated in Figure 1. Mathematically,

\[
\Pr(\text{payoff}(t) < \text{VaR}) = 0.01,
\]

for the 99% confidence level. In this formula, VaR will typically be negative, but its absolute value is reported.
Figure 1: The probability distribution, \( \Pr(\text{payoff}) \) for the payoff at time \( t \). The shaded region has area 1%.

### 3.2 EaR

The EaR can also be calculated for any time, \( t \), in the future. For times \( t < T \), the retailer must sell \( Q(t) \) units of electricity to the customers at price \( P_c \), and buy \( Q(t) \) units at the spot price, \( P(t) \). Thus, its cashflow is

\[
\text{cash flow}(t) = Q(t)(P_c - P(t)).
\] (3)

The \( t \)-day EaR is calculated from this cash flow by constructing the probability distribution function for the cash flow, \( \Pr(\text{cash flow}) \), (at time \( t \)) as illustrated in Figure 2. Mathematically,

\[
\Pr(\text{cash flow}(t) < \text{EaR}) = 0.01,
\] (4)

for the 99% confidence level calculation.

For time \( t = T \), the retailer must sell \( Q(T) \) units to the customers at price \( P_c \), so its revenue will be \( Q(T)P_c \), as before. However, now the retailer must buy the \( Q(T) \) units of electricity as specified in the forward contract, which costs \( Q(T)P_f \). In addition, if \( Q(T) > Q(T) \) it must also buy the excess units at the spot price, \( P(T) \). Its total cash flow is therefore

\[
\text{cash flow}(T) = Q(T)P_c - \left[ Q(T)P_f + (Q(T) - Q(T))P(T)\Theta(Q(T) - Q(T)) \right],
\] (5)

where

\[
\Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}.
\]

In the following section, for simplicity we shall use equation (3) to calculate EaR. Use of the more complicated equation (5) is not expected to alter our conclusions. EaR is sensitive to those situations where cash flow is drastically negative. In both situations (equation (3) and (5)) this is when \( Q \) is large, implying a large \( P \), and the cash flow is dominated by the large \(-QP\) term.
3.3 Summary

Equations (1), (3) and (5) demonstrate that in this situation:
- the $t$-day VaR depends on the probability distribution of the forward prices of the contracts in the portfolio $F(t, T)$;
- the $t$-day EaR depends on the probability distributions of the demand, $Q(t)$, and the spot price $P(t)$.

The problem thus reduces to finding relations between integrals of the three probability distributions. It is immediately obvious that if these three are completely uncorrelated then VaR and EaR will be unrelated. This conclusion may be generalised to more complicated portfolios containing more stochastic variables. In the following sections, simple correlations are assumed which lead to potentially useful relationships between VaR and EaR.

4 Detailed analysis of special cases

In this section we present formulae for VaR and EaR for special situations which are mathematically nongeneric, but which are hopefully relevant to financial situations.

To make some headway, first assume that the spot price and the forward price are equal:

$$P(t) = F(t, T) .$$

This is widely used in other markets, such as gold, but may be less relevant to the electricity markets.

Furthermore, in many markets the price is strongly correlated with the demand. This is certainly true for electricity markets where generators bid into the pool at certain prices and the demand dictates whose electricity gets drawn from the pool. This means that at any one time — say $t = 0$ — the spot price is a deterministic increasing function of demand; in fact, as described in Sec 2, total price is a piecewise linear function of demand:

$$\text{Total price} = Q(0)P(0) = \text{piecewise linear}(Q(0)) .$$
Unfortunately, for forecasting purposes, the generators’ bids are unknown in the future. They can be modelled by a stochastic variable. This implies that at time $t$ in the future the relationship between spot price and demand is unknown, but can be modelled by a positive linear correlation. We present the linearly correlated case: more general correlations can be similarly studied.

### 4.1 Correlations in general

#### 4.1.1 Perfect linear correlation

Two variables, $P$ and $Q$, are perfectly (linearly) correlated if and only if their joint probability distribution function, $\Pr_{P,Q}$, is of the form

$$\Pr_{P,Q}(P,Q) = \delta(P - \alpha Q - \beta) \Pr_Q(Q) ,$$

where $\Pr_Q$ is normalised to unity: $1 = \int \Pr_Q$, and $\delta$ is the Dirac delta function. The variables are positively correlated if and only if $\alpha > 0$. As the name suggests, $\Pr_Q$ is the marginal probability distribution for $Q$ after integrating over all $P$. That is,

$$\Pr_Q(Q) = \int dP \, \Pr_{P,Q}(P,Q) .$$

Similarly, the distribution function for $P$ reads

$$\Pr_P(P) = \frac{1}{\alpha} \Pr_Q \left( \frac{P - \beta}{\alpha} \right) .$$

Evidently the delta function is enforcing the linear relation $P = \alpha Q + \beta$.

#### 4.1.2 Completely uncorrelated variables

Conversely, two variables, $P$ and $Q$, are completely uncorrelated if and only if their joint probability distribution function, $\Pr_{P,Q}$, is a product:

$$\Pr_{P,Q}(P,Q) = \Pr_P(P) \Pr_Q(Q) .$$

We shall study these two cases — correlated and uncorrelated — below.

#### 4.1.3 Expectation values and the correlation coefficient

Denote the expectation value of a variable by $\mathbb{E}$. So

$$\mathbb{E}(P) = \int dP \, P \Pr_P(P) = \int dPdQ \, P \Pr_{P,Q}(P,Q) ,$$

for instance. Statisticians sometimes use the correlation coefficient of two variables, $P$ and $Q$, which is defined as

$$\text{Corr}(P,Q) = \frac{\mathbb{E}(PQ) - \mathbb{E}(P)\mathbb{E}(Q)}{\sqrt{\mathbb{E}(P^2) - \mathbb{E}^2(P)}\sqrt{\mathbb{E}(Q^2) - \mathbb{E}^2(Q)}} .$$
Substituting the two cases above into this formula yields
\[
\text{completely correlated } \Rightarrow \text{ Corr} = \text{sign}(\alpha) , \\
\text{completely uncorrelated } \Rightarrow \text{ Corr} = 0 .
\]
It is not true that if Corr = 0 then the variables are uncorrelated, for Corr only picks up \textit{linear} correlations. In our case, Corr is less useful than considering the underlying probability distribution functions.

4.1.4 \textbf{Partially correlated variables}

Now introduce the joint probability distribution function for partially correlated variables:
\[
\text{Pr}_{P,Q}(P, Q) = \lambda f(P)g(Q) + (1 - \lambda)\delta(P - \alpha Q - \beta)h(Q) .
\]
(7)

If \(\lambda = 0\) then the variables are completely (linearly) correlated, while if \(\lambda = 1\) the variables are completely uncorrelated. The functions \(f\), \(g\) and \(h\) must satisfy the normalised condition:
\[
1 = \lambda \left( \int f \right) \left( \int g \right) + (1 - \lambda) \int h .
\]
We show below how the correlation coefficient may be expanded in powers of \(\lambda\) or \((1 - \lambda)\).

4.1.5 \textbf{VaR and EaR}

We shall perform an expansion of VaR and EaR around \(\lambda = 0\) and \(\lambda = 1\). In general, VaR (at the 99\% level), is defined by equation 2, which in this case reads
\[
0.01 = \lambda \int_{-\infty}^{P_{f+VaR/Q}} dPdQ \text{Pr}_{P,Q}(P, Q) \\
= \lambda \int_{-\infty}^{P_{f+VaR/Q}} dP f(P) \left( \int g \right) \\
+ (1 - \lambda) \int_{P_{f-\beta+VaR/Q}/\alpha}^{\infty} dQ h(Q) .
\]
(8)

The last expression holds only if \(\alpha > 0\) (positive correlation). Otherwise the formula reads
\[
0.01 = \lambda \int_{-\infty}^{P_{f+VaR/Q}} dP f(P) \left( \int g \right) + (1 - \lambda) \int_{P_{f-\beta+VaR/Q}/\alpha}^{\infty} dQ h(Q) .
\]
Similarly, EaR, defined by equations 3 and 4, reads
\[
0.01 = \int_{Q(P_{e}-P)} dPdQ \text{Pr}_{P,Q}(P, Q) \\
= \lambda \int_{Q(P_{e}-P)} dPdQ f(P)g(Q) \\
+ (1 - \lambda) \int_{Q(P_{e}-\alpha Q-\beta)} dQ h(Q) .
\]
(9)

Now let us examine particular cases more thoroughly.
4.2 Perfect positive correlation between price and demand
In this case, the parameter $\lambda$ in equation (7) is zero, $\alpha > 0$, and $h$ becomes the probability distribution function for $Q$. Equations (8) and (9) read

$$0.01 = \int_{-\infty}^{(P_f - \beta + \text{VaR}/Q)/\alpha} dQ \ Pr_Q(Q) \quad (10)$$

$$0.01 = \int_{Q(P_c - \alpha Q - \beta < \text{EaR})} dQ \ Pr_Q(Q)$$

$$= \int_{Q_+}^{\infty} dQ \ Pr_Q(Q) + \int_{-\infty}^{Q_-} dQ \ Pr_Q(Q) . \quad (11)$$

Here

$$Q_\pm = \frac{P_c - \beta \pm \sqrt{(P_c - \beta)^2 - 4\alpha \text{EaR}}}{2\alpha} . \quad (12)$$

In most financial situations the probability distributions will be such that EaR is negative (it is greater than 1% likely the firm will have negative cash flow), so $Q_- < 0$ and the final term vanishes, since physically we should have $Pr_Q(Q) = 0$ for $Q < 0$.

