– Proceedings –

43rd Annual Conference of the Operations Research Society of South Africa

14–17 September 2014
Stonehenge in Africa, Parys, South Africa

ISBN: 978-1-86822-656-6

Editorial Board

Editorial

It is a pleasure to present to you the proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa (ORSSA). The review process for the proceedings was as follows. Eighteen (18) manuscripts were submitted for possible inclusion in the proceedings. All submitted papers were double-blind peer-reviewed by at least two independent reviewers and in some instances even third and fourth opinions were obtained. Papers were reviewed according to the following criteria:

- Contribution to Operations Research, i.e. knowledge of field, significance of contribution, suitability for conference proceedings and quality and consistency of referencing.
- Technical quality, i.e. correct use of language, clarity of expression and quality and justification of arguments.
- General, i.e. clarity and quality of illustrations, usefulness of paper to OR practitioners, suitability and length of title, abstract and complete paper.

Of the eighteen (18) submitted papers, eleven (11) were ultimately, after consideration and incorporation of reviewer comments, judged to be suitable for inclusion in the proceedings - an acceptance rate of 61%. The proceedings will also be published at:

http://www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

I would like to thank the Associate Editor, Hans Ittmann, for his professional help and guidance during the whole process of reviewing and producing the final proceedings. I would also like to single out Fanie Terblanche who was a tremendous help and played a key role during the whole process. Fanie was responsible for a large number of administrative functions and for the typesetting of the final version of the proceedings - thanks for all your help and the professional management of the process.

No proceedings can be produced without reviewers or people who submit their work - thanks to all the reviewers who gave generously of their time and expertise to review all the submissions and also a special thank you to everybody who submitted their papers. Please keep on supporting ORSSA's conference and conference proceedings by submitting your research work again next year for possible inclusion in the proceedings. Finally, thanks to the Local Organizing Committee, the conference sponsors and all conference participants for making the conference and the conference proceedings a success. I hope that you all will enjoy reading the proceedings and that it may be useful in your future research endeavors.

Reviewers

The editorial would like to thank the following reviewers

R Bennetto	OPSI Systems, South Africa
M Bester	XTranda, South Africa
I Campbell	University of the Witwatersrand, South Africa
S Das	Council for Scientific and Industrial Research, South Africa
A De Villiers	Stellenbosch University, South Africa
I Durbach	University of Cape Town, South Africa
T Du Toit	North-West University, South Africa
M Fisher	Sasol, South Africa
J Holloway	Council for Scientific and Industrial Research, South Africa
H Ittmann	HWI Consulting, South Africa
D Jordaan	North-West University, South Africa
R Koen	Council for Scientific and Industrial Research, South Africa
H Kruger	North-West University, South Africa
J Kruger	University of South Africa, South Africa
M Kruger	North-West University, South Africa
D Lotter	Stellenbosch University, South Africa
H Nel	Stellenbosch University, South Africa
N Pillay	University of Kwazulu Natal, South Africa
H Raubenheimer	North-West University, South Africa
T Stewart	University of Cape Town, South Africa
F Terblanche	North-West University, South Africa
L Van Dyk	North-West University, South Africa
J Van Vuuren	Stellenbosch University, South Africa
E Willemse	University of Pretoria, South Africa

Best wishes,

Hennie Kruger
(e) hennie.kruger@nwu.ac.za
(t) +27 18 299 2539
Editor-in-chief: ORSSA Proceedings 2014
Operations Research Society of South Africa

Table of contents

Proceedings of the 2014 ORSSA Annual Conference
pp. 1–8
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Design of a detailed microscopic traffic simulation modelling framework for signalised intersections

AP Burger*, MD Einhorn† & JH van Vuuren‡

Abstract

There are numerous advantages to using simulation when investigating the effectiveness of novel traffic control strategies at signalised intersections. If the level of detail required for the investigation is not too demanding, a commercially available traffic simulation model may suffice. If, however, a high level of realism (such as the incorporation of explicit vehicle accelerations and decelerations, vehicle turning parameters and heterogeneous vehicle sizes) is required, it may be necessary to build a purpose-made traffic simulation model satisfying the specific requirements of the investigation. In this paper, a microscopic traffic simulation modelling framework is presented which may be employed as a stand-alone and customizable traffic simulation tool for testing the effectiveness of existing and novel traffic control algorithms, some of which require individual vehicle characteristics, such as vehicle speed and their positions on road segments as input data.

Key words: Microscopic, Simulation, Traffic, Model.

1 Introduction

Numerous strategies have been proposed in recent years for mitigating the debilitating effects of traffic congestion. One such approach, which is especially applicable to inner city commuting, is the attempted optimisation of traffic signal timings at signalised intersections. Improved and efficient signal timings have the ability to reduce driver delay times by effectively utilising intersection capacity and allowing for the formation and propagation of "green waves" (platoons of vehicles travelling unimpeded through several adjacent intersections displaying green signals). This reduces the stop-and-go driving patterns associated with congested traffic which drivers in Los Angeles, Mexico City, India, China, Singapore, and Johannesburg listed as their most serious commuter pain in the IBM 2011 Global Commuter Pain Survey [7].

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: apburger@sun.ac.za

†Corresponding author: Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: einhorn@sun.ac.za

‡**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

Before any novel traffic signal control strategies may, however, be implemented on public roads, their effectiveness and reliability should be tested extensively. Simulation modelling is a powerful tool which may be used in the design, implementation and evaluation of traffic signal control strategies. There are three distinct classes of traffic simulation models, *i.e. macroscopic, mesoscopic* and *microscopic models*. Macroscopic traffic simulation models are typically modelled from an aggregated point of view, based on a hydrodynamic analogy and regard traffic flows as a particular fluid process whose state is characterised by aggregate macroscopic variables such as density, volume and speed [3]. Mesoscopic traffic simulation models, on the other hand, have the ability to account for individual vehicles, but are still primarily concerned with traffic dynamics of the vehicles as a whole and do not explicitly consider the details of vehicle lane changing and vehicle following behaviour, nor changes in vehicle speeds [3, 8]. Finally, microscopic traffic simulation models explicitly account for individual vehicle motion characteristics (*i.e.* acceleration, deceleration and lane changes) and typically employ some form of *vehicle following model* [3]. In this paper, the design of a microscopic traffic simulation modelling framework is described. The framework is intended to be used for the investigation of novel self-organising traffic signal control algorithms which make use of live, real-time data associated with individual vehicles, such as vehicle speed and distance from an intersection, thus necessitating the accurate modelling of vehicle acceleration and deceleration, vehicle following distances, lane changes and turning profiles.

2 Simulation modelling paradigms

A simulation model is described by Banks *et al.* [2] as the imitation of a real-world process or system over time such that the behaviour of the system can be studied. If the model is a sufficiently realistic imitation of the real-world process, then data may be collected from this model as if it were collected directly from the real system under observation. Over time, simulation models have become extremely useful— almost indispensable, in fact — when analysing and verifying theoretical models which may be too difficult to analyse on a purely conceptual level [5]. When building a simulation model, there are four main distinguishable approaches that may be taken to replicate a real-world system. These four approaches are *system dynamics modelling, discrete event modelling, agent-based modelling* and *dynamic systems modelling* [1]. Because of space constraints, only agent-based modelling will be discussed further here as this is the method typically employed when building microscopic traffic simulation models. The reader is, however, referred to [6] for descriptions of the remaining three simulation approaches.

Agent-based modelling is most simply defined as a decentralised, individual-centric approach to model design [1]. With agent-based modelling the behaviour of the global system as a whole is not defined, but rather the behaviour of its constituent entities, or agents. These agents can be anything from people at a train station to companies in a specific business sector, or, as is the case in a microscopic traffic simulation model, vehicles on a road network. It is from the interactions among these agents that the global behaviour of the system emerges and may be studied [4].

3 A microscopic traffic simulation modelling framework

The agent-based traffic simulation framework described in this paper was developed in *AnyLogic*, a Java-based multi-method simulation software package. AnyLogic was chosen as the modelling platform as it supports agent-based modelling and contains a basic road traffic library [1]. The framework was designed to be used as a testing mechanism facilitating the investigation and comparison of the effectiveness of previously proposed or novel traffic control strategies. Building a model within this framework may be accomplished in two separate stages. The first stage involves the design of the road network itself as well as basic traffic signals at each intersection. The second stage involves populating the road network with vehicles and defining the logic responsible for the movement of these vehicles through the system (*e.g.* the desired speed of vehicles, when vehicles should accelerate or decelerate, what this rate of acceleration or deceleration should be, and the origins and destinations of vehicles). Once built, the model may be used for data collection and analysis purposes. The data thus collected may be used to evaluate metrics which act as performance measure indicators for traffic signal control strategies. Examples of the data that may be collected include the number of vehicles present along a given lane of a given road segment, the speeds of individual vehicles, the distances of individual vehicles from an intersection, or whether vehicles are queued or travelling at their desired speeds. Examples of performance measure indicators include the delay times experienced by vehicles in the system as well as the number of stops made by the vehicles.

3.1 Building the road network and traffic signals

Before building the road network, it is required that certain global parameters are defined which dictate the appearance and connectivity of the road network. These parameters are the scale of the road network, the connection tolerance and the lane widths. The scale defines the number of pixels per metre, thereby linking the unitless display of the modelling framework graphic with an actual unit of length. The connection tolerance (measured in pixels) is the maximum distance between two lane ends for which the two lanes are considered to be connected, *i.e.* if two lane ends are closer than the connection tolerance and form an obtuse angle, they are considered as connected, and a vehicle that exits one lane may continue travel on the other. The lane width (measured in metres) defines how wide each lane in the road network will be, and as a result, how many lanes each road segment will contain. For example, if a line has a width of 60 pixels and the scale of the road network is 10 pixels per metre, and the lane width is set to 3 metres, then the corresponding road segment will comprise two lanes. The default speed limit for the road network is also user-defined and is measured in metres per second.

The traffic signals positioned at each intersection are modelled as individual agents and potentially operate independently of one another in order to facilitate the use of self-organising traffic control strategies. The signal switching logic is controlled by means of a state chart which comprises various states and state transitions. The number of different states in the state chart is determined by the number of phases which comprise a complete signal cycle at the intersection. The transition from one state to another is determined by the type of signal control implemented. In the case of fixed pre-timed control, a

time-out function is employed such that the transition from the current state to the next state is triggered once a user-specified amount of time has elapsed since the current state was entered. For more advanced, vehicle-actuated traffic signal control strategies, state transitions may be triggered when a specified boolean condition is true, or upon receipt of a specific message string.

3.2 Populating the road network

With the road network and traffic signals in place, the next step is to introduce vehicles into the simulation model. Vehicles enter the road network at designated entry points. These vehicle arrivals may be defined according to one of four user-specified methods. The vehicles may arrive at a user-specified rate, in which case arrivals are stochastic and follow a Poisson distribution with a mean equal to the chosen rate. This is equivalent to specifying exponentially distributed interarrival times between vehicles with a mean equal to the inverse of the chosen rate. Alternatively, the user may specify an interarrival time which would be identical for all arriving vehicles. The user may also choose to implement a stochastic rate schedule which defines how the arrival rate changes over time. Finally, the user may define a deterministic arrival schedule, in which case the arrivals of vehicles are generated according to the exact times defined in the arrival schedule.

When a vehicle is generated, several vehicle-specific parameters are defined instantaneously. These include the origin-destination pairing of the vehicle, the size of the vehicle, the vehicle's rates of acceleration and deceleration, and the vehicle's desired speed of travel. Vehicles are generated at each entry point to the road network, and upon generation the final destination of the vehicle is determined by Monte Carlo simulation. This origin-destination pairing of the vehicle dictates when and where a vehicle must change lanes, as well as whether it should turn left or right at an intersection, or carry on travelling straight. Monte Carlo simulation is used to determine the size of the vehicle generated. The user decides on the probabilities associated with the different sizes of vehicles which ultimately determines the number of small, medium and large vehicles present in the road network. A vehicle's size determines its rates of acceleration, deceleration and desired speed. Typically, the larger the vehicle, the slower its rates of acceleration and deceleration, and the lower its desired travel speed. These trends may, however, be overridden by the user.

Apart from the logic which determines how fast a vehicle travels, or at what rate it accelerates or decelerates, logic has also been implemented which determines when and where a vehicle must accelerate or decelerate. Associated with each vehicle are minimum and maximum allowable distances to the vehicle in front of it, which depend on the vehicle's speed, as well as minimum and maximum allowable speeds, which, in turn, depend on the distance to the vehicle in front of it. There is also a maximum speed allowed on curved roads (*e.g.* corners). Let v_i be the speed of vehicle i and let $s_{i,i-1}$ be the distance between vehicle i and vehicle $i-1$ in front of it. Now, if $s_{i,i-1}$ is less than the value of some function f of v_i which determines the minimum allowable distance between two vehicles or if v_i is greater than the value of some function g of $s_{i,i-1}$ which determines the maximum allowable speed of a following vehicle, then vehicle i will decelerate. On the other hand, if $s_{i,i-1}$ is greater than the value of some function f' of v_i which determines the maximum

allowable distance between two vehicles or if v_i is less than the value of some function g' of $s_{i,i-1}$ which determines the minimum allowable speed of a following vehicle, then vehicle i will accelerate. The maximum speed on curved roads is determined according to a function h of the radius of the arc of the curve. The functions f, f', g, g' and h are all user-defined.

The logic responsible for a vehicle's interaction with traffic signals operates in much the same manner. When a red or a late amber signal is displayed, vehicles decelerate as if there were a stationary vehicle at the stop line of the intersection. In the case of permissive right-turning vehicles, while a green signal is displayed, the vehicle turning right will wait in the intersection until the intersection and a portion of road, which extends a user-specified distance on the opposite side of the intersection, are free of any on-coming traffic, at which point the vehicle will complete its turn. For the case in which a right-turning vehicle is still present in the intersection when the traffic signal changes from green to amber, the vehicle need only wait until the intersection is free of any oncoming traffic before completing its turn.

3.3 Data collection and assimilation

The traffic simulation modelling framework described above was designed to allow for testing traffic control algorithms which assume the use of radar detection technology mounted at the intersection. Such radar sensors typically achieve a detection range of up to 275 metres [9] and are capable of detecting and tracking the speeds and positions of individual vehicles along a road segment, thereby enabling them to determine the vehicles' estimated times of arrival at the intersection. It was therefore necessary to incorporate this logic into the model.

Three lists are associated with each lane adjoining an intersection:

CarList. This list contains all the vehicles present on a lane. As a vehicle enters the lane, be it at the lane's entry point or as a result of a lane change, it is added to this list. A vehicle is removed from the list when it reaches the end of the lane or when it changes onto an adjacent lane. This list provides the user with information on the number of vehicles present along a specific lane. It makes it easier for a user to access individual vehicles and their associated characteristics, such as speed.

Queue. This list contains all queued, motionless vehicles along a lane and is a subset of the previously mentioned list, *CarList*. A vehicle is added to this list as soon as its speed equals zero. It is removed from the list as soon as it begins accelerating from rest. The list provides the user with information on the queue length along a specific lane.

QPred. This list contains all vehicles both currently stopped and queued as well as those which have not yet stopped, but will become queued before the existing queue has been cleared, and is again a subset of the first list, *CarList*.

In order to predict which vehicles will become queued, it is necessary to predict where the back of the queue will be. The predicted vehicle queue and back-of-queue position are

calculated continually by an algorithm. For every vehicle not in the predicted queue list *QPred*, the algorithm calculates the amount of time it will take the vehicle to reach the current back-of-queue position. The algorithm then compares this time to either the sum of the remaining red time and the time required to clear the current predicted queue of vehicles (if the traffic signal displayed is not green) or just the time required to clear the current predicted queue of vehicles (if the signal displayed is green). If this time is found to be shorter, then the vehicle is added to the predicted vehicle queue list *QPred* and the back-of-queue position is incremented by the length of the vehicle plus the minimum space gap between stationary vehicles. For the case in which the front vehicle along a lane is not yet queued (*i.e.* the predicted queued vehicle list *Queue* is empty) the vehicle is added to the predicted queued vehicle list under one of two conditions. If the traffic signal displayed is green and the vehicle cannot clear the intersection before this signal changes to amber and ultimately to red, then it is added to the list. Analogously, if the traffic signal displayed is not green and the vehicle will arrive at the intersection before the signal changes from red to green, then again, it is added to the list. Vehicles are removed from the predicted queued vehicle list when they depart from the associated lane, at which point in time the predicted back-of-queue position is decremented by the length of the vehicle together with the minimum space gap between stationary vehicles. An example of a typical intersection scenario may be seen in Figure 1.

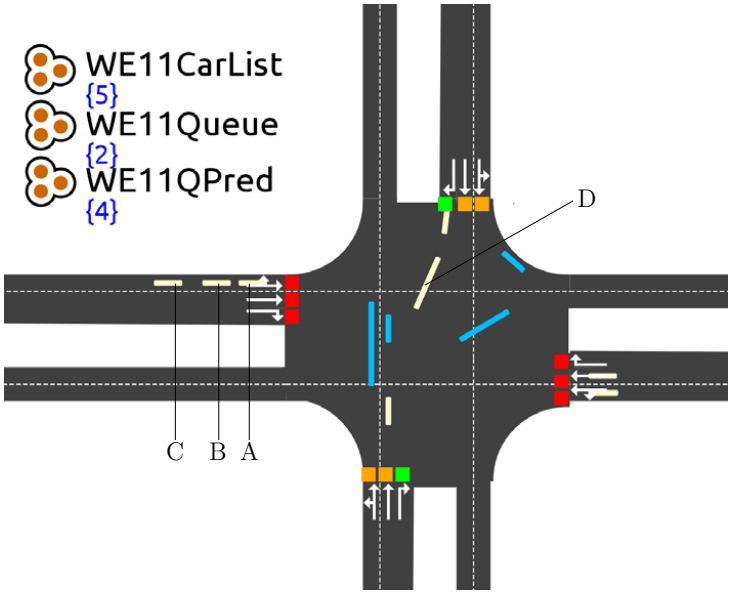
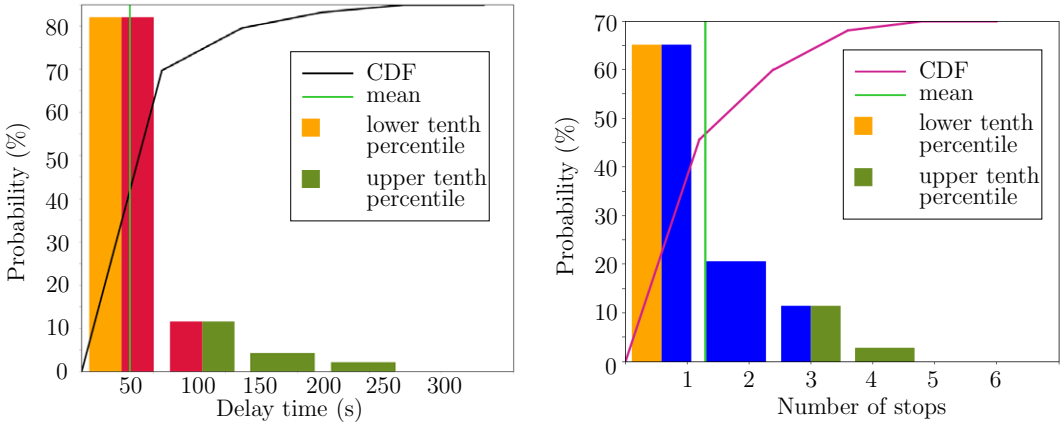


Figure 1: An example of an intersection scenario implemented in the modelling framework described in this paper. The list labelled *WE11CarList* is a list of all the vehicles present along lane 1 approaching the intersection (intersection 1) travelling in a West-to-East direction, and in this scenario has a size of five. The list labelled *WE11Queue* is a list of all the stationary vehicles along the lane, which comprises the two stationary vehicles in this scenario (labelled A and B). The vehicle labelled C has not yet come to a complete stop and therefore has not been added to the list. The third list, labelled *WE11QPred*, is a list of all the vehicles that are predicted to become queued and thus delayed. In the scenario depicted this list contains four vehicles. This means that the fifth vehicle along the lane will clear the intersection without becoming queued. The vehicles travelling in South-to-North and North-to-South directions are receiving an exclusive right-turn phase. The vehicle labelled D is currently waiting in the intersection while the three vehicles travelling in the opposite direction clear the intersection.

Due to the fact that the desired speed, as well as the origin-destination pairing of a vehicle, is known upon its generation and therefore, the total distance the vehicle is to travel, the delay time a vehicle experiences while travelling through the road network may be calculated by subtracting the time it would take the vehicle to move from its origin to its destination without being impeded by any traffic signals and resulting queues or slower moving vehicles from the actual time it spends travelling through the road network. The minimum time a vehicle can spend travelling through the road network is calculated by dividing the distance the vehicle has to travel from its origin to destination by its desired speed. The actual time spent by a vehicle travelling through the road network is captured by a timing mechanism which records the time the vehicle enters the road network as well as the time it leaves the road network.

The average delay time of all vehicles which pass through the road network is an important performance measure indicator as it provides the user with an idea of how different traffic signal control algorithms perform in respect of their ability to minimise driver delay under various prevailing traffic conditions. This feature also provides information on the maximum amount of time a driver was delayed, another important performance measure indicator to consider. A third performance measure indicator implemented is that of the number of stops a vehicle makes while travelling through the road network. An integer value is associated with each vehicle and is initialised as zero. Each time a vehicle comes to a complete stop, this value is incremented by one. The average number of stops made by vehicles can provide the user with an idea of the efficiency of traffic signal control algorithms in respect of their propensity to facilitate green waves, because the fewer vehicles that are required to stop as a result of red traffic signals, the lower their delay time is likely to be.

The framework allows for real-time analysis to take place as output is generated and the results of such an analysis can be displayed while the model is running. An example of this output is shown in Figure 2.



(a) Probability density function and cumulative distribution function of the delay time experienced by vehicles.

(b) Probability density function and cumulative distribution function of the number of stops made by vehicles.

Figure 2: An example of dynamic output generated by a model in the framework described in this paper while it runs.

4 Conclusion

The traffic simulation modelling framework described in this paper forms an integral part of an ongoing study into the feasibility and effectiveness of self-organising traffic signal control algorithms. It is intended that the framework, as well as all associated code, be made publicly available in the future. The model provides the user with an analytic tool which may be adjusted to suit the specific modelling requirements of the user.

Although the modelling framework described in this paper was built with the intention of investigating and comparing various traffic signal control strategies, it is not limited to such investigations alone. It may be used to investigate the effects of other factors on the system as a whole, such as the addition or removal of lanes from road sections, disruptions as a result of vehicle breakdowns, building of pedestrian crossings, or the installation of speed cameras, to name but a few.

It is the intention of the authors to implement novel self-organising traffic signal control strategies within the traffic simulation modelling framework described here so as to showcase the potential and benefits of the application of self-organisation to traffic control optimisation and congestion reduction as well as the benefits of using radar detection as opposed to conventional electromagnetic induction loop detectors. Another potential focus area for future work is the development of road network topologies of varying size and configuration so as to investigate under what conditions the various signal control strategies, as well as the types of detection equipment are most, and least effective.

References

[1] The AnyLogic Company, 2014, *AnyLogic help*, [Online], [Cited April 23rd, 2014], Available from `http://www.anylogic.com/anylogic/help/`

[2] Banks J, Carson JS, Nelson BL & Nicol DM, 2001, *Verification and validation of simulation models*, pp. 367–397 in Fabrycky WJ & Mize JH (Eds), *Discrete-event system simulation*, 3^{rd} Edition, Prentice-Hall, Upper Saddle River (NJ).

[3] Barceló J, 2010, *Models, traffic models, simulation and traffic simulation*, pp. 1–62 in Barceló J (Ed), *Fundamentals of traffic simulation*, Springer, New York (NY).

[4] Borshchev A & Filippov A, 2004, *From system dynamics and discrete event to practical agent based modeling: Reason, techniques, tools*, Proceedings of the 22nd International Conference of the System Dynamics Society, pp. 25–29.

[5] Craig DC, 1996, *Extensible hierarchical object-oriented logic simulation with an adaptable graphical user interface*, Doctoral Dissertation, Memorial University of Newfoundland, St. John's.

[6] Einhorn MD, 2012, *An evaluation of the efficiency of self-organising versus fixed traffic signalling paradigms*, MSc Thesis, Stellenbosch University, Stellenbosch.

[7] Gyimesi K, Vincent C & Lamba N, 2011, *Frustration rising: IBM 2011 commuter pain survey*, International Business Machines, [Online], [Cited April 23rd, 2014], Available from `http://www.ibm.com/press/us/en/presskit/35314.wss`

[8] Papacostas CS & Prevedouros PD, 2001, *Transportation software*, pp. 626–652 in Curless L (Ed), *Transportation engineering and planning*, Prentice-Hall, Upper Saddle River (NJ).

[9] Wavetronix, 2014, [Online], [Cited May 5th, 2014], Available from `http://www.wavetronix.com/en/products/smartsensor/advance/features`

Proceedings of the 2014 ORSSA Annual Conference
pp. 9–17
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Developing long-term scenario forecasts to support electricity generation investment decisions

R Koen* T Magadla† P Mokilane‡

Abstract

Many decisions regarding capital investment in electricity generation technologies need to be made well in advance, usually when there is still a large amount of uncertainty regarding the favourability of future conditions. There may be uncertainty about the amount of electricity required in future as well as the the variability in the demand, and both of these uncertainties can affect decisions pertaining to such capital investment decisions.
This paper presents an approach that uses multilevel models to develop scenario forecasts for South African load profiles (hour-to-hour changes in the electricity demand), which can then be used to support decisions regarding the electricity generation capacity required. Although historical load profile patterns are known, there is uncertainty about how future patterns will deviate from historical ones. By developing scenarios that represent different views about future load profile patterns, forecasts can be obtained for each scenario and, in turn, these scenario forecasts can be used to investigate the effect of changes in demand patterns on future electricity generation requirements. The approach of using multilevel modelling to obtain long-term hourly forecasts for a particular scenario has not been seen elsewhere in the literature, but shows promise for providing appropriate support electricity generation expansion decisions.

Key words: electricity load profiles, forecasting load profiles, long-term load forecasting, multilevel models, scenario forecasts.

1 Introduction

Decisions regarding capital investments in the electricity sector, *i.e.* decisions regarding investment in electricity generation technologies, generally need to be taken a long time before the investment is made, since building projects are typically large and take long to complete. Therefore, there may be substantial uncertainty regarding the future conditions under which the generation technologies will be operating at the time that decisions need

*Corresponding author: Council for Scientific and Industrial Research (CSIR), South Africa, PO Box 395, Pretoria, 0001, email: rkoen@csir.co.za

†Council for Scientific and Industrial Research (CSIR), South Africa, PO Box 395, Pretoria, 0001, email: tmagadla@csir.co.za

‡Council for Scientific and Industrial Research (CSIR), South Africa, PO Box 395, Pretoria, 0001, email: pmokilane@csir.co.za

to be made, and decisions have to rely heavily on forecasts of operating conditions which estimate the total amount of electricity that will be required in future as well as the expected variability in demand. Investment in an appropriate mix of technologies requires knowledge regarding periods of low and high demand so that costs can be minimised, for instance, by limiting technologies that have high operating costs to being used only during very short, high demand (peak) periods.

Fluctuations in electricity demand are measured by so-called load profiles, which records the hour-to-hour electricity demand. It seems to be standard practice to develop scenario forecasts for total annual demand in order to allow for uncertainty in the forecasts to be taken into account in the decision-making process, but not always to consider changes in future load profiles. Usually, total annual demand is forecasted using scenarios for the required future period, and then a "typical" or "average" annual load profile is superimposed on the annual demand scenarios. This paper presents a possible way of determining scenario forecasts for load profiles in order to also include uncertainties regarding future load profile changes in capital investment decisions.

2 Data used

Load profile patterns differ for different countries and regions in the world. These differences can be attributed to differences in social, economic and climatic conditions, but also due to differences in the way electricity may be used in response to such factors (see, for example, Pilli-Sihvola *et al*, 2010, for a discussion on different usage patterns due to climate differences in countries in Europe). Data on load profiles in South Africa was obtained from Eskom, the main producer of electricity and operator of the national transmission grid. Historical load profile data was supplied by Eskom for the period 1 January 1997 – 31 December 2013. The data for the period 1997 – 2012 was used to build forecasting models, while the 2013 data was used to test the forecasts obtained from the models.

The data consisted of a date, an hour of the day (ranging from 0 to 23) and the total electricity supplied during that hour, measured in MegaWatts. The hourly demand data was adjusted by the relevant total annual demand to ensure that the load profile pattern that was used was not influenced by year-to-year changes in the overall demand. The forecasting models were developed using only information related to the electricity demand, the date and time. In addition, information related to the date, namely which day of the week it represented; whether the particular day fell on a public holiday; whether the day fell within the "slow" December period after the 16^{th} of December public holiday or whether it fell within the "typical" peak winter usage period (weeks 25 to 29) was added to the dataset. Note that although weather patterns are known to have an influence on load profiles, data on climatic conditions such as temperature and humidity was not included in the forecasting models.

3 Method

In reviewing forecasting methods for load profiles, Hahn, Meyer-Nieberg and Pickl (2009) point out that "... up to now, the main focus in load forecasting has been on short-term load forecasting, since it is an important tool in the day-to-day operation of utility systems" and indicate that fewer studies are reported on long-term load forecasts of 20 –30 years ahead. Hyndman and Fan (2010) agree that "In the literature to date, short-term demand forecasting has attracted substantial attention... medium- and long-term forecasting have not received as much attention, despite their value for system planning and budget allocation." Although literature references were found that used various methods for medium to long-term forecasting of load profiles, such as regression (Al-Hamadi and Soliman, 2005, and Hyndman and Fan, 2010), probabilistic forecasts using cross-correlations (McSharry, Bouwman and Bloemhof, 2005), functional data analysis (Besse, Cardot and Stephenson, 2000, and Aguilera, Ocaa and Valderrama, 1999), expert systems (Kandil, El-Debeiky and Hasanien, 2002), support vector machines (Trkay and Demren, 2011), grey dynamic systems (Morita, Kase, Tamura, Iwamoto, 1996) and neural networks (Kermanshahi, 1998, Yue, Zhang, Xie and Zhong, 2007 and Carpinteiro *et al*, 2009, amongst others), these reported methods have not been used to develop forecasts for more than 10 years into the future, most of them for one year ahead only. Therefore, it seems as though there are very few methods that are available in the literature for forecasting long-term load profile patterns for 20-30 years into the future, and particularly not based on only date and time data.

One challenge was therefore to find an appropriate method to obtain a set of long-term forecasts, *i.e.*, 20 - 30 years, for South African electricity load profiles. A further challenge was to obtain a method that could be used to develop different sets of alternative long-term load profile forecasts. These alternative sets are potentially very valuable for decision-making – perhaps not for describing all possible future patterns, since there is ignorance rather than uncertainty regarding what will happen in the future (see discussion in Stirling, 2010), but rather for determining to what extent possible changes in future patterns could change decisions regarding generation technology requirements.

3.1 Load profile patterns

The typical load profile patterns for South Africa are illustrated in Figure 1 and Figure 2.

Figure 1 shows that more electricity is consumed during the winter months than during the summer months of any year. Figure 2 illustrates the electricity usage pattern during a day. Although there is considerable variation in the hourly demand on different days throughout a year, the spline curve superimposed on the scatter plot indicates the "average" pattern observed over the course of a day, with the typical "two peaks" observed in the early morning and late afternoon.

Investigation of the data showed a high degree of correlation within a day (correlation coefficients of above 0.95 between all hourly values). This correlation pattern was seen as important for the development of forecasting models, and led to the consideration of a multilevel modelling approach.

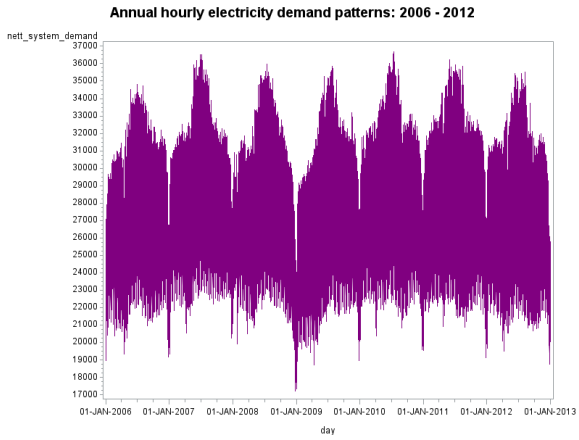


Figure 1: *Annual pattern in hourly consumption data for South Africa (Source: Eskom).*

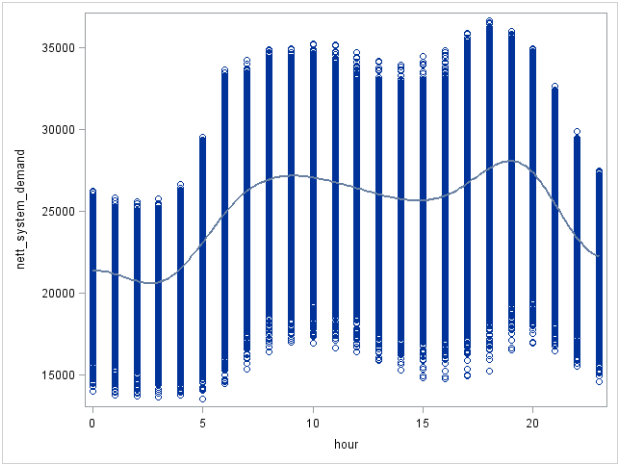


Figure 2: *Daily electricity demand patterns in South Africa (Source: Eskom).*

3.2 Multilevel models

Multilevel models (also called mixed effect models, mixed level models, hierarchical models, or random effects models) have been used in a wide range of domains. A group of applications originated in medical domains, with most of these applications reported as Longitudinal Data Analysis (LDA) applications in the literature. LDA was developed for the analysis of "longitudinal" data, *i.e.* data collected over time from a set of patients or to monitor changes over time for particular patients. Another group of applications can be found within the social sciences (see, for example, Gentry and Martineau, 2010, or Ozkaya *et al*, 2013, or Singer, 1998) where they are mostly referred to as hierarchical models or multilevel models, thereby emphasizing the fact that observations may be grouped within a hierarchy, and that observations within the same level of a hierarchy may be correlated.

The emphasis of multilevel modelling is usually on the assessment of change over time. According to Davidian and Giltinan (2003), a common challenge across different domains is to understand features underlying profiles of continuous, repeated measurements taken on

a sample of individuals. Such profiles over time can be linear or non-linear. According to Wang, Xie and Fisher (2012), if a regression model is fitted to data that is characterised by repeated measures within a hierarchical structure, the assumption of uncorrelated observations may be violated and therefore the regression model may not provide valid results. Although no literature references that applied multilevel models to load profile data were found, the obtained load profile data seemed to show a pattern of repeated, correlated (hourly) measurements nested within a hierarchical structure (*i.e.* a day), similar to repeated blood pressure measurements on patients or student performance measurements nested within schools.

The approach taken in this study followed the method suggested by Wang, Xie and Fisher (2012), but books by Frees (2004), Verbeke and Molenberghs (2000) as well as the article by Singer (1998) also provide useful descriptions of the application of multilevel models to a variety of practical situations.

3.3 Describing non-linear daily patterns

Although Davidian and Giltinan (2003) describe a non-linear multilevel approach, there seemed to be limits to the types of non-linear functions that could be used. It was therefore decided to determine the functional form of the non-linear daily pattern first and then to use transformations to linearise the data before applying a linear multilevel modelling approach. In order to find a functional form to describe the daily pattern, symbolic regression, which is a form of genetic programming, was used.

Given a set of input data x, and a set of output data y, the aim of symbolic regression is to find a function f such that $y = f(x) + e$. Unlike traditional linear and non-linear regression techniques that involve determining the parameters that best make a pre-defined function fit a set of observed data, symbolic regression makes no assumption about the structure of the function (Lew *et al*, 2006). It involves searching the space of mathematical expressions for the structure of the function as well as the parameters that best describe the data while minimising various error metrics, and works in the following manner (Poli *et al*, 2008):

1. A set of mathematical expressions, e.g. {+, -, *, /, v, ^, ...} is specified.
2. An evolutionary algorithm to evolve both the structure of the function and the parameters is then applied.
3. Functions that minimise a certain error metric are retained and further evolved.
4. This process continues until an appropriate function has been found.