4.3 Completely uncorrelated price and demand
In this case, the parameter $\lambda$ in equation (7) is unity, $f$ is the probability distribution function for $P$, and $g$ for $Q$. Equations (8) and (9) read

$$0.01 = \int_{-\infty}^{P_f + \text{VaR}/Q} dP \ Pr_P(P) \quad (13)$$

$$0.01 = \int_{Q(P_c - P < \text{EaR})} dP dQ \ Pr_P(P) Pr_Q(Q)$$

$$= \int_{0}^{\infty} dQ \ Pr_Q(Q) \int_{P_c - \text{EaR}/Q}^{\infty} dP \ Pr_P(P)$$

$$+ \int_{-\infty}^{0} dQ \ Pr_Q(Q) \int_{-\infty}^{P_c - \text{EaR}/Q} dP \ Pr_P(P) . \quad (14)$$

Once again in real financial situations the probability distributions will be zero for $Q < 0$ (negative demand), the final term vanishes, so the EaR expression takes the form

$$0.01 = \int_{0}^{\infty} dQ \ Pr_Q(Q) \int_{P_c - \text{EaR}/Q}^{\infty} dP \ Pr_P(P) .$$

Finally note that solving the first expression for VaR would probably yield $\text{VaR} < 0$, but that we would report, by convention, $|\text{VaR}|$ to the regulatory bodies.

4.4 Strongly positively correlated price and demand
In this case, the parameter $\lambda$ in equation (7) is close to zero and $\alpha > 0$. Further simplification of equations (8) and (9) is not possible without assuming some form for the functions $f$, $g$ and $h$. 

109
Assume
\[ \Pr_{P,Q}(P,Q) = \frac{1}{\sqrt{\pi \epsilon}} \exp \left( -\frac{(P - \alpha Q - \beta)^2}{\epsilon} \right) h \left( \frac{P + \alpha Q - \beta}{2\alpha} \right). \] (15)

The probability distribution must be normalised which implies (after changing variables to \( u = P - \alpha Q - \beta \) and \( v = (P + \alpha Q - \beta)/2\alpha \)),
\[ 1 = \int h. \]

This probability distribution implies
\[ \mathbb{E}(P - \alpha Q - \beta) = 0, \]
\[ \mathbb{E}(P - \alpha Q - \beta)^2 = \epsilon/2. \]

Hence, this probability distribution is describing the situation when, on average, \( P \) and \( Q \) are linearly related \((P = \alpha Q + \beta)\), but their perfect correlation is spoiled by Gaussian white noise, with zero mean and variance \( \epsilon/2 \).

A useful result is that in the limit of small \( \epsilon \),
\[ \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{\pi \epsilon}} \exp \left( -(x - x_0)^2 / \epsilon \right) f(x) = \int_{-\infty}^{\infty} dy \frac{1}{\sqrt{\pi}} \exp \left( -y^2 \right) f(x_0 + \epsilon y) \]
\[ = f(x_0) + \frac{1}{2} \epsilon f''(x_0) + O(\epsilon^2). \]

for (integrable, etc) functions \( f \). This may be used to expand \( \Pr_{P,Q} \) in small \( \epsilon \) so that it looks more like the previous definition for partially correlated variables, Eq (7):
\[ \Pr_{P,Q}(P,Q) = \delta(P - \alpha Q - \beta)h(Q) + \frac{1}{4} \epsilon \delta''(P - \alpha Q - \beta)h \left( \frac{P + \alpha Q + \beta}{2\alpha} \right) + O(\epsilon^2). \]

This motivates the use of the notation ‘h’ for the probability distribution function in equation (15) when comparing with the general expression of equation (7). It also demonstrates that \( \epsilon \) is \( O(\lambda) \) in the notation of the previous sections. We shall now perform a power series expansion in \( \epsilon \) of equations (8) and (9).

The formula for VaR reads
\[ 0.01 = \int_{-\infty}^{(P - \beta + \text{VaR}/\alpha)/\epsilon} dQ \left( h(Q) + \frac{\epsilon}{16\alpha^2} h''(Q) + O(\epsilon^2) \right). \] (16)

The formula for EaR reads
\[ 0.01 = \left( \int_{Q^+}^{\infty} dQ + \int_{-\infty}^{Q^-} dQ \right) \left( h(Q) + \frac{\epsilon}{16\alpha^2} h''(Q) + O(\epsilon^2) \right), \] (17)

where \( Q_{\pm} \) is given in equation (12). Boundary terms encountered during the derivation of these expressions are zero through assuming a well behaved \( h \). As mentioned before, it is likely that \( Q_- < 0 \), so for financially realistic cases with \( h(Q) = 0 \) for \( Q < 0 \), the second integral is zero (more generally, much smaller than the first integral).
4.5 Weakly positively correlated price and demand

In this case, the parameter $\lambda$ in equation (7) is close to unity and $\alpha > 0$. Further simplification of equations (8) and (9) is not possible without assuming some form for the functions $f$, $g$ and $h$. It is not possible to make the large $\epsilon$ expansion of equation (15), since this does not yield $\Pr_{P,Q}(P, Q) = f(P)g(Q)$ as the first term.

Instead, assume

$$\Pr_{P,Q}(P, Q) = f(P - \epsilon \alpha Q)g(Q - \epsilon P/\alpha) ,$$

with $\alpha > 0$. For $\epsilon = 0$ this reduces to the completely uncorrelated situation, $\Pr_{P,Q} = fg$. The probability distribution must be normalised. Upon changing variables, $u = P - \epsilon \alpha Q$ and $v = Q - \epsilon P/\alpha$, and expanding to first order in $\epsilon$, the normalisation condition reads $1 = (\int f)(\int g)$. The functions $f$ and $g$ may be arbitrarily scaled without changing the form for $\Pr_{P,Q}$, so let us choose

$$\int f = 1 = \int g .$$

The normalisability also implies that both $f$ and $g$ vanish at $\pm \infty$, which shall be employed below when we use integration by parts.

Expanding to first order in $\epsilon$ gives

$$\Pr_{P,Q}(P, Q) = f(P)g(Q) - \epsilon \alpha Q f'(P)g(Q) - \epsilon \frac{P}{\alpha} f(P)g'(Q) + O(\epsilon^2) .$$

Hence

$$\Pr_{P}(P) = f(P) - \epsilon f'(P) \int dQ \, \alpha Q g(Q) + O(\epsilon^2) ,$$

$$\Pr_{Q}(Q) = g(Q) - \epsilon g'(Q) \int dP \, P f(P)/\alpha + O(\epsilon^2) ,$$

which yield

$$\mathbb{E}(P) = \int dP \, Pf(P) + \epsilon \alpha \mathbb{E}(Q) ,$$

$$\mathbb{E}(Q) = \int dQ \, Qg(Q) + \epsilon \mathbb{E}(P)/\alpha ,$$

to first order. The expectation value of $P$ depends on the expectation value of $Q$, and vice versa, so there is a weak positive correlation between the two variables.

VaR, equation (8), becomes

$$0.01 = \left( \int_{-\infty}^{P_f + \text{VaR}/Q} dP \, f(P) \right) - \epsilon \alpha f(P_f + \text{VaR}/Q) \mathbb{E}(Q) \tag{18}$$

EaR, equation (9), becomes

$$0.01 = \int_{-\infty}^{\infty} dQ \, g(Q) \int_{-\infty}^{P_f - \text{EaR}/Q} dP \, f(P) + \int_{-\infty}^{0} dQ \, g(Q) \int_{-\infty}^{P_f - \text{EaR}/Q} dP \, f(P)$$

$$- \epsilon \alpha \left( \int_{-\infty}^{0} - \int_{-\infty}^{\infty} \right) dQ \, Qg(Q)f(P - \text{EaR}/Q)$$

$$- \epsilon \alpha \left( \int_{-\infty}^{P_c} - \int_{-\infty}^{\infty} \right) dP \, Pf(P)g(EaR/(P_c - P)) \tag{19}$$

111
Once again, the integrals over the positive domains dominate since it is likely that the functions $f$ and $g$ are small or zero for negative argument.

4.6 Discussion

In this main section of the paper, we have studied a number of situations in which VaR and EaR might be able to be related. Examining equations (10), (13), (16) and (18) demonstrates that VaR depends upon the left tail of a distribution. Conversely, equations (11), (14), (17) and (19) demonstrate that EaR depends upon the right tail of a distribution. Physically this corresponds to VaR measuring the likelihood of a large decrease in derivative price (which is equal to spot price in this section), while EaR measures the likelihood of a large increase in demand or spot price.

5 Conclusions

VaR and EaR are calculated from integral equations involving (joint) probability distribution functions for spot prices, forward prices and demands, as well as potentially other more exotic option prices. The MISG team was asked by Technology Trading Australia whether VaR and EaR were related in some way which might reduce time spent computing these risk measures while additionally making risk reporting more streamlined, transparent and risk-compliant.

In some cases the probability distribution functions involved are related and this implies some simplification of the computations required to extract the two measures of risk, since in real-life situations a large amount of computational time is spent using historical data to build the probability distribution functions. However, in all but the most trivial situations, such as symmetric probability distribution functions, this does not appear to imply simple algebraic expressions relating VaR to EaR.