4 Results

The stepwise approach recommended by Wang, Xie and Fisher (2012) was used to carry out the analysis. The first step is to calculate the Intra-class Correlation Coefficient (ICC),

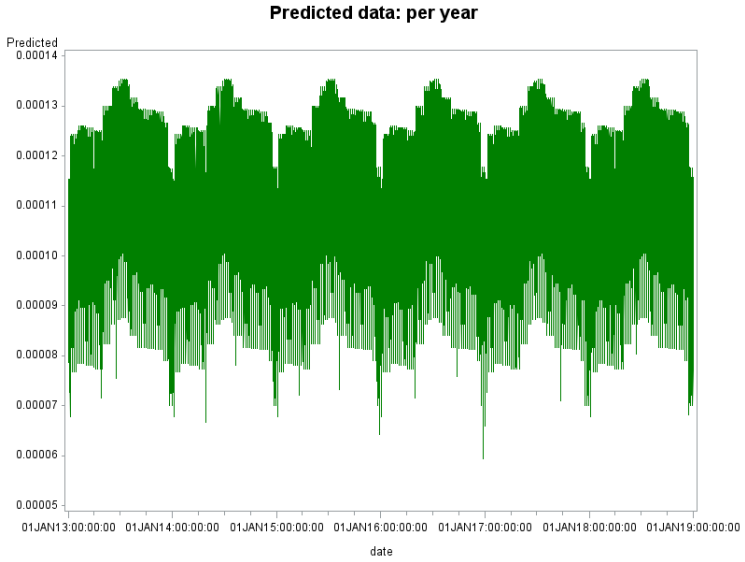


Figure 3: *Load profile predictions from preferred model for 2013 - 2018.*

which indicates whether there is sufficient correlation among observations within the levels of the hierarchy to justify the use of multilevel modelling. If the ICC is close to 1, then multilevel modelling is justified, if it is close to 0, then ordinary regression may be used. According to Wang, Xie and Fisher (2012) the classical definition of the ICC was given by Shrout and Fleiss (1979) as the ratio of the between-group variance to the total variance, where the total variance is the sum of the within-group and between-group variances, *i.e.*, $ICC = \frac{s_b^2}{s_b^2+s_w^2}$. For this data, the ICC was calculated as 0.6536, or 65.36%, therefore using multilevel modelling was indeed justified.

A multilevel model was fitted using SAS Proc MIXED. Symbolic regression was used to identify the type of non-linear relationship present within the daily profile (*i.e.* to describe the functional form of the relationship between hourly electricity demand and the hour of the day). Eurequa, which is software that is specially designed for conducting symbolic regression, was used for this purpose. The symbolic regression indicated that using log, sine and cosine functions of the hourly values resulted in a better fit than a polynomial or power function. Therefore these functions were used in obtaining suitable transformations of hourly values to include as explanatory variables. The Log Likelihood, Akaike information criterion (AIC) and the Bayesian information citerion (BIC) were used to select the model with the most suitable combination of explanatory variables. Forecasts were generated for 2013 using the selected model, and these forecasts were compared to the actual 2013 hourly values in order to assess the forecast accuracy. A Mean Absolute Percentage Error (MAPE) was calculated in order to compare the 2013 forecasted values against the 2013 actual values. The preferred model had a MAPE value of 3.4% - Lewis (1982) indicates that a MAPE of less than 10% can be classified as a highly accurate forecast. Figure 3 shows an extract of the forecasts obtained from the preferred model.

5 Implementation in scenarios

The multilevel modelling approach seems to deliver models that fit the historical load profiles successfully, and that seem to provide relatively accurate forecasts. It is also well suited to developing scenario forecasts. Firstly, a multilevel model produces parameters very similar to those from an ordinary regression model, indicating the effect of different explanatory variables on the load profile patterns. In addition, the non-linear functions used to model the daily pattern can be manipulated or studied separately. Finally, the fact that the model seems to be successful in modelling the electricity demand based on explanatory variables related to the calendar date and hour only, and does not seem to require variables such as hourly temperature, means that the model can produce forecasts far ahead into the future without requiring difficult to obtain forecasts for explanatory variables over the same future period.

While there is uncertainty about how future load profiles will deviate from historical patterns, this modelling approach can be used to provide scenario forecasts for load profiles based on different views regarding potential future patterns. The scenario forecasts can then be used to investigate the effect of potential changes in demand patterns on future electricity generation requirements. Scenario forecasts produced by the multilevel modelling approach are in a suitable format for using as inputs into currently existing capacity planning Linear Programming models, and therefore can be used to do sensitivity testing of decisions. This approach therefore shows promise for providing appropriate support for electricity generation expansion decisions, and its usefulness is currently being tested.

6 Acknowledgements

The research underlying this paper was conducted during a contract research project funded by Eskom. The team would therefore like to thank Eskom for funding the research, and would particularly like to thank the following individual Eskom staff members for the supply of data, the explanation of historical data and electricity terminology, as well as support during the analysis and modelling work: Moonlight Mbata, Letu Moti, Lisinda Du Plessis and Ferdi Kruger.

7 References

Aguilera, A. M., Ocaa, F. A. and Valderrama, M. J., 1999 *Forecasting with unequally spaced data by a functional principal component approach*, Test, Vol 8, pp 233-253.

Al-Hamadi, H.M., Soliman S.A., 2005, *Long-term/mid-term electric load forecasting based on short-term correlation and annual growth*, Electric Power Systems Research, Vol 74, pp 353–361

Besse, P., Cardot, H. and Stephenson, D., 2000, *Autoregressive forecasting of some functional climatic variations*, Scandinavian Journal of Statistics, Vol 27, pp 673–687.

Carpinteiro,O.A.S., Lima, I., Moreira, E.M., Pinheiro, C.A.M., Seraphim, E., Pinto, J.V.L., 2009, *A hierarchical hybrid neural model with time integrators in long-term load forecasting*, Neural Computation and Applications, Vol 18, pp 1057–1063

Davidian, M. and Giltinan, D.M*., Nonlinear models for repeated measurement data: an overview and update (invited article*), 2003, Journal of Agricultural, Biological and Environmental Statistics, Vol 8, No 4, pp 387 – 419

Frees, EW, 2004, *Longitudinal and Panel Data: Analysis and applications in the Social Sciences*, Cambridge University Press

Fan, S., Hyndman, RJ, 2012, *Forecasting Electricity Demand in Australian National Electricity Market*, IEEE Power & Energy Society General Meeting, pp 1- 4

Gentry, W.A., Martineau, J.W., 2010, *Hierarchical linear modeling as an example for measuring change over time in a leadership development evaluation context*, The Leadership Quarterly, Vol 21, pp 645-656

Hahn, H., Meyer-Nieberg, S., Pickl, S., 2009, *Electric load forecasting methods: Tools for decision-making*, European Journal of Operational Research, Vol 1999, pp902 – 907

Hyndman, R.J., Fan, S., 2010, *Density Forecasting for Long-term Peak Electricity Demand*, IEEE Transactions on Power Systems, Vol 25, Issue 2, pp 1142-1153.

Kandil, M. S., El-Debeiky, S. M., Hasanien, N. E., 2002, *Long-Term Load Forecasting for Fast Developing Utility Using a Knowledge-Based Expert System*, IEEE Transactions on Power Systems, Vol 17, No 2, May 2002, pp 491 – 496.

Kermanshahi, B., 1998, *Recurrent neural network for forecasting next 10 years loads of nine Japanese utilities*, Neurocomputing, Vol 23 , pp 125-133

Lew, T. L., Spencer, A. B., Scarpa, F., Worden, K., Rutherford, A. and Hemez, F., 2006, *Identification of response surface models using genetic programming*, Mechanical Systems and Signal Processing, Vol 20, No 8, pp 1819-1831.

Lewis, C.D., 1982, *Industrial and business forecasting methods: A radical guide to exponential smoothing and curve fitting*, Butterworth Scientific: London, Boston.

McSharry, P.E., Bouwman, S., Bloemhof, G., 2005, *Probabilistic Forecasts of the Magnitude and Timing of Peak Electricity Demand*, IEEE Transactions on Power Systems, Vol 20, No 2, pp 1166-1172.

Morita, H., Kase,T., Tamura,Y., Iwamoto, S., 1996*, Interval prediction of annual maximum demand using grey dynamic model*, Electrical Power & Energy Systems, Vol 18, pp. 409-413

Ozkaya, H.E., Dabas, C., Kolev, K., Hult, G.T.M, Dahlquist, S.H., Majeshwar, S.A., 2013, *An assessment of hierarchical linear modeling in international business management and marketing,* International Business Review, Vol 22, pp 663 - 677

Pilli-Sihvola,K., Aatola, P., Ollikainen, M., Tuomenvirta, H., 2010, *Climate change and electricity consumption - Witnessing increasing or decreasing use and costs?*, Energy Policy, Vol 38, pp 2409–2419

Poli, R., Langdon, W. B., McPhee, N.F., Koza, J. R., 2008, *A field guide to genetic programming*, Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk [accessed July 2014].

Shrout, P.E. and Fleiss, J.L., 1979, *Intraclass correlations: Uses in assessing rater reliability*, Psychological Bulletin, Vol 86, pp 420-428

Singer, JD, 1998, *Using SAS PROC MIXED to fit multilevel models, hierarchical models and individual growth models*, Journal of Educational and Behavioral Statistics, Vol 24, pp 323-355

Stirling, A., 2010, *Multicriteria diversity analysis: A novel heuristic framework for appraising energy portfolios*, Energy Policy, Vol 38, pp 1622–1634

Trkay,B.E., Demren, D., 2011, *Electrical Load Forecasting using Support Vector Machines*, 7th International Conference on Electrical and Electronics Engineering (ELECO), pp I-49 - I-53

Verbeke, G., Molenberghs, G, 2000, *Linear Mixed Models for Longitudinal Data*, Springer-Verlag New York

Wang, J., Xie, H., Fisher, J.H., 2012, *Multilevel models: Applications using SAS*, Higher Education press & Walter de Gruyter GmbH & Co

Yue, L., Zhang, Y., Xie, H., Zhong, Q., 2007*, The Fuzzy Logic Clustering Neural Network Approach for Middle and Long Term Load Forecasting*, Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, November 18-20, 2007, Nanjing, China

Proceedings of the 2014 ORSSA Annual Conference
pp. 18–26
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Development of an optimiser for a simulator of an electric utility: Challenges and approach

MN Hatton* JF Bekker†

Abstract

An efficient and reliable energy generation capability is vital to any country's economic growth. Many strategic, tactical and operational decisions exist along the energy supply chain. Shortcomings in this developing nation's energy production industry have led to the development of an Energy Flow Simulator (EFS). The simulator is claimed to incorporate all significant factors involved in the energy flow process, from coal to consumption. Currently, a study is done to add an optimisation capability to the simulator. The paper serves three main purposes: To summarise literature on energy market modelling, to provide an overview of the simulator, and to pave the path for optimisation of the simulator.

Key words: Electricity industry, Energy sector, Simulation, Simulation optimization.

1 Introduction

Throughout the world, electric power utilities form an essential foundation for nations' economies. The planning thereof is a highly complex process as it is characterised by decentralized decision-making and involves many processes structured in an intricate hierarchy [17, 18]. In developing countries, planning is even more challenging. With the rapid economic growth of emerging economies, electricity demand is rapidly increasing [1]. Commission of power systems to meet such demand places huge pressure on already constrained capital reserves. Adding to the problem, developing countries have to keep energy policies in-line with the worldwide push to cleaner energy policies, which come at a far greater cost than traditional coal-fired power stations [5]. Therefore, energy utility planners are assigned a daunting decision-making task.

Strictly speaking, electricity falls under the greater field of energy. However, in this paper electricity and energy will be used interchangeably. Nuclear, gas and renewable energy sources are included, but the focus is on coal-fired power stations coal-fired as they produce

*Department of Industrial Engineering, University of Stellenbosch, Private Bag X1, Stellenbosch, 7602, South Africa.
†Department of Industrial Engineering, University of Stellenbosch, Private Bag X1, Stellenbosch, 7602, South Africa.

the majority of electricity. It is also important to note that this electric utility is a vertically integrated monopoly. A simple example of how it could be important is the effect it could have on the modelling process. Because it is a monopoly, cost minimization could be the objective, as opposed to the more standard approach of maximising a utility's profit.

The article is structured as follows. Energy sector modelling techniques are investigated in §2. §3 provides an overview of the simulator. Subsequently, the proposed optimiser is presented in §4, and finally, the paper is concluded in §5.

2 Modelling techniques in literature

This section is grouped into two parts. At first, modelling techniques applied to specific problems within the energy sector are analysed. Subsequently, the holistic modelling of the energy sector is investigated, treating the energy sector in a like manner to an economic supply and demand market.

2.1 Modelling problems within the sector

In energy industry literature, there are many specific optimisation problems. Some examples of such problems are economic dispatch [31], generator maintenance scheduling [6, 22, 31], coal handling process scheduling [4], transmission losses [31], hydrothermal coordination [31], coal stockpile simulation [20], optimal power flow [31], generation expansion planning [26, 29], and state estimation problems [31].

Additionally, there have been countless studies on load forecasting and demand reduction, namely: Modelling household electricity-saving using a modified SMAA approach [8] and system dynamics [7], including variants of ARIMA modelling for electricity demand [3, 10], and a seasonal demand forecasting hybrid procedure[33]. [2, 28] provide surveys of energy demand models.

Initially, power generation studies focused upon very specific problems, such as optimal power flow [31]. However, currently models are being developed for integrated resource plans. Integrated resource plans aim to incorporate all facets of the energy supply chain, within a specific region [19, 25, 30]. The future trend has an increased focus towards modelling the entire global energy supply chain, with an emphasis on emissions [19, 25].

2.2 Modelling the sector as a whole

Given the importance of efficient and reliable energy planning, many decades of research have culminated into countless energy flow planning models [30]. Essentially, the aim of electric power utility planning is to provide reliable electricity at an acceptable economic and environmental cost [17]. Ventosa *et al.* [30] survey three types of models: Simulation, equilibrium and optimisation. Equilibrium models incorporate multiple utilities, and are thus irrelevant to this single utility study. Usually, either a simulation model or an optimisation model is applied. However, in this study both are incorporated.

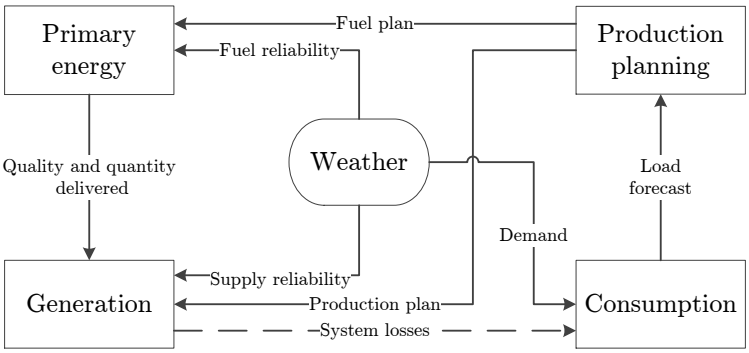


Figure 1: *High-level representation of the Energy Flow Simulator. (Source [12]).*

Advances in electricity market planning methods can be separated into five main groups: Traditional programming, mathematical programming, metaheuristics, agent based modelling, and integrated resource planning [14, 32]. Firstly, in the pre-1960s, traditional programming was used to determine when and where to locate generation units based on available capacity, with no focus upon consumption. Secondly, in the post-1960s, developments in operations research and increased computing power allowed for the application of mathematical programming techniques. Thirdly, metaheuristic techniques have been applied since Holland's genetic algorithm was formulated in 1975 and they are still being applied in energy market modelling to this day. Fourthly, integrated resource planning places greater emphasis on incorporating environmental cost by including emissions and greener technologies in modelling. Finally, recent studies apply agent-based modelling techniques to energy markets [32]. In this study, the simulator is a holistic representation of the energy flow supply chain. Thus, it can be viewed as an integrated resource plan.

3 Existing simulator for the energy utility

"Simulation" is a word used in many contexts, such as flight simulation package, simulation training, and computer games. For the purposes of this study, *simulation* is defined as the imitation of the energy flow from primary energy to end-use consumption.

When uncertainty exists, so does risk. Essentially, the simulator takes the form of a decision support system, allowing energy planners to manage risk. The simulator makes use of the Monte Carlo method. By incorporating the randomness of and interdependence between parameters, the Monte Carlo method models the complexity of energy flow which would otherwise be too hard to model analytically. However, in this day and age, given the scale and complexity of energy sector systems, it would still be impossible to incorporate all the detailed aspects of the system. Therefore, the simulator has been developed with the intention of modelling only the key, medium-to-high level characteristics.

Figure 1 represents an overview of the interaction between the main components of the EFS. The simulator begins with *consumption*. Simply put, consumption (demand) is fore-

cast per region and customer type. The country is split into three main regions: Central, Eastern, and Southern. The three regions are representative of three main climate regions in the country. Customers are segmented into four types: Residential, manufacturing, mining, and other. Segmenting the customers allows for more accurate demand forecasting, because each type of customer exhibits a different demand curve [12]. In addition to region and customer type, demand is forecast based on the selected Gross Domestic Product (GDP) scenario and weather scenario. GDP scenarios are defined as high, medium, and low. The EFS assumes a positive correlation between GDP and electricity demand. Weather scenarios are created by analysing historic weather data and determining profiles for hot, normal, and cold years. Based on the chosen GDP scenario (high, medium, or low) and weather scenario (hot, normal, or cold), hourly profiles per region, customer type, and month are created. Developing hourly profiles, amongst other benefits, helps to test the effect of time-of-use tariffs and demand-side management technologies.

The second main component is the *production planning* module. The load forecast is aggregated into monthly intervals. Smaller time intervals would be preferable, although in reality energy utility planners tend to plan on a month-to-month basis. The production planning module schedules the planned energy production per power station (including coal, nuclear, gas-turbine, hydro-electric, and renewable) to minimise production cost. Power stations that have cheaper production costs are scheduled first. Demand must be met whilst taking into account production capacity. A linear programming solver is the main method used for production planning.

The third main component is the *primary energy* module. Primary energy deals exclusively with coal, because coal-fired power plants produce most of the utility's electricity. The aim of the primary energy module is to manage the risk of unreliability and provide what-if scenarios. Examples of unreliability include unplanned power station maintenance, variation in the calorific value of coal, variation in the quantity of coal delivered, and variation in the coal burnt at each coal-fired power station. Furthermore, weather plays an important role in the reliability of coal, because open cast mines and stockpiles at the coal-fired power stations are not sheltered from the rain. The primary energy module incorporates and builds upon the work of Micali & Heunis [20] who developed a coal stockpile simulator. They describe the coal stockpile simulator as a dynamic model that enables what-if analysis to evaluate different plans and scenarios.

The fourth and final main component is the *generation* module. The production plan, supply reliability, and the quality and quantity of coal to be delivered are fed into the generation module. The generation module then quantifies emissions, such as sulphur and nitrogen oxides. The generation module also provides a time-of-use tariff report. Furthermore, system losses are incorporated. System losses are estimated as percentage losses and reduce the supply of electricity. Correctly estimating system losses is important to make sure enough electricity is supplied to meet demand.