Our main conclusion is that for electricity retailers, VaR and EaR are not related in any useful way. In rough financial terms, the VaR for an electricity retailer is sensitive to drops in (forward) prices, since then the value of a portfolio is reduced. However, the EaR is usually more sensitive to increases in electricity prices and demands since these will cause losses in earnings. This strongly suggests there is no simple map from VaR to EaR. Other markets may not share these features. However, in these cases any theoretical, simple relationships between VaR and EaR may be inaccurate in practice due to the large variety of rather independent stochastic variables that they depend upon.

Our recommendations for reducing time spent on computing VaR and EaR are:

- Identify the probability distribution functions required for VaR and EaR, and compute any that are common between VaR and EaR only once.
- It may be easier to use the stochastic set (demand, bid structure) than the set (demand, spot price) since in some cases the bid structure fluctuates only a little, or may even be considered fixed.
- In some cases it may be appropriate to assume the forward prices are equal to the spot prices, which will reduce the number of variables in the problem.
- The probability distributions would be simple to calculate if the underlying stochastic processes could be found and calibrated from historical data. In this method, it is the calibration (and continual recalibration required by regulators) that is difficult.
Acknowledgements

This paper summarises the work done at MISG2007 by the project moderators and the MISG team. The moderators are very grateful to the team: Jeff Dewynne, Chulmin Kang, Eungyoung Kim, Yan-Xia Lin, Jaekyu Lee, Michael McCrae, Song-Ping Zhu and Andrew Zogas (in alphabetical order).

References


Maximizing the contribution of wind power in an electric power grid

*Geoff Pritchard*  
*University of Auckland*

*Winston L. Sweatman*  
*Massey University*

*Kim Nan*  
*University of Auckland*

*Mike Camden*  
*Statistics New Zealand*

*Bill Whiten*  
*University of Queensland*

1 Introduction

The operator of an electric power grid must make continual adjustments in response to fluctuating conditions. On a short time scale (less than one hour), there have traditionally been two main kinds of variability for which the operator must compensate:

- Fluctuations in loads, offset by certain power plants with the ability to change their output quickly (frequency keepers).
- Unplanned outages (i.e. failures) of generators or transmission lines, offset by uncommitted power plants with the ability to begin generating quickly (spinning reserve).

Wind farms are also a source of variability, due to the fluctuating strength of the wind. In power grids where wind farms constitute only a small fraction of the total generating resource, the resulting variability can be included in the first category above, and dealt with as part of frequency-keeping. However, in a power grid with a large wind component (10-50% of time-averaged generation), wind-related variability may have to be treated as a third category of fluctuation, with specific operating procedures required to compensate for it. Both the amount of capacity needed, and the frequency with which it is required, are likely to fall somewhere between the traditional frequency-keeping and the spinning-reserve requirements.

This paper treats several different approaches to the problem of optimal grid operation in the presence of wind power. All the analyses are essentially economic in nature. That is, the cost of reserving spare capacity (either generation or transmission) is balanced against its expected benefits. Longer-term problems of optimal capital investment are not considered, although the methods and results here may be relevant to such problems.

The situation of the New Zealand power grid is particularly relevant. Although wind farms contributed only 1.5% of system energy in 2005 [1], it is anticipated that this share may increase to more than 10% over the next decade. New Zealand is also well-endowed
with hydroelectric generation, which can play a key role in compensating for wind-related variability.

The essential problem at hand is the dispatch problem (also known as the optimal power flow problem). Suppose that each power station offers one or more tranches of power to the market, with each tranche being a fixed quantity of power at a fixed offer price. The problem is then to choose which tranches to accept (in whole or in part) to meet demand while minimizing the cost of power at the stated offer prices. (A similar problem would be solved even if there were no market; in that case, the offer prices would be replaced by some other estimates of the short-run marginal cost of generation.)

In practice, a system operator wishes to minimize the cost of power purchased over a fixed period of time (of duration $T$ typically of the order of 5 minutes). If the behaviour of the power system is steady during this period, it suffices to solve the dispatch problem just once. The solution gives the constant power outputs that each station must maintain throughout the period.

Now consider the situation where some parameters of the dispatch problem vary with time. In particular, the quantity offered in a tranche (such as from a wind farm) may be a function of time (usually unpredictable in advance, i.e. a stochastic process). In this context, the dispatch problem becomes one of stochastic control. The actual cost to be incurred over the period is a random variable, and our objective is to minimize its expectation.

We consider several different ways to formulate such a time-varying dispatch problem. In Section 2, we study a simple model with three power generators. We explore the optimisation process to minimise the cost of meeting a power demand load. In Section 3, the available wind power may be a general stochastic process, but the response of the rest of the system is somewhat idealized (all ramp rates are treated as either zero or infinity). In Section 4, we briefly consider the problems of collection and analysis of wind speed data for use in wind-power generation planning. We conclude in Section 5.

2 The three-power-station model

Our initial model (Figure 1) incorporates just three power stations: a wind farm, a low-cost (thermal) generator that can adjust only slowly to new power levels, and a fast-ramping but high-cost (hydro) generator. To simplify the issues we assume that there are no significant transmission losses or constraints, so that all power is effectively generated and consumed at a single location. We also assume a constant load. As the rapid fluctuations in wind-power generation occur, the power output of the low-cost generator is adjusted to try to balance the load. As this low-cost generator is slow-ramping, it cannot immediately compensate for rapid changes in the wind, and the balance must be met by the fast-ramping high-cost generator. (A more general interpretation for this system could be as a model for the interaction of power provision sectors rather than of individual power stations.)

Power losses due to a drop in wind need high-cost power while the low-cost generator ramps up to its new required level. The costs due to the use of this high-cost power can be reduced by running the low-cost generator at a higher base level, and by not using some of the available wind power. Similarly an increase in wind power can be anticipated by using a base of high-cost power to allow more rapid adjustment to use the additional resource.

Within this model we may study how best to balance the sources of power for the cheapest long-term operation. This will depend upon the costs of the different generators, and the
nature of the variation in wind power. The mathematical solution typically involves finding the minimum point of a U-shaped function.

Notation for this section:
- $L$: Load, the amount of power required or that can be transmitted.
- $W(t)$: Wind speed by time.
- $P_w(t)$: Wind power available.
- $P_u(t)$: Wind power used.
- $h$: Wind-power headspace, the maximum amount of wind power allowed.
- $\Delta P_w$: Step change in wind power.
- $P_L(t)$: Power from low-cost generator by time.
- $\Delta P_L$: Excess power from low-cost generator.
- $c_L$: Cost of power from low-cost generator.
- $r$: Maximum rate of change for low-cost generator.
- $P_H(t)$: Power used from high-cost generator by time.
- $\Delta P_H$: Excess power from high-cost generator.
- $c_H$: Cost of power from high-cost generator.
- $T$: Total time considered.

We make the following general assumptions. Power output from the low-cost generator may be increased or decreased at a constant rate $r$. The changes in output of the high-cost generator, and the wind generator, are taken to be instantaneous. The model is parametrised by $h$, the portion of the load which wind power is allowed to serve. The system attempts to cover any load not served by wind power from the low-cost power source, with occasional use of high-cost power to match fluctuations in the wind.

In order to optimise $h$ we require the costs for each of the power generators. For the purposes of these models, the cost of wind power was taken to be 0, while the costs of units of low- and high-cost power were taken to be $c_L$ and $c_H$, respectively. (We can set the cost of wind power generation equal to zero as it is the marginal costs that are involved in the optimisation. Wind power is likely to be the cheapest power source.)

The simplest kind of model is reactive, in which the system responds to current wind-power generation. Upon a sudden drop in wind power, it takes time to ramp up the low-cost generator to meet the power demand. The deficit must be met by the high-cost power station. Even with unlimited wind power, there is likely to be a point at which it is more cost effective to balance wind power with a contribution from the low-cost generator and to constrain the
amount of high-cost power needed should the wind drop.

A more sophisticated approach can be taken if changes in wind-power generation can be anticipated. Then, it may be worthwhile to deliberately increase power output from the high-cost generator and ramp down the low-cost generator prior to the onset of increased wind-power generation. This enables the wind to be more fully used as soon as it is available. Similarly, preparations could be made for the drops in wind-power generation.