The simulator consists of many more elements, which are not discussed due to limited space. The simulator has approximately 300 types of variables. Each type of variable has tens, hundreds or thousands of instances — stored in a large database. Many variables are defined by a statistical distribution, whilst other variables are assigned a deterministic value. Examples of variables are *average ash content of coal delivered*, *residential area*

demand, *mining sector time-of-use tariff*, and SO_2 *factor*.

The simulator is not an optimiser. Many variables within the simulator are known to be sub-optimal [11]. Thus, a need exists to let the simulator optimise decision-making. For this study, *simulation optimisation* is defined as the iterative process, whereby: 1) The simulator is run, and 2) The values of the chosen variables from the simulator are varied; with the process repeating itself for a set number of iterations or until solutions converge.

4 Proposed optimisation

The component with the greatest potential for optimisation is the primary energy module [11]. The primary energy module was built around the coal stockpile simulator developed by Micali & Heunis [20]. The coal stockpile simulator was specifically developed with the intention that it be optimised at a later stage.

The traditional means of determining stockpile size was to set the minimum level at 20 stockpile days and the maximum level at the capacity of the stockpile yard. A stockpile day differs between power stations. It is equivalent to the average daily mass of coal burnt. Hence, 20 stockpile days will, on average, provide a coal-fired power station with 20 days of fuel. The stockpile acts as a buffer to mitigate the effect of unforeseen events.

In the past decade, the growth in demand reduced the reserve margin (system's overall capacity). The load factor at each power station was subsequently increased. The increased load factor resulted in more coal needed to be burnt at many of the power stations, which in turn placed more pressure on coal contracts. Many coal contracts could not provide for the increased coal requirements and hence additional suppliers were contracted. Many of the additional suppliers were contracted on a short-term basis, placing more risk in the system and calling for a sophisticated coal management system. Therefore, the coal stockpile simulator was developed [20]. The coal stockpile simulator provides what-if scenarios for decision makers.

In 2008, increasing demand and a lack of additional capacity coupled with unexpected events caused nationwide blackouts. Some of the problems were attributed to depleted stockpiles. One of the actions taken by senior management was to increase the minimum stockpile level from 20 to 42 stockpile days, for all powerstations. Still the traditional approach, but now with a much higher lower limit on stockpiles levels. The traditional approach does not take into account the fact that each power station exhibits different behaviour due to variation [11, 12, 20]. Some stockpiles may require larger buffers due to high levels of variation, whilst other stockpiles may exhibit low variation and thus require much smaller stockpiles. Thus, the purpose of this research is to investigate and determine the optimal (or at least near-optimal) stockpile levels. This includes the stockpile warning limits. Due to random variation these warning limits are required.

Figure 2 depicts an example situation of the proposed optimiser, for a single power station. The 'UWL' and 'LWL' represent the upper and lower warning limits. 'DS' refers to the desired stockpile level. 'A' and 'B' are explained in the following paragraph. Before that, a short note on how coal deliveries and coal burnt are handled in the model: The deliveries are estimated by equating them to the average coal burnt. The EFS is run a set

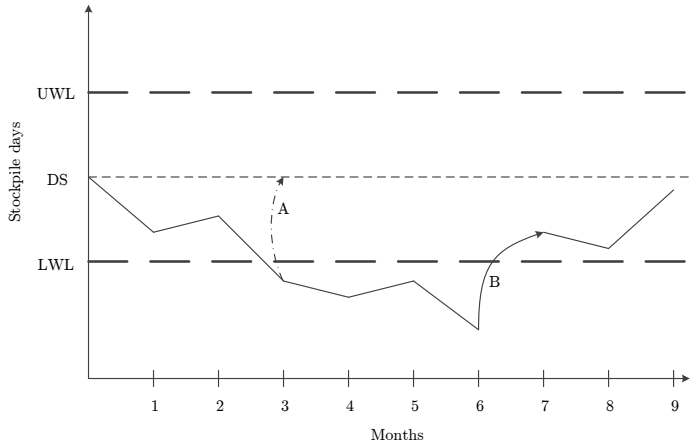


Figure 2: *The workings of the proposed optimiser.*

number of times (say 1 000) and then the average coal burnt is calculated. Subsequently, the deliveries are equated to the average coal burnt. There are other ways to determine deliveries, but this method seemed the most practical because it builds upon the work of the EFS. Even when equating the deliveries to the average coal burnt, random inherent variation in both the coal deliveries and coal burnt — especially the coal burnt — causes a resultant random variation in the stockpile levels.

Getting back to the how the proposed optimiser works. The warning limits serve as alarm sirens in two-cases: 1) Lower warning limit — to order emergency coal if the stockpile levels are too low, and 2) Upper warning limit — to cancel coal deliveries if the stockpiles are too high. Both cases are penalised. If the upper warning limit is too high, not enough cancellations will occur and then the stockpile levels have the potential to become excessively large. Likewise, if the lower warning limit is too low, then the stockpiles may stoop drastically low and be at risk of a stockpile shortage.

However, on the other hand, if either of the warning limits are too "tight" around the desired stockpile level, then the optimiser will continually penalise even the smallest variation. Figure 2 illustrates the situation when the lower warning limit comes into play. At point 'A', the stockpile drops below the lower warning limit, which results in an emergency order of coal. The quantity of the emergency order is equal to the difference in the desired level and current level. However, the order has a randomly distributed lead time (in this case three months). Between the time of order (point 'A') and delivery (point 'B'), variation continues to take place in the system. When the order arrives (point 'B') the stockpile level is closer to the desired level, but may not be exactly at the desired level.

The proposed objective of the optimisation model is to minimise coal stockpiles, coal shortage, emergency deliveries, and cancellation of deliveries. The objectives could be modelled in a multi-objective manner. However, because they can all be modelled as costs, a single objective approach is chosen. An inventory holding cost will be assigned. The coal shortage penalty is equivalent to the unserved energy penalty — which is estimated by the national government [27] — because if there is no coal available to burn, electricity

cannot be produced.

Difficulties arise in simulation optimisation because the objective function can: Exhibit various levels of simulation noise, be non-differentiable, and be computationally expensive. Fu *et al.* [13] classify six commonly used approaches in simulation optimisation: 1) Ranking and selection, 2) Response surface methodology, 3) Gradient-based procedures, 4) Random search, 5) Sample path optimisation, and 6) Metaheuristics. Metaheuristics are best suited for this simulation optimisation, because i) The search space is not limited, and ii) Near optimal solutions can be computed in a relatively short time period [15]. Metaheuristics perform well in non-linear, stochastic, dynamic environments; such as in the case of the EFS. Four metaheuristics are commonly applied to simulation optimisation: Genetic algorithms, scatter search, tabu search, and simulated annealing [13]. However, Fu *et al.* [13] proposed using another metaheuristic for simulation optimisation, namely, the cross entropy method (CEM).

The CEM fits a probability distribution on the space of solutions [21], thus making it more versatile than other metaheuristics. Fu *et al.* [13] state that the CEM shows great promise in the field of simulation optimisation, because it is not dependant explicitly on the current set of simulated values. Such versatility is beneficial in a stochastic environment where much simulation noise exists. Therefore, the CEM is the proposed algorithm for the optimiser of the simulator.

5 Road forward

As previously mentioned, this paper provides the groundwork for the second phase of the two-part study. In the second part of the study, the optimiser (in the form of the CEM algorithm) will be developed. The optimiser will be integrated with the EFS, tested and validated, and implemented.

References

[1] Asif M & Muneer T, 2007, *Energy supply, its demand and security issues for developed and emerging economies*, Renewable and Sustainable Energy Reviews, **11(7)**, pp. 1388–1413.

[2] Bhattacharyya S & Timilsina G, 2009, *Energy demand models for policy formulation: A comparative study of energy demand models.*

[3] Chikobvu D & Sigauke C, 2012, *Regression-SARIMA modeling of daily peak electricity demand in South Africa*, Journal of Energy in South Africa, **23(3)**, pp. 23–30.

[4] Conradie D, 2007, *Scheduling coal handling processes using metaheuristics*, MSc thesis, University of Pretoria.

[5] D' Sa A, 2005, *Integrated resource planning (IRP) and power sector reform in developing countries*, Energy Policy, **33(10)**, pp. 1271–1285.

[6] Dahal K & Chakpitak N, 2007, *Generator maintenance scheduling in power systems using metaheuristic-based hybrid approaches*, Electric Power Systems Research, **77(7)**, pp. 771–779.

[7] Davis S & Durbach I, 2010, *Modelling household responses to energy efficiency interventions via system dynamics and survey data*, ORiON: The Journal of ORSSA, **26(2)**, pp. 79–96.

[8] Durbach I & Davis S, 2013, *Decision support for selecting a shortlist of electricity-saving options: A modified SMAA approach*, ORiON: The Journal of ORSSA, **28(2)**, pp. 99–116.

[9] Eberhard A, 2011, *The future of South African coal: Market, investment, and policy challenges.*

[10] Ediger V & Akar S, 2007, *ARIMA forecasting of primary energy demand by fuel in Turkey*, Energy Policy, **35(3)**, pp. 1701–1708.

[11] Enerweb, 2014, *Interview with Gerard van Harmelen (Demand Intelligence Business Area Manager at Enerweb).*

[12] Enerweb, 2014, *Introduction to the energy flow simulator*, (Unpublished) Technical Report.

[13] Fu M, Glover F & April J, 2005, *Simulation Optimization: A Review, New Developments, and Applications*, Proceedings of the Winter Simulation Conference, **(1)**, pp. 83–95.

[14] Ge J, Zhang C & Du M, 2011, *Research on modeling and optimization methods of integrated resource planning*, 4th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), pp. 639–644.

[15] Gendreau M & Potvin J.-Y, 2005, *Metaheuristics in combinatorial optimization*, Annals of Operations Research, **140(1)**, pp. 189–213.

[16] Heron A, 1985, *Financing electric power in developing countries*, International Atomic Energy Agency Bulletin, **27**, pp. 44–49.

[17] Hobbs B, 1995, *Optimization methods for electric utility resource planning*, European Journal of Operational Research, **83(1)**, pp. 1–20.

[18] Liu M, Yang L & Gan D, 2005, *A survey on agent based electricity market simulation*, Power System Technology, **29(4)**, pp. 76–80.

[19] Loulou R & Labriet M, 2007, *ETSAP-TIAM: the TIMES integrated assessment model Part I: Model structure*, Computational Management Science, **5(1-2)**, pp. 7–40.

[20] Micali V & Heunis S, 2011, *Coal Stock Pile simulation*, Proceedings of the 8th Industrial and Commercial Use of Energy (ICUE), pp. 198–203.

[21] Rubinstein R & Kroese D, 2004, *The Cross-Entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*, Springer, Berlin.

[22] Schlunz E, 2011, *Decision support for generator maintenance scheduling in the energy sector*, MSc thesis, Stellenbosch University.

[23] Schramm G, 1993, *Issues and problems in the power sectors of developing countries*, Energy Policy, **21(7)**, pp. 735–747.

[24] Scott T & Read E, 1996, *Modelling hydro reservoir operation in a deregulated electricity market*, International Transactions in Operational Research, **3(3)**, pp. 243–254.

[25] Seebregts A, Goldstein G & Smekens K, 2001, *Energy/environmental modeling with the MARKAL family of models*, Proceedings of the Operations Research, pp. 44–49.

[26] Sirikum J, Techanitisawad A & Kachitvichyanukul V, 2007, *A new efficient GA-benders' decomposition method: For power generation expansion planning with emission controls*, IEEE Transactions on Power Systems, **22(3)**, pp. 1092–1100.

[27] South African Department of Energy, 2013, *Integrated resource plan for electricity (IRP) 2010-2030.*

[28] Suganthi L & Samuel A, 2012, *Energy models for demand forecasting: A review*, Renewable and Sustainable Energy Reviews, **16(2)**, pp. 1223–1240.

[29] Tekiner H, Coit D & Felder F, 2010, *Multi-period multi-objective electricity generation expansion planning problem with Monte-Carlo simulation*, Electric Power Systems Research, **80(12)**, pp. 1394–1405.

[30] Ventosa M, Baillo A, Ramos A & Rivier M, 2005, *Electricity market modeling trends*, Energy Policy, **33(7)**, pp. 897–913.

[31] Wood A, Wollenberg B & Sheble G, 2013, *Power generation, operation, and control*, 3rd Edition, John Wiley and Sons, Inc, New York.

[32] Zhou Z, Chan W & Chow J, 2009, *Agent-based simulation of electricity markets: a survey of tools*, Artificial Intelligence Review, **28(4)**, pp. 305–342.

[33] Zhu S, Wang J & Zhao W, 2011, *A seasonal hybrid procedure for electricity demand forecasting in China*, Applied Energy, **88(11)**, pp. 3807–3815.

Proceedings of the 2014 ORSSA Annual Conference
pp. 27–35
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Implementation challenges associated with a threat evaluation and weapon assignment system

DP Lötter* & JH van Vuuren†

Abstract

A threat evaluation and weapon assignment system is typically employed in a military surface-based air defence environment to provide real-time decision support to fire control officers when they have to classify incoming aircraft as threats and evaluate the perceived level of threat that these aircraft pose to defended assets on the surface. In addition, such a system is also employed to aid the operator when he has to decide on the assignment(s) of available surface weapon system(s) to neutralise these threats. In this paper, a brief review is given of the current state of a large research project aimed at threat evaluation and weapon assignment decision support designed for a surface-based air defence environment. A number of shortcomings and implementation challenges associated with this decision support system are identified and possible ideas for overcoming these shortcomings are proposed.

Key words: Threat evaluation, weapon assignment, decision support.

1 Introduction to a TEWA decision support system

A military *Surface-Based Air Defence* (SBAD) environment typically consists of *Defended Assets* (DAs) on a ground or water surface which require protection from enemy aircraft. Command centres on the surface rely on a network of sensors to detect aircraft entering the defended airspace surrounding the DAs and to provide them with important aircraft attributes, such as their courses of direction, the speeds at which they are travelling and their altitudes. A collection of surface-based *Weapon Systems* (WSs) are deployed to provide protection to the DAs from possible attacks by these aircraft.

The problem of defending DAs is commonly known in the military operations research literature as *Threat Evaluation and Weapon Assignment* (TEWA). This problem is twofold: A TE subproblem is concerned with classifying observed aircraft as hostile or friendly, evaluating the level of threat posed by hostile aircraft to the DAs and prioritising these

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: danielotter@sun.ac.za

†**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

threats accordingly, while a WA subproblem is concerned with assigning available WSs to engage these prioritised threats effectively.

A *Fire Control Officer* (FCO) is responsible for decisions related to the assignment of WSs to threats in real-time. While solving the above-mentioned subproblems may be simple when only a small number of aircraft enter the defended airspace, it becomes extremely challenging for the FCO when the defended airspace is saturated with aircraft, an attack strategy often adopted by enemy forces in an attempt to overwhelm operators. Furthermore, the speed at which enemy aircraft travel typically results in a very short time-frame during which the FCO has to solve these subproblems and make assignment decisions. Combined with the severely stressful situations in which these decisions usually have to be taken, the rapidly unfolding attack scenario may require almost super-human effort on the part of the FCO to identify good WS-to-threat assignment pairs.

One way of providing relief to the pressure experienced by an FCO is to furnish him with a computerised TEWA decision support system [10]. The aim of such a system is first to classify aircraft observed in the defended airspace as friendly, unknown or hostile, and to provide the FCO with a prioritised list of all the hostile aircraft, each with a suitable threat value assigned to it. This threat value may be an estimation of the perceived level of threat that a specific aircraft poses to a specific DA or a collection of DAs. A second aim of such a decision support system is to provide the FCO with a proposed WS-to-threat assignment list for the engagement of hostile aircraft by available WS(s) with a view to optimising some objective, such as minimising the overall accumulated survival probabilities of the hostile aircraft. The FCO may then use the information provided by the system in conjunction with his own experience and judgment to make final WS-to-threat assignment decisions.

Extensive basic research has been conducted with respect to the design of an SBAD TEWA decision support system at the TEWA Centre of Expertise of Stellenbosch University during the period 2005–2014. Although the system has been greatly refined over this period, a number of implementation challenges and shortcomings still remain. The aim in this paper is to briefly review the current local state of knowledge related to the design of a TEWA decision support system within an SBAD environment and to put forward a number of suggestions for overcoming these implementation challenges.

The paper is organised as follows. A review of the current state of local SBAD-based TEWA knowledge is provided in §2, followed, in §3, by a detailed description of the prevailing implementation challenges. The paper closes, in §4, with a number of conclusions.

2 The current state of local TEWA knowledge

Each computational cycle within a typical TEWA decision support system consists of a series of events which occur consecutively and these computational cycles are repeated until a stopping criterion is reached [10]. The natural progression of these events are illustrated graphically in Figure 1 and are discussed in some detail in this section.

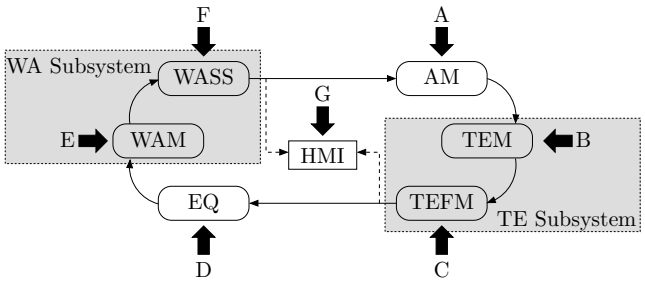


Figure 1: *Implementation challenges related to the design of an SBAD TEWA system.*

2.1 Research conducted on the TE subsystem

In 2008, Roux and Van Vuuren [11] designed a first-order automated TE subsystem architecture. In this architecture, an *Attribute Management* (AM) component analyses aircraft attributes (*e.g.* the speed or altitude at which aircraft are travelling) obtained from sensor systems, calculating a number of derived aircraft attributes (*e.g.* the acceleration of a threat) for each aircraft. This is also the first process in a TEWA computation cycle, as depicted in Figure 1. Next, the architecture includes a *Threat Evaluation Model* (TEM) component, which is the heart of the TE subsystem, and contains a number of mathematical TE models which utilise the output from the AM component to estimate the level of threat posed by each of the hostile aircraft.

Later in 2008, Heyns [3] developed four deterministic TE models for populating the TEM component described above. These models for fixed-wing aircraft are designed to evaluate a threat based solely on derived model-related attributes such as bearing, course, or course variation of aircraft — no specific consideration was given to possible weapon delivery profiles that the aircraft may execute.

Roux and Van Vuuren [11] also developed a variety of fixed wing TE models, thereby expanding the range of TE models of Heyns [3] for inclusion in the TEM component. They classified these models into three distinct hierarchical groups, based on model complexity, and proposed that the models function concurrently, with the more sophisticated, data-hungry models being phased in as high-quality data become available and they start producing realistic results.