2.1 The square-wave wind-power generation model

If we assume that the potential wind-power generation is a square-wave of period $\tau$ we can conduct an exact analysis of the optimal strategy. Of course the actual function will be much more complicated but this simpler form still enables us to see some of the effects of variation. For time duration $\beta$ the wind power $P_w$ is at its upper value $P_{w\text{max}}$ and for $\tau - \beta$ it is at its lower value $P_{w\text{min}}$. If the maximum wind-power generation $P_{w\text{max}}$ is always below the load $L$, then the low-cost generator permanently meets the difference $L - P_{w\text{max}}$. So to further simplify the problem, without losing the features of primary interest, we assume $P_{w\text{max}} = L$ and likewise $P_{w\text{min}} = 0$, so that

$$P_w(t) = \begin{cases} 
L, & \text{if } 0 < t < \beta \\
0, & \text{if } \beta < t < \tau
\end{cases} \quad P_w(t + \tau) = P_w(t). \quad (1)$$

First, consider what happens when the low-cost generator changes in direct response to the wind-power fluctuations. The maximum wind power used is $h$ and the remainder $L - h$ of the load is always met by the low-cost generator. This simple reactive model is illustrated in Figure 2(a) for a case where the slow-ramping is more rapid than the fluctuations. Specifically, $h/r < \beta$ and $h/r < \tau - \beta$. With the square-wave of equation (1), the wind power used $P_u$, 

Figure 2: The square-wave model. The upper part of each diagram shows the square wave output of the wind farm. The lower part shows how wind (green), low-cost (grey) and high-cost (red) generators meet the load.
low-cost power \( P_l \) and high-cost power \( P_h \) are given by

\[
\begin{align*}
 [P_u, P_l, P_h](t) &= \begin{cases}
 \left[ \tau t, L - \tau t, 0 \right], & \text{if } 0 < t < h/\tau \\
 \left[ h, L - h, 0 \right], & \text{if } h/\tau < t < \beta \\
 [0, L - h + r(t - \beta), h - r(t - \beta)], & \text{if } \beta < t < \beta + h/\tau \\
 [0, L, 0], & \text{if } \beta + h/\tau < t < \tau 
\end{cases}

de (2)
\end{align*}
\]

Initially, there is no wind-power generation and the load is entirely met by the low-cost generator. As there is potential for wind-power generation the low-cost generator ramps down to allow the demand to be met by a mixture of wind and low-cost power. When the wind-power generation ceases again the high-cost power generator must initially meet the demand while the low-cost generator ramps up to meet the full load. This cycle is repeated. In order to determine the optimum value of \( h \), we calculate the total cost per time period \( \tau \), which is a quadratic,

\[
c_l (L \tau - h \beta) + c_h h^2 / (2 \tau).
\]

The minimum for this is readily found to be \( h = r \beta c_l / c_h \), which corresponds to a total cost:

\[
c_l \tau - c_l^2 \beta^2 \tau / (2 c_h). \quad (4)
\]

If the difference in cost between low-cost and high-cost generators is too small this may not be the cheapest strategy: the combination of power sources switching directly between wind and high-cost power may be collectively cheaper than low-cost power \( c_l \tau < c_h (\tau - \beta) \). A further minor complication for this reactive model occurs when there are ranges of \( h \) for which the slow ramping is too slow for the low-cost generator to fully decrease power by \( h \) during the bursts of wind (of duration \( \beta \)), or to fully increase power by \( h \) during the lulls in the wind (of duration \( \tau - \beta \)) or both of these cases. That is, if either \( h/\tau > \beta \) or \( h/\tau > \tau - \beta \), or both. For these ranges of \( h \) the total cost becomes a linear function and so reaches its extreme values at the limits of the domain.

In the limiting case there is no ramping of the low cost generator. (This is used in the analysis in Section 3.) Then the total cost function is linear over the entire range of \( h \) from 0 to \( L \) and so the optimum value of \( h \) will be either 0 or \( L \). When \( h = 0 \), the entire power production is by the low-cost generator and will cost \( c_l \tau \) over the period \( \tau \). When \( h = L \), this low-cost generator is never used, the production being entirely by the wind generator when this is operating or by the higher cost expensive generator when it is not. The total cost for this case is \( c_h L (\tau - \beta) \) per period and so \( h = 0 \) is cheaper than \( h = L \) when \( c_l \tau < c_h (\tau - \beta) \).

The linearity, for no ramping, has resulted in the lowest cost option occurring at one of the extremes of either having no wind generation or no low-cost generation. However, for more general wind forms, this will not be the case. For example, if, instead of the square wave form for wind power, we use a triangular wave form, then the optimal value of \( h \) in the no-ramping case is again given by a quadratic expression.

This simple square-wave model also illustrates the possible savings if changes in wind generation can be predicted. For example suppose that we anticipate the rises in wind power (but not the falls, although their anticipation can also be used for savings (Section 2.3)). In this case we may increase power generation from the high-cost generator and ramp down the low-cost generator prior to the onset of the increased wind generation (Figure 2(b)). This enables the wind to be more fully used as soon as it is available. We will also suppose that
the slow-ramping is fast enough to complete between changes in the wind-power generation and that the high-cost power is sufficiently costly to justify ramping. Let the slow-ramping down of the low-cost generator begin at \( r/h_2 \) before the onset of the wind-power generation so that at the beginning of wind-power generation the low-cost generator is producing \( L - h_2 \) power. Then the cost function equation (3) is modified to

\[
c_f(Lr - h\beta - rh_2/r) + c_h(h^2 + h_2^2)/(2r). \tag{5}
\]

This is a quadratic form in the two parameters \( h \) and \( h_2 \). The minimum cost now occurs at the higher value of \( h = r\beta c_f/(c_h - c_l^2/c_h) \) with \( h_2 = c_l h/c_h \). Substituting this into equation (5) above we see that the second (negative) term in the total cost (equation (4)) has increased in magnitude by the same factor \((1/(1 - c_l^2/c_h^2))\) as \( h \). So the total cost with anticipation is

\[
c_fLr - c_l^2\beta^2r/(2(c_h - c_l^2/c_h)). \tag{6}
\]

This is a saving of \( c_l^2\beta^2r/(2c_h(c_h^2 - c_l^2)) \) per period \( r \) over not anticipating the wind.

2.2 Numerical simulations

The three-power-station model was simulated for the reactive case. For each simulation a constant minimum level of low-cost generation was maintained. As far as possible, wind
generation is used to meet the remaining load $h$. If, during operation, low-cost and wind outputs together are insufficient to meet the load then high-cost power is used to meet the balance while the low-cost generation ramps up to compensate. If, instead, the wind and low-cost power is greater than the load, then wind is spilled and the low-cost generator ramps down (unless it is already at its minimum level).

![Graphs showing power contributions and headroom](image)

Figure 4: The four graphs on the left show the components of different power sources used to meet the load and the wind-power spilled for a sample wind-speed time series. The curve on the right shows the relationship between average total cost and $h$ for an illustrative set of power prices.

The simulations were conducted with $h$ values from 1% to 100% of the load. For each value of $h$ there were 100 random realisations of the realistic wind speed time-series generated from the model

$$W(i + 1) = 0.99W(i) + 2N(0, 1)$$

(7)

where the $W(i)$ are the values of the wind at discrete time intervals and $N(0, 1)$ is a Gaussian random sample [2]. The power generation response $P_g$ to wind speed $W$ is not linear (although this was used initially as a first approximation). During the initial rise in the wind the
response is cubic but for high winds the power generation is bounded and for very high winds
the wind must be spilled to protect the generator and so there is no power output. The
power-generation response used takes the form

\[
P_w = \begin{cases} 
W^{3} P_{\text{max}}, & \text{if } 0 < W < W_{\text{limit}} \\
P_{\text{max}}, & \text{if } W_{\text{limit}} < W < W_{\text{max}} \\
0, & \text{if } W > W_{\text{max}} 
\end{cases}
\]

which is typical of a single wind generator. The response curve is shown in Figure 3 together
with a sample wind speed time series and its power output.

The cost of meeting the load's power demand is found for each realisation. These costs
were averaged over the 100 simulations for each of the different strategies. This average is shown
in Figure 4. Again the relationship between cost and \( h \) is a U-shaped curve for which we have
an optimal \( h \). Numerical simulations of this kind can be readily extended and adjusted for
more sophisticated models, as in Section 2.3.

2.3 Optimum settings for accommodating variation in wind power

Here we revisit the problem of choosing optimum settings for accommodating variation in
wind-power for more realistic wind power forms which essentially change at random with the
available wind. A step down in wind power needs to be replaced by other power, whereas a
step up in wind power need not be used. For this reason it is necessary to consider separately
the cases of a step down and a step up. This is done in the next two subsections.

Figure 5: Power sources during a drop in wind power: (a) (left) no additional low-cost power;
(b) (right) additional low-cost power replaces some wind power, reducing the need for high-
cost power. The top diagram shows the total power output composition and the lower diagram
shows the individual power outputs. (Green - wind power used; Cyan - unused wind power;
Blue - low-cost power; Red - high-cost power; Magenta - available wind power).
2.3.1 Step change down in wind power \((\Delta P_w < 0)\)

First, we consider the use of additional low-cost power to provide a buffer for the variation in wind power. This low-cost power replaces some of the wind power, reducing the amount of wind power used. However, it also reduces the need for high-cost power immediately following a drop in the wind.

Figure 5 shows the effect of a step down in wind power. A portion of high-cost power (red) is needed to replace the wind power until the low-cost generator can ramp up to take over from the reduced wind power. In Figure 5(a), the wind power is fully used, whereas in Figure 5(b), a higher level of low-cost power is used and this replaces a portion of the wind power. For the expense of replacing some of the wind power with low-cost power, the requirement for high-cost power is significantly reduced.

Thus additional low-cost power \(\Delta P_l\) is used to accommodate changes downwards in wind power \(\Delta P_w\) at the expense of not using all the available wind power during normal operation. Initially we assume one step change down in the period \(T\). It is noted that although there are continuing random changes in the wind the expected change is close to zero. So, we initially assume that the average of wind changes following the step change is zero.