At the lowest level of sophistication a suite of binary flagging models was proposed. These models are qualitative in nature and are only able to flag an aircraft for operator attention if there is a significant change in the observed kinematic behaviour of aircraft. These models are not able to distinguish between different levels of threat posed by aircraft.

The next suite of TE models are the deterministic models developed by Heyns [3]. These models are quantitative in nature, and are able to distinguish in a deterministic, kinematic-based manner between different levels of threat posed by aircraft.

Finally, the most sophisticated level of TE models contain a suite of stochastic models. These models are also quantitative in nature, also being able to distinguish between different levels of threat posed by aircraft, and further take into consideration enemy arsenal intelligence and doctrine when estimating a single threat value for each aircraft with re-

spect to each DA. This estimation is typically the probability that an aircraft will attack and/or kill a particular DA.

In 2013, Van Staden [14] developed a mathematical model for classifying the so-called *Formative Element Combinations* (FECs) associated with enemy aircraft — that is, the aircraft type, the weapon types carried and the aircraft attack technique adopted rather than estimating these parameters individually based on expert judgement and pre-deployment intelligence reports, as originally suggested by Roux and Van Vuuren [11]. The model is based on a hidden Markov modelling paradigm and predicts the most probable attack technique adopted by aerial threats, based on their observed kinematic data. This information may then be used in conjunction with the enemy's known arsenal of aircraft and WS types carried to determine the most probable FEC for each threat. Incorporating the results of this model into the stochastic TE models is expected to yield more reliable estimated threat values of aircraft.

The final component in the TE subsystem architecture of Roux and Van Vuuren [11] is a *Threat Evaluation Model Fusion* (TEFM) component. This component is designed to combine the results produced by the various models in the TEM component so as to produce a global threat value for each aircraft. This is achieved by a multi-criteria decision analysis technique, such as a value function procedure or additive model.

The results of the TEFM are relayed to the FCO via a *Human Machine Interface* (HMI), which displays the airpicture as well as the threat values of the various aircraft and other TEWA-related information on a series of computer screens.

2.2 Research conducted on the WA subsystem

In 2008, Potgieter [8] proposed a basic first-order WA subsystem architecture. The design included an *Engagement Efficiency Matrix* (EEM) component in which the efficiency values achieved by WSs, when assigned to engage threats, are discretised and filtered for external elements (such as adverse weather conditions and/or terrain feature interference). Furthermore, the design included a model framework component, which is the heart of the WA subsystem, and contains a variety of mathematical assignment models for solving the WA subproblem. This component uses the results of the TE subsystem and the output from the EEM component to propose the assignment of WSs to threats. A number of WA models (including models of a static[1] and dynamic[2] nature) as well as rule-based weapon assignment heuristics, which may be used in this framework, were also presented.

Du Toit [1] built on the dynamic WA models of Potgieter [8] by formulating the WA problem dynamically in two different ways in 2009 — first under the assumption that the number of threats and locations of targets are all known in advance and secondly under the assumption that not all targets are observable (at each time interval the locations of targets present is only known stochastically).

[1] In the context of WA models, the term *static* refers to models in which the numbers and locations of WSs and threats are known with certainty at some time instant τ and a single assignment of WSs to threats is sought at time τ such that all the WSs are committed [1].

[2] In contrast, the term *dynamic* refers to the class of WA models in which suitable future time instants are sought at which to assign a subset of the available WSs to the threats observed [13].

In 2013, Lötter *et al.* [6] modelled the WA problem as a multi-objective decision problem in which a number of objectives are pursued simultaneously. The research included the identification of useful objectives by applying objective identification techniques from the multi-criteria decision analysis literature to a carefully selected audience of military experts. Two of these objectives were used to formulate a bi-objective, static WA model.

Also in 2013, Van der Merwe and Van Vuuren [13] modelled the WA problem in a dynamic framework as a vehicle routing problem with time windows in which WSs are modelled as vehicles having to deliver commodities (ammunition) to customers (threats). The use of time windows, in the sense that the model is required to suggest time frames during which WSs should assign threats, adds a scheduling element to the WA problem. A hybrid approximate solution approach towards solving the model was also proposed, based on the metaheuristics of simulated annealing and tabu search.

Lötter and Van Vuuren [7] went on to design an improved WA subsystem architecture for use in the context of a SBAD environment in 2014. The design provides for an *Engagement Quantisation* (EQ) component, a *Weapon Assignment Model* (WAM) component and a *Weapon Assignment Solution Selection* (WASS) component, as depicted in Figure 1. This is also the order in which events occur in a TEWA computational cycle. The core of the EQ component rests on the EEM component proposed earlier by Potgieter [8], while the WAM component serves the same purpose as the model framework component proposed by him. However, the WAM component is designed to include four classes of WAMs ranging in different levels of complexity, from which the FCO may configure a model for use in the WAM component before or during a combat situation.

The least complex class of WAMs proposed by Lötter and Van Vuuren [7] involve a single-objective WAM in a static framework. The next class of WAMs contains multi-objective, static WAMs. They differ from the first class of WAMs in the sense that they consider multiple objectives simultaneously when proposing the assignments of WSs to threats. The next level of WAMs contains single-objective, dynamic WAMs. Although these models accommodate only a single objective when considering the assignment of WSs to threats, they include a dynamic scheduling element. The final class of WAMs is also the most complex class of WAMs, containing multi-objective, dynamic WAMs. These models consider multiple objectives over the entire time continuum to propose assignments of WSs to threats as well as the scheduling of appropriate time windows for the assignments.

The final component in the WA subsystem architecture of Lötter and Van Vuuren [7] is the WASS, which employs various solution techniques to solve the WAM configured by the FCO in order to produce a collection of WS-to-threat assignment decision alternatives. The WASS component combines all these results and filter out dominated solutions by employing a sorting algorithm, so as to present the FCO with a set of Pareto-optimal solutions via the HMI.

3 Implementation challenges associated with a TEWA DSS

A number of shortcomings and implementation challenges have been identified within the components of the TEWA system design described in §2. These challenges involve (1)

the quality and quantity of TE-related input data (which may affect the working of the AM and TEM components, as indicated by the arrows labelled A and B in Figure 1), (2) the problem of potentially overwhelming the FCO with excessive decision support information (which may affect the implementation of the TEFM, WASS and HMI components, as indicated by the arrows labelled C, F and G in Figure 1), (3) incorporating FCO preferences and biases into TEWA results (which may affect the results presented by the WASS component, as indicated by the arrow labelled F in Figure 1), (4) the problem of potentially rapid switching of TEWA decision support results over time (as a result of the suggested working of the TEFM and WASS components, as indicated by the arrows labelled C and F in Figure 1), and (5) the requirement of testing and evaluating the level of performance of the TEWA system as an integrated system (which involves all the components, as indicated by the arrows labelled A – F in Figure 1).

3.1 Quality and quantity of TE-related data

A TE subsystem will only be able to perform to its full potential if sufficient quality and quantity of input data are available. These data are typically provided by sensor systems and intelligence reports and should be analysed and preprocessed thoroughly in order to provide high-quality input to the TE subsystem. A number of ways in which the quantity and quality of these input data may be improved are outlined in this section.

The scope of the TE subsystem design of Roux and Van Vuuren [11], as well as the FEC model proposed by Van Staden [14], were restricted to only include fixed wing aircraft. It may, however, be beneficial to expand the scope of the aerial threats by additionally including models for other platform types, such as rotary wing aircraft [9, 14]. By expanding the range of platform types, a more realistic TE subsystem may be obtained.

Furthermore, the identification of influential measured and derived aircraft attributes (for all aircraft types) obtainable from sensors and intelligence reports may also result in more accurate and reliable FEC classification of aircraft as well as a more appropriate estimation of the level of threat posed by aircraft. However, the availability and classification of such data are typically restricted [9]. If such information were available, existing data mining procedures may be employed for extracting significant measured attributes or discovering derived attributes from the data which influence aircraft threat values significantly.

3.2 Overwhelming the operator with information

The objective of a TEWA system is ultimately the provision of high-quality TEWA solution suggestions at any given time stage. It is, however, important to provide the FCO with only the information that he needs rather than to provide him with excessive information which may overwhelm him and may compromise his ability to make effective WS-to-threat assignment decisions. On the other hand, a sufficient store of information should be available in case the FCO wishes to access more detailed information in order to motivate decisions. When designing a TEWA HMI display, careful consideration should thus be given as to what information is deemed important to provide to the FCO. Gruhn [2] suggests that an effective HMI should be based on a user-centered design which integrates information in ways that fit the tasks and needs of the user. This implies that the FCO

should be included in the design of an effective HMI and that he should be able to configure the HMI pre-deployment and even during a mission.

One way of minimising clutter when designing an HMI is to hide excess information by employing pop-up windows on the HMI display screen. The FCO may then use a computer mouse to hover over a particular solution, resulting in the opening of a pop-up window which displays more detailed information related to a suggested WS-to-threat assignment suggestion. For example, in the case where a list or graph of approximately Pareto-optimal solutions (in objective space) is presented to the FCO, the pop-up windows may display the actual assignments of WSs to threats (in solution space) when hovering over one of the solutions. When the FCO then chooses one of these solutions, the progress of unfolding assignments may then be displayed on the screen over time.

3.3 Incorporating FCO preferences and biases

Furthermore, care should also be taken in the implementation of the WASS component proposed by Lötter and Van Vuuren [7], since presenting the FCO with too many approximately Pareto-optimal solutions may cause indecision on the part of the FCO when picking one of these solutions for implementation. One way of reducing the number of solutions presented is to ask the FCO to specify sufficient bounds on the objective function values.

Another way of reducing the number of solutions presented to the FCO is to filter out approximately Pareto-optimal solutions from the suggested list according to the operator's biases and preferences, by employing a pre-determined FCO utility function.

3.4 Switching of results between consecutive time stages

An undesirable phenomenon, which may occur when WS-to-threat assignment suggestions are reported to the FCO during a combat situation, is one called *switching*. Switching refers to the excessively rapid changing of assignment suggestions during a small subset of consecutive time stages. This kind of behaviour may be ascribed to small changes in the single shot hit probabilities that WSs are capable of achieving with respect to threats during these time stages. However, switching may cause confusion and compromise the FCO's confidence in the results produced by the TEWA system which may, in turn, lead to the FCO making sub-optimal decisions when choosing to rely on his own judgment rather than trusting the seeming indecision of the decision support system.

The problem of switching may be solved by implementing threshold values in the system in such a way that assignment suggestions are only allowed to change from one time stage to another once a variation in the results equivalent to the threshold value is reached (*i.e.* if the two solutions in question are deemed significantly different).

3.5 Evaluating the performance of the system

Although the designs of the TE and WA subsystems proposed by Roux and Van Vuuren [11] and by Lötter and Van Vuuren [7], respectively, seem to be able to provide acceptable quality decision support to FCOs, the performances of these subsystems have not

yet been tested in an integrated manner. It may be useful to employ a military expert or a group of military experts to evaluate and analyse the results produced by these subsystems. A more robust method of evaluating TEWA system performance is, however, required. It is also important that this method of evaluation be generic in the sense that should future changes be made to any of the components in the TEWA subsystems, the method should be easily adaptable to incorporate these changes and to re-evaluate the system's performance.

Truter and Van Vuuren [12] are currently designing various measures for evaluating the performance of TEWA systems. Roux [9] suggested that a simulation environment be used to test and evaluate a TEWA system's performance and that testing procedures be performed in an incremental manner. First-order examples of evaluating the performance of TEWA systems in this manner were put forward by Kok [5] and by Johansson and Falke [4].

4 Conclusion and discussion

A brief review of the design of a TEWA decision support system within an SBAD environment was provided in this paper, touching on the functions of and interactions between the various components and substructures comprising such a system. Furthermore, a number of important concerns were raised in terms of testing and implementing the system. Finally, some suggestions were made with respect to overcoming these concerns.

References

[1] du Toit FJ, 2009, *The dynamic weapon target assignment problem in a ground-based air defence environment*, MSc Thesis, Stellenbosch University, Stellenbosch.

[2] Gruhn P, 2011, *Human Machine Interface (HMI) design: The good, the bad and the ugly*, 66th Annual Instrumentation Symposium for the Process Industries, 27–29 January 2011.

[3] Heyns AM, 2008, *Measuring the threat value of fixed wing aircraft in a ground-based air defence environment*, MSc Thesis, Stellenbosch University, Stellenbosch.

[4] Johansson F & Falkman G, 2009, *Performance evaluation of TEWA systems for improved decision support*, (Unpublished) Technical report, Department of computer Science, University of Skövde, Skövde.

[5] Kok BJ, 2009, *Performance evaluation of a threat evaluation and weapon assignment system*, MSc Thesis, Stellenbosch University, Stellenbosch.

[6] Lötter DP, Nieuwoudt I & van Vuuren JH, 2013, *A multiobjective approach towards weapon assignment in a ground-based air defence environment*, ORiON, **29(1)**, pp. 31–54.

[7] Lötter DP & van Vuuren JH, 2014, *Weapon assignment decision support in a surface-based air defence environment*, Military Operations Research, Submitted.

[8] Potgieter JJ, 2008, *Real-Time weapon assignment in a ground-based air defence environment*, MSc Thesis, Stellenbosch University, Stellenbosch.

[9] Roux JN, 2008, *Design of a threat evaluation subsystem in a ground-based air defence environment*, PhD Dissertation, Stellenbosch University, Stellenbosch.

[10] Roux JN & van Vuuren JH, 2007, *Threat evaluation and weapon assignment decision support: A review of the state of the art*, ORiON, **23(2)**, pp. 151–187.

[11] Roux JN & van Vuuren JH, 2008, *Real-time threat evaluation in a ground-based air defence environment*, ORiON, **24(1)**, pp. 75–101.

[12] Truter ML & van Vuuren JH, 2014, *Prerequisites for the design and evaluation of a novel threat evaluation and weapon assignment system*, Proceedings of the Annual ORSSA Conference, pp. 54–61.

[13] van der Merwe M & van Vuuren JH, 2013, *The weapon assignment scheduling problem in a ground-based air defence environment*, Submitted.

[14] van Staden HE & van Vuuren JH, 2013, *Attack technique classification in a ground-based air defence environment*, MComm Thesis, Stellenbosch University, Stellenbosch.

Proceedings of the 2014 ORSSA Annual Conference
pp. 36–44
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Maintenance scheduling for the generating units of a national power utility

BG Lindner* JH van Vuuren†

Abstract

Reliable energy provision is a major force in shaping the economic welfare of a developing country. For a power utility in such a country one of the key focus areas is the planned preventative maintenance of the power generating units in its generation system so as to ensure that it is in a position to satisfy power demand in a reliable manner. In the *generator maintenance scheduling* (GMS) problem, the objective is to find a schedule for the planned maintenance outages of generating units in a power system which minimises maintenance costs or maximises the probability of meeting a safety margin over and above the national power demand, which is a function of time. Previous work on the GMS problem includes the use of mixed integer programming techniques and metaheuristics to find good generator maintenance schedules. This paper builds on these approaches by advocating use of a decision support system aimed at determining good generator maintenance schedules by taking into account (1) the levels and qualities of fuel stockpiles at generating units, (2) unplanned and other energy loss factors, (3) adopting a multi-objective optimisation approach instead of a single-objective approach as is usual in the literature and (4) analysing the possible interaction between inputs and outputs from the GMS problem and other energy components of the energy supply chain of a power utility.

Key words: Electricity industry, energy sector, maintenance scheduling, simulated annealing.

1 Introduction

One of the key focus areas for a power utility is the planned preventative maintenance of the power generating units in a power system [4, 5, 9, 13] so as to satisfy demand as efficiently and effectively as possible, a problem often referred to as the *generator maintenance scheduling* (GMS) problem. The maintenance scheduling of generators has been studied and analysed by many researchers [14] and this paper contains four suggestions as to how the level of realism of GMS problem formulations may be improved. These suggested improvements are (1) taking cognisance of the fuel stockpile levels associated with generating

*Department of Industrial Engineering, Stellenbosch University, Private bag X1, Matieland, 7602, South Africa, fax: 021 8082409, email: 15150526@sun.ac.za

†**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

units when scheduling generator maintenance, (2) including unplanned and other generation loss factors in generator maintenance planning, (3) adopting a multi-objective optimisation approach toward generator maintenance planning and (4) analysing the possible interaction between generator maintenance planning and other energy planning models.

Power utilities often employ large-scale energy flow simulation models of their energy supply chains to inform decisions on operational and strategic levels. These energy system models typically interconnect the conversion and consumption of energy [18] and include operations involved with primary fuel supplies (*e.g.* mining, petroleum extraction), conversion and processing (*e.g.* power plants, refineries), and end-use demand for energy services (boilers, residential space conditioning). The demand for energy is normally disaggregated by sector (*i.e.* residential, manufacturing, transportation, and commercial) and by specific functions within a sector (*e.g.* residential air conditioning, heating, lighting, hot water) [18]. These energy flow models serve to facilitate the investigation of what-if scenarios for decision makers [12], with some utilising optimisation techniques [6, 18]. It is within this decision support framework that maintenance scheduling solutions are expected to be incorporated in a dynamic fashion.

This paper is organised as follows. After conducting a brief survey of GMS problem formulations and solution techniques from the literature in §2, descriptions of the abovementioned four proposed improvements to GMS formulations are described in §3. This is followed by a discussion on the feasibility of the proposed model improvements in §4 and finally some concluding remarks in §5.

2 Literature review

Although the GMS problem is related to a number of classical optimisation problems, such as the assignment problem, the travelling salesman problem and the vehicle routing problem, it is not one of these [15]. Factors complicating formulations of the GMS problem result from attempts at incorporating the fact that generated electricity cannot be stored; that the transmission network is limited and hence that a required amount of electricity must be generated at every instant; that an adequate amount of reserve capacity has to be available at all times; and the parallel nature of electricity supply within a power system (due to multiple generating units) [9, 15].

2.1 GMS problem formulation

In the literature related to the GMS problem, a dominant objective is usually included in model formulations as a single function to be optimised, while the lesser important objectives are incorporated as constraints. The most typical objectives found in literature are based on economic criteria, reliability criteria, and convenience criteria [9, 16].

Economic criteria. The most common economic objectives consist of minimising the total operating cost associated with a generator maintenance schedule, which includes energy production and maintenance cost [4]. Energy production costs include fuel costs, salaries and wages, costs related to energy production and generator start-up

and shut-down costs. Maintenance costs, on the other hand, include replacement part costs and salaries and wages related to unit maintenance [15]. These economic costs typically vary from generating unit to generating unit and data related to these costs are sometimes difficult to obtain [15].

Reliability criteria. Reliability objectives include minimising the expected lack of peak net reserve, expected energy not supplied or loss of load probability [4]. These reliability criteria may be either stochastic or deterministic in nature [13]. The most common choice is to minimise the reserve load, usually formulated as the sum of squares of the reserve [13, 17], because data for this criterion are usually more easily obtainable [15]. Another option is to maximise the smallest reserve load during any time period. For some power utilities, reliability objectives are more important than economic considerations [15].

Convenience criteria. Examples of convenience criteria are minimising soft constraint violations or minimising possible disruptions to the power generation schedule [15].

As mentioned, researchers have mostly adopted a single-objective optimisation approach towards formulating instances of the GMS problem [16]. Since the criteria mentioned above are, however, often conflicting in nature, the problem is inherently multi-objective.

The constraints included in formulations of the GMS problem can vary significantly, depending on the nature and assumptions of the power utility's operations [15]. Typical constraints employed in the literature include the following [1, 9].

Maintenance window constraints ensure that each unit is serviced between an earliest and latest time period. These time windows are typically dictated by annual generating unit service frequencies, as imposed either by power utility policy or by operational service levels.

Load constraints ensure that the load demand is met during each time period. This demand must, of course, be met by generating units that are not scheduled for maintenance during the relevant time period.

Reliability constraints may be incorporated by specifying a reserve/safety margin over and above the load constraints.

Service contiguity constraints are imposed to ensure that the number of time periods required to service a particular generating unit run consecutively over time.

Resource constraints specify a limit on the amount of resources available for the purpose of maintenance. These resources may involve service budgets, the availability of adequately qualified service personnel and the availability of spare parts.

Exclusion constraints are used when certain generating units are not allowed to be taken out of service simultaneously (*e.g.* two units in the same power station or too many units in the same geographical region).

Transmission/network constraints have been incorporated recently and seek to ensure the transmission capabilities of the electrical network (*e.g.* maintaining voltage levels) or that a power station meets the demands of the geographic regions within its service area via the transmission network infrastructure.

Maintenance plans usually span an annual time horizon [10, 15], but this can vary, and planning horizons in the literature range from eight weeks to five years. Common time intervals include one week [10], but this also varies with values ranging in the literature from single-day and five-day to monthly intervals [15].

2.2 Solution techniques

According to [1, 15] the most prevalent solution methods applied to solve instances of the GMS problem include heuristic search algorithms, mathematical programming techniques, dynamic programming, expert systems, fuzzy systems, and metaheuristics:

Heuristic search algorithms search and improve upon the quality of solutions based on trial and error, and are comparatively seldom used [1] due to the inferior quality solutions that they often produce.

Mathematical programming techniques are typically used for single objective instances of the GMS problem, and mostly include variations of the branch-and-bound method. Further methods include the generalised reduced gradient algorithm for nonlinear programming problems and successive linear programming, amongst other methods.

Dynamic programming ideally suits the temporal nature of maintenance scheduling problems [15] and has been used in the context of the GMS problem in [7, 8].

Expert systems develop an automated solution methodology by imitating the many years of experience of experts in the field [16].

Fuzzy set theory is employed to address multiple objectives and uncertainties in the constraints [16] and has been used in [4, 7].

Metaheuristics are used when the dimensions of a GMS problem instance increases to the point where exact solution methodologies take too long to implement. These techniques then often obtain very good (although not necessarily optimal) solutions within more acceptable computation time frames. Recently, metaheuristics have been used to solve GMS problem instances close to optimality within very limited computational times [15]. Typical metaheuristics applied to the GMS problem include genetic algorithms, simulated annealing, tabu search and ant colony optimisation.

3 Proposed adaptations to the GMS problem

We propose to build upon previous GMS-related work, most notably by Schlünz [15], by incorporating two important additional notions into the GMS problem formulation, namely (1) the level and quality of fuel stockpiles used for electricity generation and (2) unplanned and other loss factors related to energy generation.

In addition, we also advocate the simultaneous adoption of two main scheduling objectives, namely to seek acceptable trade-offs between minimising the cost associated with a generator maintenance schedule and maximising the reliability of the generating programme which results from a maintenance schedule.

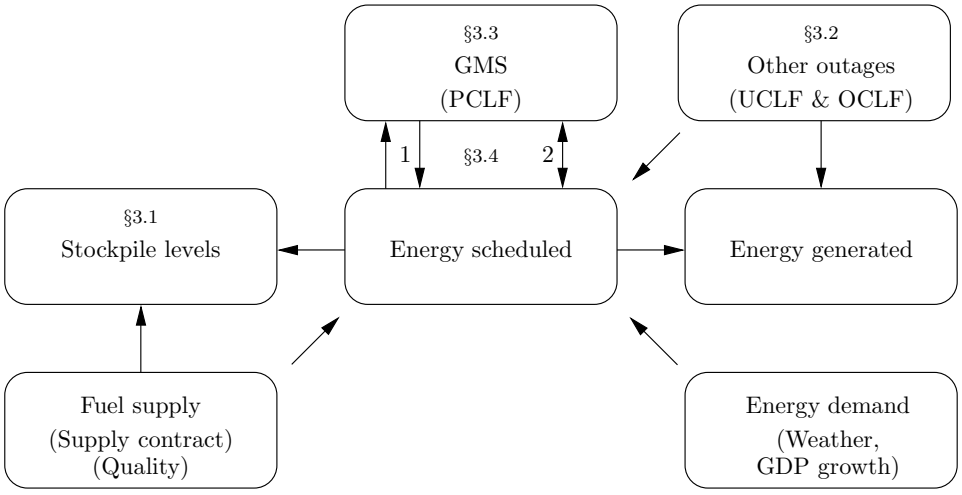


Figure 1: *Interactions between the GMS problem and the typical supply chain components of a national power utility.*

Further elucidation of how this improved GMS formulation is expected to interact with the most important components of a typical simulation model of a power utility's energy supply chain may be found in Figure 1.

3.1 Including fuel reserves in the formulation

It is important for a power utility to plan and adequately manage its fuel stockpile levels, because excess quantities of coal (one of the main fuel types used; the others being natural gas, water and uranium [15]) constitute an asset which is not producing revenue and hence incurs lost interest charges. Large coal stockpiles also require more careful management to ensure that the commodity is stored safely and does not deteriorate unduly [20]. One major problem experienced is that the coal sometimes becomes too wet as a result of intentional irrigation so as to avoid the possibility of spontaneous combustion within a stockpile [2]. Significantly increasing the moisture level of coal can reduce its combustion efficiency [15, 20]. Further coal-quality deterioration can occur on sunny days or as a result of intermittent rain [2].

Very low levels of stockpiles, on the other hand, raise the risk of a generating unit running out of fuel, which means that it will not be able to meet its expected demand [20].

The fuel stockpile of a coal power station, for example, varies according to the conservation law

$$\begin{pmatrix}\text{Stockpile levels}\\ \text{at end of period } t\end{pmatrix} = \begin{pmatrix}\text{Stockpile levels}\\ \text{at start of period } t\end{pmatrix} + \begin{pmatrix}\text{Coal delivered}\\ \text{during period } t\end{pmatrix} - \begin{pmatrix}\text{Coal burnt}\\ \text{during period } t\end{pmatrix}, \quad (1)$$

where

$$\begin{pmatrix}\text{Coal delivered}\\ \text{during period } t\end{pmatrix} = \begin{pmatrix}\text{Coal contract}\\ \text{during period } t\end{pmatrix} \pm \begin{pmatrix}\text{Delivery uncertainty}\\ \text{during period } t\end{pmatrix} \quad (2)$$

and

$$\begin{pmatrix}\text{Coal burnt}\\ \text{during period } t\end{pmatrix} = \begin{pmatrix}\text{Energy generated}\\ \text{during period } t\end{pmatrix} \times \begin{pmatrix}\underline{\text{Heat rate}}\\ \text{CV}\end{pmatrix}. \tag{3}$$

In (3), the *calorific value* (CV) [measured in MJ/kg] is the potential energy locked up in the coal that can be converted to actual heating ability. The *heat rate* [measured in MJ/kWh] is the amount of thermal energy required to generate one KWh of electrical energy [10] and thus indicates a generating unit's efficiency to convert its fuel to electricity. Finally, the *energy generated* in (3) is governed by the relationship

$$\begin{pmatrix}\text{Energy generated}\\ \text{during period } t\end{pmatrix} = \begin{pmatrix}\text{Energy scheduled}\\ \text{during period } t\end{pmatrix} + \begin{pmatrix}\text{Additional load}\\ \text{during period } t\end{pmatrix} - \begin{pmatrix}\text{Outages}\\ \text{during period } t\end{pmatrix}. \tag{4}$$

The *energy scheduled* in (4) is usually determined by solving the problem of meeting specific electricity sector demands from the power generating units that are not scheduled for preventative maintenance during period t.

It is proposed that both the qualities and levels of fuel stockpiles should be taken into account when solving the GMS problem. It may, for example, be advantageous, in terms of time bought for stockpile replenishment to move the service time of a generating unit forward (within its window of acceptable service times) if its stockpile level or quality is observed to be dangerously low. Previous work by Schlünz and Van Vuuren [16, 17] did not include these values, but assumed that stockpile levels would always be within adequate margins.

3.2 Including generation loss factors in the formulation

Outages are a function of the *planned capability loss factors* (PCLFs), the *unplanned capability loss factors* (UCLFs), *other capability loss factors* (OCLFs) and the installed energy. This function often takes the form

$$\text{Outages} = (\text{PCLF}+\text{UCLF}+\text{OCLF}) \times \text{Installed energy}, \tag{5}$$

where PCLFs are power generation losses specifically planned by the management of a power utility for maintenance purposes and other shutdowns, UCLFs include losses due to weather conditions and transmission line failures, and OCLFs are other losses due to events outside the control of the management of a power utility [11]. The *installed energy* in (5) is the combined generating capacity of all generating units under consideration.

Schlünz and Van Vuuren [16] took unplanned and other capacity loss factors into account by representing them all as a single safety factor. However, we propose a more detailed analysis of what these values typically are, and incorporation of the *outages* value in (5) into the GMS problem formulation.

3.3 Adopting a multi-objective optimisation approach

Simulated annealing has previously been used [3, 4, 14] to solve instances of single-objective formulations of the GMS problem. We propose rather formulating the problem as a bi-objective problem, simultaneously minimising the cost associated with generator maintenance scheduling and maximising the reliability of the resulting generating programme,

and using a multi-objective version of a neighbourhood metaheuristic search technique, such as simulated annealing, to find acceptable trade-offs between the values of these objectives.

3.4 Analysis of interaction with other simulation models

It is important to note that energy scheduled and power loss outages in (4) have further interlinked inputs and outputs in the energy supply chain of a power utility, as illustrated in Figure 1. The generator maintenance scheduling problem is typically solved based on parametric values representing the amount of energy generation required, or scheduled, at a power station. These values are, in turn, determined by the maintenance schedule's output, in terms of expected outages. This interaction is one-way (the arrows labelled 1 in Figure 1), *i.e.* when solving for typical decision variables of the energy scheduling problem the maintenance schedules of the various generating units are incorporated into the model as parameters, not as linked decision variables, and *vice versa*. However, we propose that these operational decision components interact dynamically (the arrow labelled 2 in Figure 1) within such a simulation and/or optimisation framework.

4 Feasibility of the proposed approach

The authors plan on demonstrating the feasibility of the GMP model formulation enhancements proposed in §3 in a real South African case study. The case company currently utilises a state-of-the-art computerised simulation tool, called the *Energy Flow Simulator* (EFS) in aid of long-term and strategic decision making [19]. This tool is currently capable of simulating the entire energy supply and demand chain in South Africa, and contains components simulating different weather conditions and economic trends, energy load scenarios, the quality and quantity of the national coal stockpile and the effectiveness of energy generation schedules for the different generating units [12, 19]. The EFS, however, does *not* currently have the capability of incorporating generating unit maintenance scheduling in its energy generation planning component. It is within the existing framework of the EFS that the authors plan to implement the above-mentioned enhanced GMP model formulation. This framework will allow for adequate testing of the robustness and efficacy of simultaneous schedule production for energy generation and generator unit maintenance downtimes.

5 Conclusion

In this paper, we proposed a number of adaptations to typical formulations of the GMS problem. Two of these adaptations were concerned with improvements to the level of realism of the formulation (incorporating the quality and level of the fuel stockpile associated with each generating unit and including a suite of power generation loss factors). Further adaptations were related to the paradigm in which the optimisation takes place (a bi-objective optimisation approach was suggested) and the way in which the inputs to and the outputs from the GMS problem interact with other components of the energy

supply chain of a power utility. The paper is a report on work in progress within a larger research project at Stellenbosch University. The next step in this research project will be to attempt implementations of these adaptations, starting with the fuel stockpiles.

References

[1] Ahmad A & Kothari DP, 1998, *A review of recent advances in generator maintenance scheduling*, Electric Machines & Power Systems, **26(4)**, pp. 373–387.

[2] Banerjee D, Hirani M & Sanyal S, 2000, *Coal-quality deterioration in a coal stack of a power station*, Applied Energy, **66(3)**, pp. 267–275.

[3] Burke E & Smith A, 2000, *Hybrid evolutionary techniques for the maintenance scheduling problem*, IEEE Transactions on Power Systems, **I(I)**, pp. 122–128.

[4] Dahal KP & Chakpitak N, 2007, *Generator maintenance scheduling in power systems using metaheuristic-based hybrid approaches*, Electric Power Systems Research, **77(7)**, pp. 771–779.

[5] Dahal K, McDonald J & Burt G, 2000, *Modern heuristic techniques for scheduling generator maintenance in power systems*, Transactions of the Institute of Measurement and Control, **22(2)**, pp. 179–194.

[6] Energy Research Centre, 2014, *Assumptions and methodologies in the South African TIMES (SATIM) energy model*, URL: `http://www.erc.uct.ac.za/Research/Otherdocs/Satim/SATIM%20Methodology-v2.1.pdf`.

[7] Huang C, Lin C & Huang C, 1992, *Fuzzy approach for generator maintenance scheduling*, Electric Power Systems Research, **24(1)**, pp. 31–38.

[8] Huang SJ, 1997, *Generator maintenance scheduling: A fuzzy system approach with genetic enhancement*, Electric Power Systems Research, **41(3)**, pp. 233–239.

[9] Kralj B & Petrović R, 1988, *Optimal preventive maintenance scheduling of thermal generating units in power systems – A survey of problem formulations and solution methods*, European Journal of Operational Research, **35(1)**, pp. 1–15.

[10] Kralj B & Petrovic R, 1995, *A multiobjective optimization approach to thermal generating units maintenance scheduling*, European Journal of Operational Research, **84(2)**, pp. 481–493.

[11] Micali V, 2012, *Prediction of availability for new power plant in the absence of data*, Proceedings of the The Industrial and Commercial Use of Energy (ICUE), Proceedings of the of the 9th IEEE Concference, pp. 1–8.

[12] Micali V & Heunis S, 2011, *Coal Stock Pile simulation*, Proceedings of the The Industrial and Commercial Use of Energy (ICUE), Proceedings of the 8th IEEE Concference, pp. 198–203.

[13] Mohanta DK, Sadhu PK & Chakrabarti R, 2007, *Deterministic and stochastic approach for safety and reliability optimization of captive power plant maintenance scheduling using GA/SA-based hybrid techniques: A comparison of results*, Reliability Engineering & System Safety, **92(2)**, pp. 187–199.

[14] Saraiva JT, Pereira ML, Mendes VT & Sousa JC, 2011, *A simulated annealing based approach to solve the generator maintenance scheduling problem*, Electric Power Systems Research, **81(7)**, pp. 1283–1291.

[15] Schlünz EB, 2011, *Decision support for generator maintenance scheduling in the energy sector*, MSc Thesis, Stellenbosch University, Stellenbosch.

[16] Schlünz EB & Van Vuuren JH, 2013, *An investigation into the effectiveness of simulated annealing as a solution approach for the generator maintenance scheduling problem*, International Journal of Electrical Power & Energy Systems, **53**, pp. 166–174.

[17] Schlünz EB & Van Vuuren JH, 2012, *The application of a computerised decision support system for generator maintenance scheduling: A South African case study*, South African Journal of Industrial Engineering, **23(3)**, pp. 169–179.

[18] Seebregts A, Goldstein G & Smekens K, 2002, *Energy/environmental modeling with the MARKAL family of models*, In Operations Research Proceedings 2001, pp. 75–82.

[19] Van Harmelen G, 2014, *Utility Analytics Business Area Manager at Enerweb*, [Personal Communication], Contactable at `gerard.van.harmelen@enerweb.co.za`.

[20] Whittington H & Bellhouse G, 2000, *Coal-fired generation in a privatised electricity supply industry*, International Journal of Electrical Power & Energy Systems, **22(3)**, pp. 205–212.

Proceedings of the 2014 ORSSA Annual Conference
pp. 45–53
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

On the q-criticality of graphs with respect to secure graph domination

AP Burger*, AP de Villiers† & JH van Vuuren‡

Abstract

A subset X of the vertex set of a graph G is a *secure dominating set* of G if each vertex of G which is not in X is adjacent to some vertex in X and if, for each vertex u not in X, there is a neighbouring vertex v of u in X such that the swap set $(X - \{v\}) \cup \{u\}$ is again a dominating set of G. The *secure domination number* of G is the cardinality of a smallest secure dominating set of G.

The notion of secure graph domination finds applications in the generic setting where the vertex set of G represents distributed locations in some spatial domain and the edges of G represent links between these locations. Patrolling guards, each stationed at one of these locations, may move along the links in order to protect the graph. A minimum secure dominating set of G then represents a smallest collection of locations at which guards may be stationed so that the entire location complex modelled by G is protected in the sense that if a security concern arises at location u, there will either be a guard stationed at that location who can deal with the problem, or else a guard dealing with the problem from an adjacent location v will still leave the location complex protected after moving from location v to location u in order to deal with the problem.

A graph G is *q-critical* if the smallest arbitrary subset of edges whose removal from G necessarily increases the secure domination number, has cardinality q. The notion of q-criticality is important in applications such as the one mentioned above, because it provides threshold information as to the number of edge failures (perhaps caused by an adversary) that will necessitate the hiring of additional guards to secure the location complex.