There is no extra benefit from additional low-cost power above \(-\Delta P_w\) and so this case can be ignored. The total cost is:

\[
c_l \int_0^T P_l(t)dt + c_l \Delta P_l T + \frac{c_h}{2r} (-\Delta P_w - \Delta P_l)^2.
\]  

(9)

To find the minimum cost, we differentiate with respect to \(\Delta P_l\) and equate this to zero to obtain:

\[
\Delta P_l = -\Delta P_w - r T c_l / c_h.
\]  

(10)

Thus the optimum \(\Delta P_l\) is always less than \(-\Delta P_w\) (which is positive), and is limited by zero when the second term becomes too large. If high-cost power is sufficiently expensive then there will be an advantage in using additional low-cost power instead of wind power. The optimum amount of excess low-cost power also depends on the time interval \(T\) used, with longer time intervals indicating less excess low-cost power. The time interval used could also include a step in the opposite direction as considered in the next subsection.

Instead of assuming a fixed size of step change over the time interval \(T\), now we assume that for step changes below \(-\Delta P_l\) (negative and of a greater magnitude) the size is determined by the probability:

\[
P(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_l, T) .
\]

Rather than equation (9) the following total cost now applies:

\[
c_l \int_0^T P_l(t)dt + c_l \Delta P_l T + \\
\frac{c_h}{2r} \int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_l, T) (-\Delta P_w(u) - \Delta P_l)^2 du.
\]  

(11)

To find the minimum cost, differentiate with respect to \(\Delta P_l\) (to get a simpler but approximate result, assume that the probability term is essentially constant, a method for correction of this and other approximations by adjusting the value of \(T\) is suggested in a later subsection):

\[
c_l T - c_h / r \int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_l, T) (-\Delta P_w(u) - \Delta P_l) du.
\]  

(12)
Set this to zero and solve for $\Delta P_l$:

$$
\Delta P_l = -\frac{\int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_l, T))\Delta P_w(u)du - rTc_l/c_h}{\int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_l, T))du}. 
$$

(13)

After dividing by the denominator, the first term gives the average size of the larger step changes, while the second term is adjusted to account for the probability of a step change down occurring. This is unfortunately an implicit equation for $\Delta P_l$, but as it involves a single variable it is easily solved numerically for particular cases.

As noted above there is a problem in the choice of the value of $T$. This is considered in Section 2.3.4.

Figure 6: Power sources during a wind power increase; (a) (left) no high-cost power; (b) (right) some high-cost power replaces low-cost power. Diagrams are as in Figure 5 (Green - wind power used; Cyan - unused wind power; Blue - low-cost power; Red - high-cost power; Magenta - available wind power).

2.3.2 Step up change in wind power ($\Delta P_w > 0$)

High-cost power (which can be altered rapidly) could be used to replace some low-cost power to enable more rapid utilisation of an increase in wind power. Figure 6 shows the effect of a step up in wind power. In Figure 6(a) some of the wind power cannot be used as the low-cost generator cannot adjust sufficiently fast. Figure 6(b) shows the same increase in wind power but with some high-cost power replacing part of the low-cost power. Now it can be seen that more of the wind power is used.

Using more high-cost power than the step in wind power is clearly a waste. The total cost when $\Delta P_h < \Delta P_w$ is:

$$
c_l \int_0^T P_l(t)dt - c_l\Delta P_h T + c_h\Delta P_h T - \frac{c_h}{2r} \Delta P_w^2 + \frac{c_h}{2r} (\Delta P_w - \Delta P_h)^2. 
$$

(14)
As before, the value of $\Delta P_h$ at the minimum is found by differentiation:

$$
\Delta P_h = \Delta P_w - r(c_h - c_l)T/c_h.
$$

(15)

This always gives $\Delta P_h < \Delta P_w$ when $c_h > c_l$. When the difference between high and low-cost power is not too large, high-cost power can be used to take advantage of rapid increases in wind-power generation.

If we consider a cycle of a step up then a step down with both allowing time for the low-cost power to adjust so that $T > 2\Delta P_w/r$. Then by applying this inequality to equation (15) we have

$$
\Delta P_h < \Delta P_w(2c_l/c_h - 1).
$$

(16)

Thus it is seen that for $\Delta P_h$ to be positive we should have $c_h < 2c_l$.

With a probability distribution for $\Delta P_w(u)$ we have:

$$
\Delta P_h = \frac{\int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < \Delta P_h,T)\Delta P_w(u)du - r(c_h - c_l)T/c_h}{\int_0^1 P(\Delta P_w(u) : \Delta P_w(u) < \Delta P_h,T)du}.
$$

(17)

Thus $\Delta P_h$ is always less than a weighted sum of the $\Delta P_w$ values.

The quantity $\Delta P_h$ is the reduction for the target power of the low-cost generator. The low-cost generator moves towards its target at its limited slow ramp rate. The actual power allocation will be the amount currently generated by the low-cost generator, plus either the wind power needed to reach the required total power, or, if this is not sufficient, the maximum wind power with the remainder coming from the high-cost generator. Again, selecting the value of $T$ is a problem.

2.3.3 Simulation of wind power generation

Sections 2.3.1 and 2.3.2 have provided formulae that show how extra amounts of low and high-cost power can help reduce the expense of wind power variation management. Equations (13) and (17) can be used for a given value of $T$ to determine the amounts of additional power to be generated. (It can be easily verified that these two formulae act independently.) These formulae will now be tested for a numerical simulation similar to that of Section 2.2.

A table can be constructed giving the values of $\Delta P_l$ and $\Delta P_h$ for for each wind speed. For every case the distribution of wind speeds after the time interval $T$ is also calculated. For the demonstration here a normal distribution with standard deviation 3.0 is used as $T$ is assumed small enough for this to be sufficiently accurate. (Another distribution of changes in wind speed could be used if thought appropriate.) From the distribution of wind speeds, and a power versus wind speed relation, the distribution of power after time interval $T$ can be determined. The power relation, typical of a single wind generator, can be seen as the magenta curve in Figure 7:

$$
P_w = \begin{cases} 
30(W/15)^3, & \text{if } W < 15 \\
30, & \text{if } 15 \leq W \leq 25 \\
0, & \text{if } W > 25.
\end{cases}
$$

(18)

Once the distribution of power changes is obtained, equations (13) and (17) can be solved to obtain target values of $\Delta P_l$ and $\Delta P_h$ as illustrated in Figure 7.
Figure 7: Targets for power generation given wind speed: Magenta - potential wind power; Green - wind power target; Cyan - unused wind power; Blue - low-cost power target; Red - high-cost power target. This illustration is for the values: $L = 50$; $r = 2$; $c_L = 40$; $c_h = 80$; $T = 2$.

Due to the slow dynamic performance of the low-cost generator, it may not be able to provide the specified target value for a given wind speed. In this case any deficit is made up first by wind power, if available, and then by high-cost power, while the low-cost generator is ramped towards the target.

Given the table of target values, a simulation can be performed of the wind generation over time. The formula for the wind speed:

$$W_i = \alpha W_{i-1} + 3 N(0, 1),$$

is similar to that given in [2]. We chose $\alpha$ (0.9798) to give 15 as the standard deviation of the wind speed. (A Fourier Analysis was done of wind speed data in [2]. The Power Spectrum had two clear peaks corresponding to 24 and 12 hour periods and a possible indication of a component with an 8 hour period. The wind speed models, equations (7) and (19), in part arise from this analysis. That paper also found several different forms for the daily variation, rather than a typical average behaviour.)

Figure 8 shows a portion of the simulation results. The variation in wind speed is seen to be amplified in the wind power $P_w$. These variations then drive the variations in the other power values.

2.3.4 Determination of the time interval

There remains the value of $T$ to be determined. This can be done by a direct search using simulated wind speeds. Ten thousand time steps were used with the same random wind profile.
\[ c_l = 40, \ c_h = 60 \]

<table>
<thead>
<tr>
<th>( T(\Delta P_h) )</th>
<th>( T(\Delta P_L) )</th>
<th>2</th>
<th>8</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1773.6</td>
<td>1765.7</td>
<td>1765.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1779.0</td>
<td>1770.1</td>
<td>1769.5</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1780.3</td>
<td>1771.6</td>
<td>1771.0</td>
<td></td>
</tr>
</tbody>
</table>

\[ c_l = 40, \ c_h = 80 \]

<table>
<thead>
<tr>
<th>( T(\Delta P_h) )</th>
<th>( T(\Delta P_L) )</th>
<th>2</th>
<th>8</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1843.4</td>
<td>1843.1</td>
<td>1843.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1836.5</td>
<td>1835.7</td>
<td>1835.9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1836.7</td>
<td>1836.2</td>
<td>1836.5</td>
<td></td>
</tr>
</tbody>
</table>

\[ c_l = 40, \ c_h = 100 \]

<table>
<thead>
<tr>
<th>( T(\Delta P_h) )</th>
<th>( T(\Delta P_L) )</th>
<th>2</th>
<th>8</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1902.2</td>
<td>1912.3</td>
<td>1914.4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1888.8</td>
<td>1898.2</td>
<td>1900.9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1888.7</td>
<td>1898.2</td>
<td>1901.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Cost of generation for different power costs and \( T \) values

for each run.

Table 1 gives the effect of a range of values for \( T \) and for the cost of high-cost power. The value of \( T \) for a step up in wind power has been assumed independent to that for a step down. Except where noted the same values as in the previous subsection have been used. In this table it should be noted that a higher value of \( T \) corresponds to reduced additional power.