Denote by Ω_n the largest value of q for which q-critical graphs of order n exist. It has previously been established that $\Omega_2 = 1$, $\Omega_3 = 2$, $\Omega_4 = 4$, $\Omega_5 = 6$ and $\Omega_6 = 9$. In this paper we present a repository of all q-critical graphs of orders $2, 3, 4, 5$ and 6 for all admissible values of q and we also establish the previously unknown values $\Omega_7 = 12$, $\Omega_8 = 17$ and $\Omega_9 = 23$. These values support an existing conjecture that $\Omega_n = \binom{n}{2} - 2n + 5$ for all $n \geq 7$.

Key words: Secure domination, graph protection, edge criticality.

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: apburger@sun.ac.za

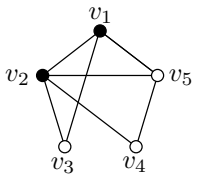
†Corresponding author: Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: antondev@sun.ac.za

‡**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

1 Introduction

A *dominating set* of a graph G is a subset X of the vertex set of G with the property that each vertex of G not in X is adjacent to at least one vertex in X. A *secure dominating set* of G is a subset X_s of the vertex set of G with the property that X_s forms a dominating set of G and, additionally, for each vertex u not in X_s, there exists a vertex $v \in X_s$ for which the *swap set* $(X_s - \{v\}) \cup \{u\}$ is again a dominating set of G. The *secure domination number* of G, denoted by $\gamma_s(G)$, is the minimum value of $|X_s|$, taken over all secure dominating sets X_s of G (*i.e.* the cardinality of a smallest secure dominating set of G). A number of general bounds have been established for the parameter $\gamma_s(G)$ in [7], and exact values of $\gamma_s(G)$ have also been established for various graph classes, such as paths, cycles, complete multipartite graphs and products of paths and cycles. Various properties of secure dominating sets of graphs have also been studied in [1, 2, 4, 5, 6].

Consider, as an example, the graph G_1 in Figure 1 for which $\gamma_s(G_1) = 2$. A minimum secure dominating set for G_1 is $\{v_1, v_2\}$; vertex v_4 is defended by v_2 while v_3 and v_5 are both defended by v_1.



G_1

Figure 1: *A minimum secure dominating set* $\{v_1, v_2\}$ *for a graph* G_1 *of order* 5.

The notion of secure graph domination finds applications in the generic setting where the vertex set of G represents a network of distributed locations in some spatial domain and the edges of G represent links between these locations. Patrolling guards, each stationed at one of these locations, may move along the links in order to protect the graph. A minimum secure dominating set of G then represents a smallest collection of locations at which guards may be stationed so that the entire location complex modelled by G is protected in the sense that if a security concern arises at location u, there will either be a guard stationed at that location who can deal with the problem, or else a guard dealing with the problem from an adjacent location v will still leave the location complex protected after moving from location v to location u in order to deal with the problem.

In this setting, the notion of edge removal is important, because one might seek the cost (in terms of the additional number of guards required to protect a location complex modelled by G) if a number of edges in G fail (*i.e.* a number of links are eliminated from the location complex, thereby disqualifying guards from moving along such disabled links).

A graph G is *q-critical* if the smallest arbitrary subset of edges whose removal from G necessarily increases the secure domination number, has cardinality q. Being able to determine the value of q for which a given graph is q-critical is important from an application point of view, because this value may be seen as a robustness threshold in the sense that the failure of some subsets of $q-1$ edges in G result in graphs that can still be dominated

securely by $\gamma_s(G)$ guards, but this is not true for the failure of q edges in G.

In this paper, we provide empirical evidence in support of a conjecture by Burger *et al.* [3] that the largest value of q for which there exists a graph of order n that is q-critical, is $\binom{n}{2} - 2n + 5$ for all $n \geq 7$. We also provide a repository of all q-critical graphs of order n and size m for all $q \in \{0, 1, \ldots, m\}$ and all $m \in \{1, 2, \ldots, \binom{n}{2}\}$, where $n \in \{2, \ldots, 6\}$.

2 The concept of q-criticality

We denote the set of all non-isomorphic graphs obtained by removing $q \in \{0, 1, \ldots, m\}$ edges from a given graph G of size m by $G - qe$. Furthermore, let $\gamma_s(G - qe)$ denote the set of values of $\gamma_s(H)$ as $H \in G - qe$ varies (for a fixed value of q). We distinguish between the graph obtained by removing a *specific* edge e from a given graph G, by writing $G - e$, and the class of graphs obtained by removing *any* single edge from G, by writing $G - 1e$.

The cost function

$$c_q(G) = \min \gamma_s(G - qe) - \gamma_s(G) \tag{1}$$

is nonnegative and bounded from above by q for all $q \in \{0, 1, \ldots, m\}$ [3]. This cost function measures the *smallest possible* increase in the secure domination number of a member of $G - qe$, relative to the secure domination number of a graph G of size m, when a set of $q \in \{0, 1, \ldots, m\}$ edges are removed from G.

A graph G is *q-critical* if $c_{q-1}(G) = 0$, but $c_q(G) > 0$ (that is, if the smallest arbitrary subset of edges whose removal from G necessarily increases the secure domination number, has cardinality q). The notion of q-criticality partitions the set of all non-isomorphic, nonempty graphs of order n in the sense that any such graph is q-critical for exactly one value of $q \in \{0, 1, 2, \ldots, \binom{n}{2}\}$, as demonstrated in the so-called edge-removal metagraph of the complete graph $\mathcal{K}_n$ of order 4 in Figure 2. The *edge-removal metagraph* of a graph G of size m is a graph whose nodes represent the non-isomorphic members of $G - qe$ for all $q = 0, 1, \ldots, m$. These nodes are arranged in $m+1$ levels, numbered $0, 1, \ldots, m$. The nodes on level q correspond to the members of $G - qe$. A node H on level $q-1$ of this metagraph is joined to a node H' on level q if H' can be obtained by removing one edge from H, for any $q \in \{1, 2, \ldots, m\}$. The only node on level 0 of the edge-removal metagraph of some graph G corresponds to G itself, while the only node on level m corresponds to the empty graph of the same order as G. The edge-removal metagraph of the complete graph $\mathcal{K}_n$ is of particular interest, because it contains nodes corresponding to all the non-isomorphic graphs of order n.

Let $\mathcal{Q}_n^q$ be the class of q-critical graphs of order $n \geq 2$ for some $q \in \{1, \ldots, \binom{n}{2}\}$. Grobler and Mynhardt characterised the graph class $\mathcal{Q}_n^1$ for all $n \in \mathbb{N}$ in 2009 [8, Theorem 2] and used their characterisation to derive a four-step construction process for computing all the members of the class $\mathcal{Q}_n^1$. Because of space constraints we do not give a full description of this (nontrivial) construction process here, but rather refer the reader to [8, Section 3.1] for the details. The following characterisation may be used to compute the class $\mathcal{Q}_n^q$ inductively from the class $\mathcal{Q}_n^{q-1}$ for any integer $n \geq 2$ and all permissible values of $q \geq 2$,

using the above-mentioned 4-step construction process by Mynhardt and Grobler [8] for the class $\mathcal{Q}_n^1$ as base case.

Proposition 1 ([3]) *A graph G of size at least $q > 1$ is q-critical if and only if*
(a) at least one graph $H^* \in G - 1e$ for which $\gamma_s(H^*) = \gamma_s(G)$ is $(q-1)$-critical, and
(b) each graph $H \in G - 1e$ for which $\gamma_s(H) = \gamma_s(G)$ is q^*-critical for some $q^* \leq q - 1$. ■

The inductive process referred to above is formalised in Algorithm 1. The algorithm commences by considering a graph $H \in \mathcal{Q}_n^{q-1}$ and proceeding to add a single edge $e \notin E(H)$ to H in Step 3, upon which the result of Proposition 1 is used to test whether or not $H + e \in \mathcal{Q}_n^q$. This process is repeated for each edge $e \notin E(H)$ and for each graph $H \in \mathcal{Q}_n^{q-1}$.

In Step 3 of Algorithm 1, another algorithm, Algorithm 2, is called to test whether $G = H + e \in \mathcal{Q}_n^q$. In Algorithm 2, each member of $G - 1e$ is examined. If a member $E \in G - 1e$

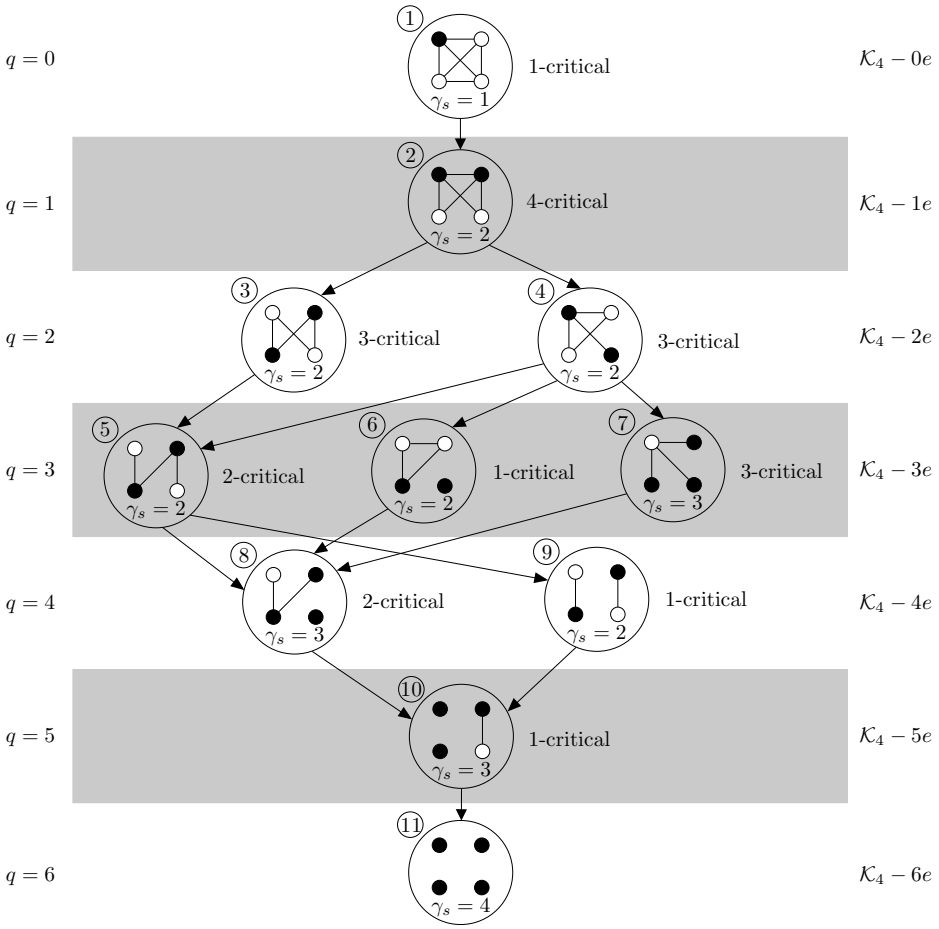


Figure 2: *The edge-removal metagraph of the complete graph $\mathcal{K}_4$ of order 4. The set $\mathcal{K}_4 - qe$ is shown on level q of the graph for all $q = 0, \ldots, 6$. Minimum secure dominating sets of the resulting graphs are denoted by solid vertices in each case. An arrow of the form $G \to H$ from level q to level $q + 1$ means that G is a certificate in $\mathcal{K}_n - qe$ showing that $H \in \mathcal{K}_n - (q+1)e$.*

Algorithm 1: Computing the class of $\mathcal{Q}_n^q$ of q-critical graphs of order n

Input : The graph classes $\mathcal{Q}_n^1, \ldots, \mathcal{Q}_n^{q-1}$.
Output : The class $\mathcal{Q}_n^q$ of q critical graphs of order n.

```
for each H ∈ Q_n^{q-1} do
    for each e ∉ E(H) do
        if q-Critical(H + e, q) then Q_n^q ← Q_n^q ∪ {H + e}
```

is found for which $\gamma_s(E) = \gamma_s(G)$, then $G \notin \mathcal{Q}_n^q$ by Proposition 1. Similarly, if a member $F \in \mathcal{Q}_n^p$ is found for some $p \geq q$, then $G \notin \mathcal{Q}_n^q$ by Proposition 1. If, however, no such graph F is found, then $G \in \mathcal{Q}_n^q$ by Proposition 1, since $H \in \mathcal{Q}_n^{q-1}$.

Algorithm 2: q-Critical(G, q)

Input : A graph G and the value of q.
Output : A boolean value stating whether G is q-critical.

```
if G ∈ Q_n^p for some p ≤ q − 1 then
    return [False]
for each e ∈ E(G) do
    if γ_s(G − e) = γ_s(G) and G − e ∉ Q_n^p for some p ≤ q − 1 then
        return [False]
return [True]
```

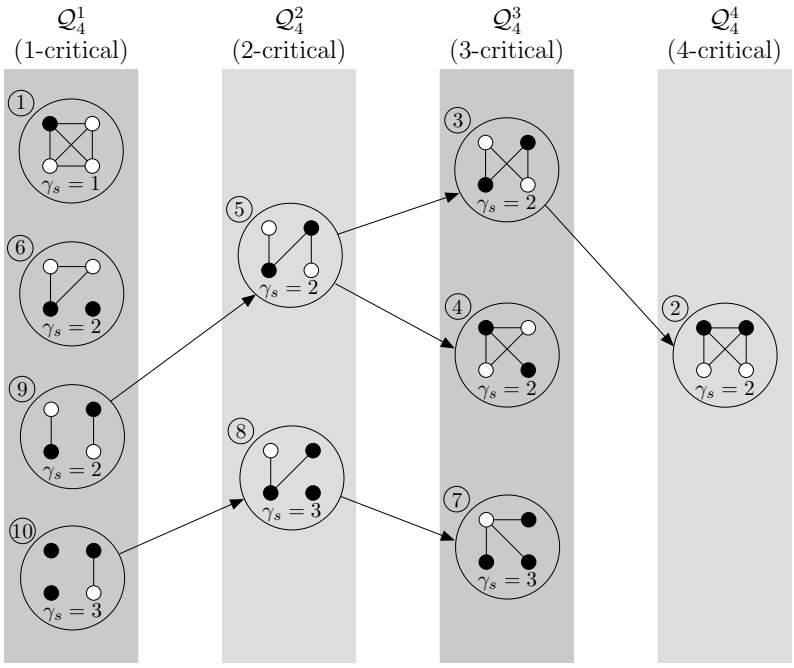


Figure 3: *The graph classes $\mathcal{Q}_4^1, \mathcal{Q}_4^2, \mathcal{Q}_4^3$ and $\mathcal{Q}_4^4$. Minimum secure dominating sets are denoted by solid vertices in each case. An arrow of the form $H^* \to G$ in the figure denotes the relationship between the graphs G and H^* in Proposition 1.*

The graph classes $\mathcal{Q}_4^1, \ldots, \mathcal{Q}_4^4$ are shown in Figure 3. The classes $\mathcal{Q}_4^5$ and $\mathcal{Q}_4^6$ are both

	Number of q-critical graphs of order n							
$n \to$	2	3	4	5	6	7	8	9
$\Omega_n \to$	1	2	4	6	9	12	17	23
$\lvert\mathcal{Q}_n^1\rvert$	1	2	4	7	14	26	52	104
$\lvert\mathcal{Q}_n^2\rvert$		1	2	6	18	50	141	394
$\lvert\mathcal{Q}_n^3\rvert$			3	9	32	111	428	1 514
$\lvert\mathcal{Q}_n^4\rvert$			1	8	34	165	910	4 424
$\lvert\mathcal{Q}_n^5\rvert$				2	28	199	1 484	10 587
$\lvert\mathcal{Q}_n^6\rvert$				1	18	195	1 875	20 144
$\lvert\mathcal{Q}_n^7\rvert$					8	153	2 010	30 849
$\lvert\mathcal{Q}_n^8\rvert$					2	93	1 847	38 831
$\lvert\mathcal{Q}_n^9\rvert$					1	37	1 520	41 620
$\lvert\mathcal{Q}_n^{10}\rvert$						10	1 088	38 341
$\lvert\mathcal{Q}_n^{11}\rvert$						3	627	30 962
$\lvert\mathcal{Q}_n^{12}\rvert$						1	260	22 864
$\lvert\mathcal{Q}_n^{13}\rvert$							76	15 934
$\lvert\mathcal{Q}_n^{14}\rvert$							19	10 053
$\lvert\mathcal{Q}_n^{15}\rvert$							5	5 222
$\lvert\mathcal{Q}_n^{16}\rvert$							2	2 048
$\lvert\mathcal{Q}_n^{17}\rvert$							1	585
$\lvert\mathcal{Q}_n^{18}\rvert$								138
$\lvert\mathcal{Q}_n^{19}\rvert$								34
$\lvert\mathcal{Q}_n^{20}\rvert$								11
$\lvert\mathcal{Q}_n^{21}\rvert$								5
$\lvert\mathcal{Q}_n^{22}\rvert$								2
$\lvert\mathcal{Q}_n^{23}\rvert$								1
Total	1	3	10	33	155	1 043	12 345	274 667
Time	$\ll 1$	$\ll 1$	< 1	2	23	531	27 208	1 069 220

Table 1: *Cardinalities of the graph classes $\mathcal{Q}_n^1, \ldots, \mathcal{Q}_n^{\Omega_n}$ for $n \in \{2, \ldots, 9\}$ as computed on a 3.4 GHz Intel(R) Core(TM) i7-3770 processor with 8 GiB RAM running in* Ubuntu 12.04 *and using a* C++ *implementation of Algorithms 1–2 in conjunction with the* Boost *graph library [11] for graph isomorphism testing. The computation times, shown in the last row, are measured in seconds and represent the time required for determining the graph class $\mathcal{Q}_n^q$ from the graph class $\mathcal{Q}_n^{q-1}$, for all $q \in \{2, \ldots, \Omega_n\}$.*

empty. The 4-step construction of Grobler and Mynhardt [8] was used to compute the class $\mathcal{Q}_n^1$ in the first column of the figure as base case. Thereafter, Algorithm 1 was used to compute the classes $\mathcal{Q}_4^2$, $\mathcal{Q}_4^3$ and $\mathcal{Q}_4^4$ inductively.

Note that it is, in view of Proposition 1 and Algorithms 1–2, not necessary to construct the entire edge-removal metagraph of the complete graph of order n in order to determine the graph class $\mathcal{Q}_n^q$ for a fixed value of q; instead only the classes $\mathcal{Q}_n^1, \ldots, \mathcal{Q}_n^q$ need be constructed inductively which, for values of q that are small compared to $\binom{n}{2}$, can be achieved in a fraction of the time required to construct the entire edge removal metagraph of $\mathcal{K}_n$.