In the first case the cost of high-cost power is not much more than that of the low-cost power. It is seen here that low values of \( T(\Delta P_h) \) give the lowest cost, corresponding to an addition of high-cost power. Additional low-cost power (low \( T(\Delta P_L) \)) increases the costs in this case.

The second case has the high-cost power twice the cost of the low-cost power. Here we see the effect of the additional low-cost power (low \( T(\Delta P_L) \)) is close to zero, but too much high-cost power (low \( T(\Delta P_h) \)) increases cost.

The final case has a higher value for the high-cost power. Here we see the advantage of using additional low-cost power to avoid the use of the high-cost power. Additional high-cost power (low \( T(\Delta P_h) \)) increases the costs.

### 2.3.5 Additional comments

This illustration uses the power curve for a single wind generator. The largest adjustment to power and hence costs, are seen close to the steepest parts of the wind power curve (Figure 7). In practice the power output of a wind farm will be the total from multiple wind generators and so will be smoother than that used here. This will reduce the benefits to be gained.
Figure 8: Simulation results showing the effect of variation in wind speed on power generation

In some cases the high-cost power cost is large, and then there are significant benefits in scheduling a margin of low-cost power. Avoiding high-cost power here decreases the marginal power costs that determine what customers pay.

By using additional low-cost power to buffer possible drops in wind generation, the wind farm is being penalised for the variability of its supply by not having all the available power scheduled for use. Replacing some low-cost power with high-cost, to better utilise possible increases in wind power, is only of benefit when the costs of high and low-cost powers are close.

3 Optimal dispatch with simple ramp rates

In this section again, as earlier, we make the important simplifying assumption that wind farms share the network with power stations of two kinds: fast and slow. The fast stations
effectively have an infinite ramp rate as before; they can quickly be re-dispatched in response to any fluctuations in the output of wind farms. However, here, the slow stations are taken to have essentially zero ramp rate; their dispatch, once chosen at time \( t = 0 \), may not be altered during the time period \( 0 \leq t \leq T \) considered. Broadly speaking, the “fast” behaviour is an idealized model of a hydroelectric station, while the “slow” behaviour models a thermal power station.

In such a model, the conventional least-cost dispatch serves only to minimize the instantaneous cost at time \( t = 0 \). However, the true objective is to minimize the expected overall cost over the period \([0, T]\), allowing for the adjustments that will become necessary as the wind fluctuates. (From the perspective of \( t = 0 \), the actual cost to be incurred is a random variable, due to stochasticity in the wind; its expectation value is our objective.) To minimize cost in this sense, a different initial dispatch may be required.

### 3.1 A one-node model

Again we begin by considering an example in which all power is effectively generated and consumed at a single location (Figure 9).

A total load of 200 must be met from the following generator offers. A wind farm offers a time-varying quantity \( P_w(t) \), at zero price (marginal cost). We suppose that \( 0 \leq P_w(t) \leq 100 \) and \( P_w(0) = 50 \). A thermal station offers 100 at price \( c_T \), and a hydro generator offers tranches of 100 at price \( c_1 \) and a further 100 at price \( c_2 \). The offer prices satisfy \( 0 < c_1 < c_T < c_2 \).

The decision to be made at time \( t = 0 \) is essentially the quantity \( x \) to dispatch from the thermal station. At subsequent times \( t \in [0, T] \), the power required will be made up as follows: \( x \) from the thermal, \( P_w(t) \) from the wind farm, and the balance from the hydro.

The least-initial-cost solution would be \( x = 50 \), i.e. dispatch the wind farm, the cheaper of the two hydro tranches, and (partially) the thermal. If the available wind power should subsequently rise, it will displace \( H_1 \) hydro generation, saving water with value \( c_1 \). In the event of a decrease in available wind power, hydro generation from \( H_2 \) will be used to make up the shortfall.

However, if \( c_2 \) is very much greater than the other prices, this seems like a poor solution. One wonders if it might be better, in that case, to choose a higher initial thermal dispatch, and reserve some of the \( H_1 \) tranche for possible future use. On the other hand, if \( c_1 \) is very low (perhaps almost zero) and \( c_2 \) is not prohibitively large, a good initial dispatch might
include some of \( H_2 \) (as well as all of \( H_1 \)), so as to increase the likelihood of fully using the cheap water. Note that both of these situations involve “out-of-order” dispatching, in which an apparently expensive tranche is dispatched while a cheaper one is not. In fact, both can occur, as the following analysis shows.

Formally, the instantaneous cost of the generation at time \( t \) is

\[
e(x, P_w(t)) = \begin{cases} 
    x c_T + 100 c_1 + (100 - x - P_w(t)) c_2, & \text{if } x + P_w(t) \leq 100 \\
    x c_T + (200 - x - P_w(t)) c_1, & \text{if } x + P_w(t) \geq 100 .
\end{cases}
\]  

(20)

The instantaneous marginal cost is found by differentiating with respect to \( x \):

\[
c_x(x, P_w(t)) = \begin{cases} 
    c_T - c_2, & \text{if } x + P_w(t) < 100 \\
    c_T - c_1, & \text{if } x + P_w(t) > 100 .
\end{cases}
\]  

(21)

We want to minimize the expected average cost over the period \( 0 \leq t \leq T \), which is:

\[
C(x) = E \left[ \frac{1}{T} \int_0^T c(x, P_w(t)) \, dt \right].
\]  

(22)

This expression involves averaging over both time (the integral) and a probability space (the expectation). We can make the notation less cumbersome by defining \( \tau \) to be a random variable distributed uniformly on \([0, T]\), independently of \( P_w(t) \), and \( V = P_w(\tau) \). Then

\[
C(x) = E \left[ c(x, V) \right].
\]  

(23)

When \( V \) has a well-behaved distribution, we may differentiate inside the expectation to obtain the marginal cost with respect to \( x \):

\[
C'(x) = E \left[ c_x(x, V) \right] = (c_T - c_2) P(V \leq 100 - x) + (c_T - c_1) P(V \geq 100 - x) = (c_T - c_1) - (c_2 - c_1) P(V \leq 100 - x).
\]

This is clearly an increasing function of \( x \), and so \( C(x) \) is convex in \( x \). The optimal \( x \) can be found by setting this expression equal to 0.

For a more specific example, assume that \( P_w(t) \) follows a Brownian motion model, so that \( P_w(t) \sim N(P_w(0), \sigma^2 t) \). (We neglect the bounds \( 0 \leq P_w(t) \leq 100 \).) Then \( V \sim N(P_w(0), \frac{1}{2} \sigma^2 T) \). Solving gives an optimal initial dispatch

\[
x^* = P_w(0) - \left( \frac{\sigma^2 T}{2} \right)^{1/2} \Phi^{-1} \left( \frac{c_T - c_1}{c_2 - c_1} \right),
\]  

(24)

where \( \Phi \) is the standard normal cumulative probability distribution function. This expression may be either more or less than the least-initial-cost solution (\( x = 50 \), here), depending on the relative costs of the various tranches into play.

### 3.2 Line reservations

The line reservation technique considers only initial dispatches that arise in a particular way: by reserving capacity on transmission lines. That is, we solve a dispatch problem which differs from the conventional one only in that the assumed line capacities may be less than the actual
Figure 10: The two-node network.

capacities. The “reserved” line capacity is then available to accommodate fluctuations in the wind during the period $0 < t \leq T$.

The line-reservation technique is only a heuristic method, and the initial dispatch found in this way may not be optimal for our real objective (that is, another initial dispatch may have lower expected overall cost over the period $[0, T]$). Notice, for example, that line reservations would be of no help with the problem in the previous subsection, in which there are no lines. However, it does at least have the advantage of being easy to implement.

### 3.3 Wind matching over a single line

We shall now consider networks in which the power stations and the load are each located at one of a number of nodes which are interconnected by transmission lines of finite capacity. We begin with the wind-matching problem on a two-node network (Figure 10). In this network, a lossless transmission line of capacity $M$ connects two nodes; a load $L$ is taken off at the right-hand node. Generators offer power as follows: at the left-hand node, a wind farm offers a time-varying quantity $P_w(t)$, at zero price (marginal cost). A thermal station offers unlimited quantities at price $c_a$. At the right-hand node, a hydro station offers a relatively cheap tranche of quantity $P_H$ at price $c_b$, and a more expensive tranche at price $c_c$. Another thermal station offers unlimited quantities at price $c_d$. We assume $M < L$, $0 < c_a < c_d$, and $0 < c_b < c_d < c_c$.

The least-initial-cost solution would dispatch the generators in the order of their offer prices, starting with the cheapest. This might mean, for example, dispatching the wind and $H_1$ hydro tranches, then Thermal 1 to the extent allowed by the transmission capacity, and finally Thermal 2 for the remainder. This leaves the transmission line constrained, with no capacity to spare. If the wind output should subsequently rise above its initial level, the excess wind power must be spilled, as there is no way to make use of it.

An alternative approach would be to leave some “headroom”, via a reduction $x$ in the quantity dispatched from Thermal 1 and a corresponding increase in the quantity dispatched from Thermal 2. This leaves unused capacity $x$ in the line. Any subsequent rise in available wind power (up to $x$ above the initial level) can then be used to displace hydro generation from $H_1$, saving water with value $c_b$. Only if the wind rises above its initial level by more
than $x$ must the excess be spilled.

Note that under either solution, a decrease in the wind farm’s output must be compensated for by increased hydro generation from the $H_2$ tranche. If $c_e$ is large, the use of $H_2$ water might represent a sizable contribution to the overall expected cost. However, having reserved line capacity does not help with this situation. (One would instead need to reserve spare capacity directly within the $H_1$ hydro tranche, as in the example in Section 3.1.)

Determining the amount of headroom to leave requires solving an optimization problem. Suppose a decision is made at time $t = 0$ to leave an initial headroom $x \geq 0$, by shifting generation from Thermal 1 to Thermal 2. This commits the system to an ongoing additional cost (per unit time) of $(c_d - c_a)x$, relative to the least-initial-cost solution. At a time $t > 0$, the instantaneous further cost (or benefit if negative) due to wind shifts is

$$f(t) = \begin{cases} (P_w(0) - P_w(t))c_e, & \text{if } P_w(t) < P_w(0) \\ -(P_w(t) - P_w(0))c_b, & \text{if } P_w(0) \leq P_w(t) \leq P_w(0) + x \\ -xz_c, & \text{if } P_w(t) > P_w(0) + x, \end{cases}$$

or, more compactly,

$$f(t) = c_e(P_w(0) - P_w(t))_+ - c_b \min(x, (P_w(t) - P_w(0))_+),$$

where the notation $z_+$ denotes $\max(z, 0)$. Thus at time 0, the expected average cost per unit time incurred over the time interval $0 \leq t \leq T$ is:

$$C(x) = \mathbb{E}\left[\frac{1}{T} \int_0^T ((c_d - c_a)x + f(t)) \, dt\right] = \mathbb{E}\left[(c_d - c_a)x + c_a(-\delta)_+ - c_b \min(x, \delta_+)\right],$$

where $\delta = P_w(\tau) - P_w(0)$, with $\tau$, as before, a random variable independent of $(P_w(t))$ distributed uniformly on $[0, T]$.

Suppose we wish to choose $x$ so as to minimize $C(x)$. Assume for simplicity that $P_w(t)$ has a continuous probability distribution. We see that for $x > 0$,

$$\frac{d}{dx} C(x) = (c_d - c_a) - c_b \mathbb{E}\left[\frac{d}{dx} \min(x, \delta_+)\right] = (c_d - c_a) - c_b \mathbb{E}[1_{\delta > x}] = (c_d - c_a) - c_b \Pr(\delta > x).$$

It is clear that this is a continuous increasing function of $x$, and hence that $C(x)$ is a smooth convex function of $x$. If $\Pr(\delta > 0) \leq (c_d - c_a)/c_b$ then $C(x)$ increases with $x$ for all $x > 0$, and so the optimal choice is $x = 0$. Otherwise, the optimal $x$ can be found by solving the equation

$$\Pr(\delta > x) = (c_d - c_a)/c_b.$$  \hspace{1cm} (27)

In some situations (e.g. when $P_w(0)$ is neither 0 nor the maximum output of the wind farm, but somewhere in between), it may be reasonable to suppose that $\Pr(\delta > 0) = \frac{1}{2}$. (In other words, the wind is as likely to rise as to fall.) In this case, it is worth reserving headroom only if

$$c_d - c_a < c_b/2.$$  \hspace{1cm} (28)

One can see this intuitively: the first unit of headroom costs $c_d - c_a$ to create, whereas the average cost of not having it is $c_b$ for half of the time.
3.4 Wind matching and sloshing in a network loop.

We now apply a similar analysis to that in Section 3.3 to a more complicated problem (Figure 11). Here we have two wind farms sharing a six-node loop network with two slow stations and a fast hydro. The only transmission capacity constraint of significance is the maximum flow of 150 on the dashed line in the diagram.

The six lines are assumed to be lossless, but with equal reactances. This means that power flowing from any station to the load will divide itself between the two possible paths in inverse proportion to the number of lines traversed. That is, \( \frac{5}{6} \) of the power from Thermal 1 flows to the load via the limited-capacity line, while \( \frac{1}{6} \) flows via the other five lines. For Wind 1, the flows divide in proportion \( \left( \frac{2}{3}, \frac{1}{3} \right) \); for Thermal 2, \( \left( \frac{1}{2}, \frac{1}{2} \right) \); for Wind 2, \( \left( \frac{3}{5}, \frac{2}{5} \right) \). The Hydro has the most advantageous position: only \( \frac{1}{6} \) of its power flows over the limited-capacity line, with the other \( \frac{5}{6} \) taking the more direct route. The network flows due to different power stations may be linearly superposed.

Let us suppose that the initial wind outputs are \( P_{w1}(0) = P_{w2}(0) = 60 \). The corresponding least-initial-cost dispatch is perhaps not apparent by inspection, but it can be found by solving a linear program. If we let \( P_1, P_2, P_3, P_4, P_5, \) and \( P_6 \) denote the quantities dispatched from, respectively, Thermal 1, Wind 1, Thermal 2, Wind 2, Hydro \( H_1 \), and Hydro \( H_2 \), then the linear program is

\[
\begin{align*}
\text{min} & \quad 40P_1 + 45P_3 + 48P_5 + 60P_6 \\
\text{s.t.} & \quad P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 300 \\
& \quad \frac{5}{6}P_1 + \frac{2}{3}P_2 + \frac{1}{2}P_3 + \frac{1}{3}P_4 + \frac{1}{6}(P_5 + P_6) \leq 150 \\
& \quad 0 \leq P_1 \leq 100, \quad 0 \leq P_2 \leq 60, \\
& \quad 0 \leq P_3 \leq 200, \quad 0 \leq P_4 \leq 60, \\
& \quad 0 \leq P_5 \leq 60, \quad 0 \leq P_6 \leq 90.
\end{align*}
\]
The solution is $P_1 = P_2 = P_3 = P_4 = P_5 = 60$ and $P_6 = 0$. (Note that even though the thermal offers are cheaper than the hydro, the transmission constraint has prevented their full use.)

However, this leaves the limited-capacity line constrained, which will make it difficult to accommodate subsequent fluctuations in wind power. If the output of either wind farm (or both together) should increase, the extra wind power cannot be used to displace hydro power from $H_1$, as this would increase the flow on our constrained line. Even a slosh from Wind 2 to Wind 1 (i.e. a decrease in output at Wind 2, and an equal increase at Wind 1) could not be dealt with, although a slosh from Wind 1 to Wind 2 could be.

It might, perhaps, be a better idea to choose a different initial dispatch which does not put the network into such a constrained configuration. Suppose we require the initial dispatch to reserve spare capacity (headroom) $x$ in the limited-capacity line, for later use in responding to changes in the wind. The best way to achieve this can be found by changing the right-hand-side of the inequality constraint to $150 - x$ in the dispatch problem, and re-solving. Furthermore, the additional cost (at the margin) of creating the headroom can also be read off: it is the value at optimality of the dual variable corresponding to the inequality constraint of the dispatch problem.

We will assume that $0 \leq x \leq 20$; to create each unit of headroom within this range requires shifting 3 units of generation from Thermal 1 to Thermal 2; the cost is thus $3 \times (45 - 40)$, or $15 per unit headroom per unit time. (To create more than 20 units of headroom, some generation must be moved from Thermal 2 to Hydro $H_3$; this requires further analysis, which we omit.)

Now consider the situation at time $t > 0$. The power available from the wind farms is now $P_{w1}(t)$ and $P_{w2}(t)$; we may adjust the Hydro dispatch, but not the Thermals, in response.
The optimal response is the solution to a new version of our linear program in which only the Wind and Hydro dispatches are variables, with the Thermal dispatches being constants. Less formally, we can observe that each unit of headroom on our limited-capacity line allows the network to carry 2 units of extra power from Wind 1, all else being equal. (The extra wind power displaces hydro power, effectively requiring a new flow from Wind 1 to the Hydro to be superposed onto the existing flows; half of this new flow travels via the limited-capacity line.) Similarly, each unit of headroom allows 6 units of hydro power to be displaced by wind power from Wind 2. The network will thus remain unconstrained provided

$$
\frac{1}{2} \Delta P_{w1}(t) + \frac{1}{6} \Delta P_{w2}(t) < x,
$$

(29)

where $\Delta P_{wi}(t) = P_{wi}(t) - P_{wi}(0)$ for $i = 1, 2$.

The responses are shown in Figure 12. The network is unconstrained in regions $R_1$ and $R_2$. In $R_1$, the total available wind power has declined ($\Delta P_{w1}(t) + \Delta P_{w2}(t) < 0$), requiring Hydro $H_2$ to be dispatched to make up the shortfall; in $R_2$, additional wind power has displaced generation from Hydro $H_1$. In regions $R_3$ and $R_4$, the network is unable to carry all of the available wind power, so some must be spilled. It is always better to spill wind at Wind 1 than at Wind 2, since the amount that must be spilled is smaller (by a factor of 3); the amount of usable wind power is thus constant along horizontal lines in these two regions. In $R_3$, there has been a decrease in hydro dispatch, while in $R_4$ there has been an increase. Note that in part of $R_4$ the total available wind power has increased, but the total usable wind power has decreased, due to sloshing from Wind 2 to Wind 1.

Let $\Delta P_w(t) = (\Delta P_{w1}(t), \Delta P_{w2}(t))$. The change since $t = 0$ in the wind power used by the system is

$$
u(x, \Delta P_w(t)) = \begin{cases} 
\Delta P_{w1}(t) + \Delta P_{w2}(t), & \text{if } \Delta P_w(t) \in R_1 \cup R_2 \\
2x + \frac{1}{2} \Delta P_{w2}(t), & \text{if } \Delta P_w(t) \in R_3 \cup R_4.
\end{cases}
$$

(30)

The instantaneous cost of generation at time $t$ (relative to the least-initial-cost solution) is thus

$$
c(x, \Delta P_w(t)) = 15x - \rho(u(x, \Delta P_w(t))),
$$

(31)

where

$$
\rho(z) = \begin{cases} 
48z, & \text{if } z \geq 0 \\
60z, & \text{if } z \leq 0.
\end{cases}
$$

(32)

The overall expected cost of reserving headroom $x$ for the time period $0 \leq t \leq T$ is then

$$
C(x) = E \left[ \frac{1}{T} \int_0^T c(x, \Delta P_w(t)) \, dt \right] = E[c(x, \delta)],
$$

(33)

where $\delta = \Delta P_w(\tau)$, with $\tau$, as usual, a random variable independent of $\Delta P_w(t)$ and uniformly distributed on $[0, T]$.

If $(\Delta P_{w1}(t), \Delta P_{w2}(t))$ has a continuous probability distribution, then

$$
C'(x) = E[c_x(x, \delta)]
$$

$$
= E \left[ 15 - \rho(u(x, \delta))u_x(x, \delta) \right]
$$

$$
= 15 - 2(48P(\delta \in R_3) + 60P(\delta \in R_4)).
$$

Note that $C'(x)$ is increasing in $x$, and hence $C(x)$ is convex in $x$. To find the optimal solution for $x$, we must attempt to solve the equation $C'(x) = 0$, that is, to find $x \in [0, 20]$ with

$$
2(48P(\delta \in R_3) + 60P(\delta \in R_4)) = 15.
$$

(34)

134
The left-hand side of this equation represents the marginal value of the water that might be saved by allowing an additional unit of headroom; the right-hand side the marginal cost of the headroom. A solution will fail to exist only if \( C'(0) > 0 \) (in which case the optimal solution is \( x = 0 \)) or \( C'(20) < 0 \). In the latter case, it may be that \( x = 20 \) is optimal, or it may be that even more than 20 units of headroom are called for.

4 Analysis of wind farm data

Data on wind velocity and power output, from existing and potential sites, is useful for two purposes. First, it is needed as input into models for parts of the grid that are affected by wind-farms. Secondly, it can provide understanding of how wind and power are related across sites, in time; it will inform us about sloshing, evening-out and possibly forecasting. During the MISG week, group members examined the wind speed data provided from the Tararua wind farm, and also the Garrad Hassan report [3]. The first item raises questions about the properties that wind speed data needs so that it is fit for purpose. The second raises questions about further analysis of the dataset that it uses.

4.1 Desirable qualities in wind-speed data

The main purpose for wind speed data is to estimate wind-power output, and to investigate how this rises and falls. Hence the need is for time-series that mimic the behaviour of a turbine (or farm). The wind meter or meters need to be in the right location and at a suitable height. The series also need to be collected at time intervals or by equipment that makes the power estimates from them behave like turbine output.

The Tararua data is measured at 10 minute intervals, and is very volatile over these intervals. The data was studied for a 21 day period, which gives 3024 ten-minute intervals, 144 per day. This data was converted to a smoothed time series, a day effect, and residuals. The smoothed series was calculated as the moving average of 144 values, and centred by averaging adjacent pairs of the moving average values. The day effect was calculated after subtracting the centred moving average.

The 21-day data is shown in Figure 13. The actual series contains: cycles of about three days in length, a small day effect and residuals that are much larger than the day effect and of similar size to the cycles. Figure 14 shows behaviour over one day. The five years of data could also be examined for an annual effect, but the MISG group did not have the resources for this.

The day effect appears to have two components: a smooth part that rises and falls in an autocorrelated way, so probably reflects a small but persistent effect over these 21 days, and a fluctuating part that one would expect in the means of 21 independent haphazard events.

The residuals show haphazard or volatile behaviour: short-term fluctuations from one reading to the next. They also show autocorrelated movements of length about half a day. The smoother has excluded these from the smoothed series of course. They need further investigation.

The main conclusion about this data is that the actual series has short-term fluctuations that are large compared with the features of real interest. For future data collections, these would need to be removed either by meter design or by logging data at short intervals and then processing it to separate the components. In wind generation, we are interested in change; if
Figure 13: A 21-day part of the Tararua wind speed time series data; (a) (above) Blue - the actual data series; Magenta - a 24-hour smoothing of this; (b) (below) Magenta - the overall day (diurnal) effect repeated for each day; Blue - the residuals after subtracting the smoothed values and the day effect.
Figure 14: Data from a single day; (a) (above) Blue - wind speed; Magenta - a 24-hour smoothing of this; (b) (below) Differences: Blue - 10-minute intervals; Magenta - 30-minute intervals.
we look at change by differencing this series, the short-term fluctuations are even more of a nuisance (Figure 14).

The study helped clarify what we should look for in wind speed data and the issues include these: how does wind speed behave over time, and in particular how does it rise and fall? Hence the series of differences is important. Some of this behaviour may be predictable. We can expect (from meteorology) cycles of irregular lengths of a few days, daily cycles and annual cycles. These three features can be planned for. We would expect plenty of autocorrelation. Unfortunately, we also found plenty of short-term variability. This dominated the series of differences.

The study also clarified how data should be collected. A turbine has inertia, and will therefore smooth out short-term effects. A collection of turbines will have a further smoothing effect. There are two solutions. The first is to use a meter that mimics the behaviour of a turbine or collection; the second is to use a meter that measures velocity at short intervals (one minute or less), with its inherent volatility, but then smooth the data with a set of weights that mimics the smoothing effect of the turbine's inertia and size. Further smoothing could imitate the smoothing effect of the collection of turbines. The time-interval needs to be much shorter than the typical ramping-up or ramping-down time for a turbine.

In mimicking the behaviour of a turbine, we need, as well as smoothing, to transform speed into power. If speed and power are collected for an existing turbine, this data can be used immediately to plot the power/speed function for this type of turbine (at this location). In converting speed into power, we need to look for a lag effect: initially when wind-speed rises, some wind-energy may be accelerating the turbine rather than producing power output.

It could be useful to log wind-speed at a shorter interval (like one second), and then examine it for the short-term variability. This may occur in a particular frequency-band. The results would assist in design of data collection (instruments and intervals) for the future.

Variation and change is very important in wind generation, and leads to ramping of power output. If one site is more variable than usual, then it will be more complex and more expensive to manage substitutes for its down-times. Hence data collection design needs to enable analysis of variation.

4.2 Data for relationships among sites

The Garrad Hassan report uses a rich dataset that contains wind-speeds (and theoretical power outputs) at 10 minute and longer intervals, for 16 North Island and 5 South Island sites, for over a year. The report contains a thorough look at two aspects of this data: “correlations” between sites, and examination of events with large rates of change.

A common question at the workshop was “Are any of the correlations negative?” . In fact the report uses $R^2$ values, which measure the strength of the linear part of the relationships. These relationships can be investigated further, by looking for non-linearities and relationships with time-lags. A second aim could be to compare variability of sites, since that affects the usefulness of the site as an energy source. North and South Islands have separate parts of the grid, so relationships within islands are the main interest.

A data visualisation approach would involve steps like the following.

Firstly, produce the matrix of scatterplots, and join the dots. The scatterplots reveal much about the relationships (strength, direction [positive or negative], linearity or non-linearity, presence of clusters and outliers). The dot-joining will reveal something about the behaviour of the relationship over time, such as whether cycles in two series are in phase or out of phase.
Secondly, plot two, more, or all series against time together, and look for whether they move in phase together, out of phase, or independently. Then lag one series by a range of time intervals and check whether this strengthens or weakens the relationship.

In the above steps, it may be necessary to choose the time interval between readings, and the length of series used, so as to best reveal the behaviours.

The approach could provide useful information on how a set of wind-farms behave, so that this behaviour can be fed into models of parts of the grid.

5 Conclusions

The team working on this MISG problem have considered issues relating to electric power grid management for the case when wind power is a significant proportion of the total power generation. The topic is of importance as the proportion of wind power generation in New Zealand may significantly increase in the future. A number of approaches and simple models have been used to study aspects of production and transmission. There is scope for further extension of the work.

The three-power generator models illustrates the problems in maintaining power supply at a reasonable cost as the wind power rapidly varies. We have approached the problem of load balance considering the case when the low-cost alternative to wind can only be ramped slowly.

In our later approach the model is simplified further by taking the low-cost power output to be constant. More complicated networks of power stations have been considered which have further included the problem of managing limited capacity in power transmission lines.

The MISG group also considered the use of measured data for predictability and time profile of the wind as this is important for wind power production.

Although wind-power generation provides great opportunities for meeting energy requirements, its planning and management presents new challenges. When these have been met we will be able to more fully utilise this resource.

Acknowledgements

We are grateful to the industry representative Conrad Edwards for his support and enthusiasm. We also thank the other people who participated in this project who included Bob Anderson, Tim Duke, Tony Gibb, Roslyn Hickson, Sayed Hoseini and Phil Neame. We are grateful to the referee for helpful comments and suggestions.

References


139