3 Numerical results

Let Ω_n denote the largest value of q for which there exist q-critical graphs of order n. Values of Ω_n have been established for small n. In particular, Burger *et al.* [3] showed that $\Omega_2 = 1$, $\Omega_3 = 2$, $\Omega_4 = 4$, $\Omega_5 = 6$ and $\Omega_6 = 9$. They also conjectured as follows.

Conjecture 1 ([3]) $\Omega_n = \binom{n}{2} - 2n + 5$ *for all* $n \geq 7$.

In this paper we provide further circumstantial evidence in support of Conjecture 1, by proving the conjecture correct for $n \in \{7, 8, 9\}$. In particular, using a C++ implementation of Algorithms 1–2, we confirmed the values of Ω_n for $n \leq 6$ mentioned above, and additionally showed that $\Omega_7 = 12$, $\Omega_8 = 17$ and $\Omega_9 = 23$. The results thus obtained are summarized in Table 1, which contains listings of the cardinalities of the graph classes $\mathcal{Q}_n^q$ for $n \in \{2, \ldots, 9\}$ and $q \in \{1, \ldots, \Omega_n\}$. The classes $\mathcal{Q}_2^1, \ldots, \mathcal{Q}_9^1$ were determined by the 4-step construction process of Mynhardt and Grobler [8], referred to above.

A repository of the members of the graph classes $\mathcal{Q}_n^1, \ldots, \mathcal{Q}_n^{\Omega_n}$ is provided in Table 2 for $n \in \{2, 3, 4, 5, 6\}$. The graphs in this table are presented in the well-known *graph6* format [9], which is ideal for storing class representatives of isomorphism classes of undirected graphs in a compact manner, using only printable ASCII characters. These graphs may be converted to adjacency matrices and other formats using the reader showg, which is available online [9]. The reader showg package is part of nauty, originally designed by McKay and Piperno [10] for graph isomorphism testing.

4 Further work

In addition to attempting a general proof or refutation of Conjecture 1, another interesting problem for future research would be to investigate the *largest* number of arbitrary edges whose removal from a graph *necessarily does not increase* the secure domination number. In this context the cost function

$$C_p(G) = \max \gamma_s(G - pe) - \gamma_s(G)$$

is applicable instead of (1), and a graph G may be defined to be *p-stable* if $C_p(G) = 0$, but $C_{p+1}(G) > 0$. This problem finds application in cases where one seeks threshold information in terms of the largest set of edges whose removal from G does not increase the secure domination number of the resulting graph.

Acknowledgements

The research towards this paper was supported financially by the South African National Research Council under grant numbers 70593, 77248 and 81558.

Class	Class members											
$\mathcal{Q}_2^1$	A_											
$\mathcal{Q}_3^1$	Bw	B_										
$\mathcal{Q}_3^2$	Bg											
$\mathcal{Q}_4^1$	C~	CJ	CK	C@								
$\mathcal{Q}_4^2$	CL	CB										
$\mathcal{Q}_4^3$	CN	C]	CF									
$\mathcal{Q}_4^4$	C^											
$\mathcal{Q}_5^1$	D~{	DJ[	DBw	DJ_	D@K	D@O	D?C					
$\mathcal{Q}_5^2$	DB{	DFw	DJc	D@S	D@o	D?K						
$\mathcal{Q}_5^3$	DF{	DJk	DK{	DLs	D@[	DBW	D@s	DBg	D?[			
$\mathcal{Q}_5^4$	DJ{	DL{	DNw	DB[	D@{	DBk	DLo	D?{				
$\mathcal{Q}_5^5$	DN{	D]{										
$\mathcal{Q}_5^6$	D^{											
$\mathcal{Q}_6^1$	E~~w	EJ\w	EB^_	E?^o	EJ]?	E@ro	E@Kw	EJaG	E?Lo	E@L?	E?CW	E@Q?
	E?C_	E??G										
$\mathcal{Q}_6^2$	E?^w	E@^o	EJ]G	E?~o	E@rw	E@vo	EBjg	EJaW	E?Lw	E?\o	E?No	E?]o
	E@LG	E@N?	E@QG	E?Cg	E?D_	E??W						
$\mathcal{Q}_6^3$	E@^w	EB^g	EJ]W	E?~w	E@vw	E@~o	EBjw	EBng	EBzg	EBzo	EJfo	EJnO
	EJeW	EJeg	E?\w	E?Nw	E?]w	E@LW	E@Pw	E@Tg	E@NG	E@QW	E@Qo	E@U_
	EBY?	E?Cw	E?Ko	E?Dg	E?LO	E?F_	E?N?	E??w				
$\mathcal{Q}_6^4$	EB^w	EJ]w	E@~w	EBnw	EBzw	EB~o	EJno	EFzg	ELrw	EJew	EJbw	EJfg
	ELv_	E@Lw	E@Tw	E@\o	E@NW	E@UW	E@Qw	E@Ug	E@Uo	E@V_	E@^?	EBYG
	EB]?	EBj?	E?Kw	E?Dw	E?LW	E@T_	E?Fg	E?NG	E?NO	E?@w		
$\mathcal{Q}_6^5$	EJ^w	EB~w	EJ~o	EFzw	EJmw	EJfw	EJnW	EK~o	ELvg	E@\w	EBXw	E@Nw
	E@Uw	E@]o	E@Rw	E@Vg	E@^G	E@^O	EBYW	EB]G	EBYg	EB]_	E@v_	EBjG
	EBj_	E?Fw	E?NW	E?Bw								
$\mathcal{Q}_6^6$	EF~w	EJnw	EK~w	ELvw	El~o	ENzg	EB\w	E@]w	E@Vw	E@^W	EB]W	EBYw
	EB]g	E@vg	EBjW	EBn_	EBz_	EJf_						
$\mathcal{Q}_6^7$	EJ~w	EL~w	ENzw	E]~o	EB]w	EBZw	EBnW	EFz_				
$\mathcal{Q}_6^8$	EN~w	E]~w										
$\mathcal{Q}_6^9$	E^~w											

Table 2: *The graph classes $\mathcal{Q}_n^1, \ldots, \mathcal{Q}_n^{\Omega_n}$ for $n \in \{2, \ldots, 6\}$. Class members are presented in the well-known* graph6 *format, which is ideal for storing undirected graphs in a compact manner, using only printable ASCII characters. These graph representations may be converted to adjacency matrices (or other formats) using the reader* showg *which is available online [9].*

References

[1] Burger AP, Cockayne EJ, Gründlingh WR, Mynhardt CM, van Vuuren JH& Winterbach W, 2004, *Finite order domination in graphs*, Journal of Combinatorial Mathematics and Combinatorial Computing, **49**, 159–175.

[2] Burger AP, Cockayne EJ, Gründlingh WR, Mynhardt CM, van Vuuren JH& Winterbach W, 2004, *Infinite order domination in graphs*, Journal of Combinatorial Mathematics and Combinatorial Computing, **50**, 179–194.

[3] Burger AP, de Villiers AP& van Vuuren JH, 2014, *Edge criticality in secure graph domination*, Discrete Applied Mathematics, Submitted.

[4] Burger AP, de Villiers AP & van Vuuren JH, 2014, *On minimum secure dominating sets of graphs*, Quaestiones Mathematicae, Submitted.

[5] Burger AP, Henning MA & van Vuuren JH, 2008, *Vertex covers and secure domination in graphs*, Quaestiones Mathematicae, **31(2)**, 163–171.

[6] Cockayne EJ, Favaron O & Mynhardt CM, 2003, *Secure domination, weak roman domination and forbidden subgraphs*, Bulletin of the Institute of Combinatorics and its Applications, **39**, 87–100.

[7] Cockayne EJ, Grobler PJP, Gründlingh WR, Munganga J & van Vuuren JH, 2005, *Protection of a graph*, Utilitas Mathematica, **67**, 19–32.

[8] Grobler PJP & Mynhardt CM, 2009, *Secure domination critical graphs*, Discrete Mathematics, **309**, 5820–5827.

[9] McKay BD, 2014, *Graph6 and sparse6 graph formats*, [Online], [Cited: April 3rd, 2014], Available from: `http://cs.anu.edu.au/~bdm/data/formats.html`.

[10] McKay BD & Piperno A, 2013, *Practical graph isomorphism, II*, Journal of Symbolic Computation, **60**, pp. 94–112.

[11] Siek J, Lee LQ & Lumsdaine A, 2014, *Boost graph library*, [Online], [Cited: April 10th, 2014], Available from `http://www.boost.org/libs/graph/`.

Proceedings of the 2014 ORSSA Annual Conference
pp. 54–61
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Prerequisites for the design of a threat evaluation and weapon assignment system evaluator

ML Truter* JH van Vuuren†

Abstract

In a military air-defence environment, ground weapon systems are responsible for defending assets against hostile aerial threats. To be able to fulfil this purpose, the *weapon systems* (WSs) are equipped with an array of sensors capable of detecting these threats. In this context, the purpose of a *threat evaluation and weapon assignment* (TEWA) system is to provide decision support to human operators tasked with WS assignment decisions, enabling them to make optimal use of the WSs. Such a TEWA system typically assigns appropriate threat values to the threats and then uses these threat values to generate a recommended list of WS assignments in such a way that the cumulative survival probability of the aerial threats is minimised. A large number of TEWA systems are already in use around the world, but due to the confidential nature of this research area, the workings of these systems are typically not available in the open literature. Despite the critical role of these systems in combat situations, there exist no standard methods in the open literature to evaluate the performance of TEWA systems. After briefly describing the subsystems of a TEWA system, various factors that potentially influence the effectiveness of a TEWA system are highlighted, and a methodology is proposed for evaluating the performance of a TEWA system within an integrated simulation modelling paradigm.

Key words: Performance Evaluation, Threat Evaluation and Weapon Assignment, Decision Support, Air-Defence, Simulation.

1 Introduction

On 22 May 2003, a *Royal Air Force* (RAF) Tornado jet was returning to its base when a *United States* (US) Patriot missile wrongly identified the fighter plane as an Iraqi anti-radiation missile. The blue-on-blue[1] confrontation which followed resulted in the US Patriot missile destroying the RAF Tornado [3].

During the build-up to this event, the Patriot battery crew were monitoring for Iraqi ballistic missiles when the Tornado plane was identified by their system. The symbol

*Department of Industrial Engineering, University of Stellenbosch, Private bag X1, Matieland, 7602, Republic of South Africa, email: 16057694@sun.ac.za

†**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

[1]Synonymous with friendly fire — an inadvertent firing toward one's own or otherwise friendly forces.

which appeared on the radar corresponded to that of an anti-radiation missile. To confirm the Patriot system's threat evaluation result, the radar track was interrogated for IFF[2], but no reply was received. After meeting all the criteria laid out by the specific rules of engagement, the Patriot crew engaged in self-defence action by destroying a friendly aircraft.

Several possible causes of this accident have subsequently been identified. The investigation board concluded that the rules of engagement were not robust enough to prevent blue-on-blue confrontations. In addition, the Patriot crews were trained to identify and react quickly, engage early and to trust the Patriot system. The problem with the training was that it focused on identifying generic threats rather than those specific to the Iraq conflict, and also not on identifying false alarms. After detailed investigations, the board concluded that both the operator training and the Patriot firing doctrine were factors contributing to the accident [3].

In this paper we advocate that perhaps the most effective way of preventing the reoccurrence of such a *threat evaluation and weapon assignment* (TEWA) decision support malfunction, is to develop a simulation model for testing the performance of the system with respect to different scenarios, thereby identifying potential system errors before commissioning the system. In the above accident, a scenario generation approach could have identified the flaws in the Patriot system's threat evaluation algorithms, making it possible to take corrective action. Since lives are at stake, it is clearly of critical importance that the performance of a TEWA system be thoroughly evaluated before commencing its industrialization phase.

This paper is structured as follows. A brief overview of the working of a TEWA system is given in §2, with a focus on the three core elements of the system: threat evaluation, weapon assignment and decision support. Thereafter, the importance of, and difficulties associated with, the performance evaluation of TEWA systems are elucidated in §3. Four TEWA performance metrics are proposed in §4. The paper concludes with some suggestions for possible further work related to the performance evaluation of TEWA systems. The research detailed in this paper is a report on work in progress. The methodology proposed forms part of an ongoing research project in which instances of a TEWA system will be simulated and evaluated, resulting in the identification of possible limitations and conflicts present in the algorithms employed by the system.

2 Current State of GBAD TEWA Systems

Several TEWA *decision support systems* (DSSs) are in use around the world, but the inner workings of these systems are typically classified [11]. The majority of TEWA systems are used on naval craft in a point-defence[3] role, while TEWA in a *ground-based air defence* (GBAD) environment requires the adoption of an area-defence[4] paradigm due

[2]Identification: Friend or Foe.

[3]Point-defence applied to the defence of a single entity; stationary or moving.

[4]Area-defence entails the protection of an area, possibly containing numerous DAs spread across the area.

to the potential spatial distribution of a number of prioritized *defended assets* (DAs) on the ground [11].

The *weapon assignment* (WA), *threat evaluation* (TE), and *human machine interface* (HMI) subsystems are three key constituents of any TEWA DSS. In order to simulate a complex TEWA system in detail, a model replicating all the formative elements (enemy aircraft, WSs, sensors, DAs and decision support modules) is required [10]. To further complicate matters, the environment in which a TEWA DSS operates, typically includes continual, dynamic and several non-linear interactions between the formative elements (feedback, looping and sudden changes are common). Furthermore, all these elements are influenced by one another, giving rise to various emergent properties of the system[5] and thus making it infeasible to test a TEWA DSS in a reductionistic manner.

2.1 The Threat Evaluation Subsystem

The TE process requires input data from ground radars and associated sensors. These sensors are responsible for detecting, tracking and identifying potential threatening aerial vehicles [4]. The TE subsystem utilizes the kinematic, tracking and attribute data collected from the sensors to estimate the level of threat posed by each aerial vehicle. During this process, a threat value is assigned to each threat-DA pair. Different measured attributes are taken into consideration when determining these threat values. These attributes can be subdivided into three classes:

Proximity parameters quantify the distance between a threat and a DA. Hence, a threat that is far away from a DA will not be classified as an imminent threat to that DA, when compared to threats that are close to the DA. A widely used example of such a parameter, is the range to the *closest point of approach* of a threat with respect to a DA [10].

Capability parameters attempt to quantify a threat's ability to cause damage to a DA. To calculate this value, it is required to know specific characteristics of the attacking aircraft. Examples of capability parameters include the threat type, its weapon envelope and its fuel capacity.

Intent parameters aim to quantify the will and determination of a threat to cause damage to a DA. Of these three parameter classes, intent is the most difficult type of parameter to estimate, but certain measured threat attributes can be used to estimate a threat's intent [11]. One method in which intent is estimated is through recognition of known attack manoeuvres from an aircraft's measured track.

A number of different algorithms of varying complexity and sophistication typically run concurrently in the TE subsystem, each assigning threat values to each threat-DA pair. This results in several threat values for each threat-DA pair. In the case where certain necessary sensor data are not available, the TE system will select scaled-down TE models

[5] *Emergent properties* are defined as those properties that derive from the interaction of the elements in the system, but cannot be reduced to them [2].

which are able to estimate threat values in the absence of very detailed threat data [8]. The different threat values for each threat-DA are then fused together to obtain a single prioritised list of threat values (*i.e.* a threat value for each threat-DA pair, typically found on a consensus basis, taking into account the results contributed by all the TE models) [8]. These threat values are used by the WA subsystem in a bid to optimise the utilization of available resources (WSs and ammunition) when weapon assignment decisions are made for engaging the aerial threats.

2.2 The Weapon Assignment Subsystem

Weapon assignment is the process of reactive allocation of weapon resources (ammunition and WSs) to counter identified threats [5]. The WA subsystem of a TEWA DSS is responsible for proposing high-quality assignment proposals of available ground-based WSs to engage aerial threats over some specified time frame [7].

Before high-quality assignments can be proposed by the WA subsystem, an operator is required to select a set of objectives that have to be optimised during weapon assignment. Lötter and Van Vuuren [7] have suggested a suite of objective functions that may be used in WA algorithms, including the minimisation of threat survivability, the minimisation of overall assignment cost and the maximisation of re-engagement opportunities (available ammunition after the engagement). Because of the typically short time frame over which decisions have to be made, the solutions generated by the WA subsystem are not always optimal; sometimes locally optimal solutions are generated. For this reason, meta-heuristic optimisation techniques are usually preferred over exact ones in a TEWA DSS [7].

The WA process usually employs different algorithms running concurrently, because certain algorithms may perform better than others under certain conditions [8]. The resulting algorithmic outputs are then fused together to obtain a single (approximately) Pareto-optimal solution front in the combined objective space. The operator can then select a specific solution from this front, depending on the situation and his/her preferences. The *fire control officer* (FCO) is presented with the Pareto-optimal solutions through an HMI, and can use this interface to interact with the WA suggestions in order to gain more clarity on the reasoning behind the WA decision suggested by the TEWA system.

2.3 Communicating Decision Support to the Operator

It is ultimately a human operator who decides whether and how each aerial threat should be engaged — not a fully automated system — because decision making in a GBAD environment can have severe (possibly catastrophic) consequences if inappropriate decisions are made, as described in the introduction. Hence, it is of utmost importance to ensure that the decision support information communicated to the human operator is as clear and uncluttered as possible. By so doing, the human operator is afforded the opportunity of effectively making use of the data for the purposes of analysis, interpretation and decision-making.

The form of decision making explained above is a highly complex task and requires the integration of various data sources [6]. To succeed in this highly stressful and dynamic decision making environment, it is required that the FCO should possess a high level

of tactical expertise and knowledge of the type of threats, prevailing legal frameworks and assessment heuristics from experience [1]. Training and experience are, however, not enough to ensure tacit decision making. According to Morrison *et al.* [9], the importance of ensuring that information is meaningful, timely and easily accessible cannot be underestimated.

3 Performance Evaluation of TEWA Systems

Testing and evaluation is an iterative process of performance measurement, correction of deficiencies and remeasuring of the resulting performance. This testing process should commence as early as possible in the design process and should be conducted throughout system development [2]. The main purpose of testing a TEWA system is to identify general design deficiencies and specific conflicts present in the internal algorithms of the system, thereby highlighting required corrective action. By following this bottom-up testing approach, it is possible to reduce the risks associated with the final commissioned system. According to Sparrius [13], the only way to demonstrate risk reduction, as a prerequisite for an increase in resource commitment, is through testing and evaluation.

As is the case with many modern systems, the evaluation and design of a TEWA system is highly complex because of the magnitude of the system and the complexities of all the subsystems involved. A TEWA system's performance depends sensitively on the synergies between its subsystems, which gives rise to the emergent properties of the system (see §2). These emergent properties cannot be accounted for by individually testing the WA and TE subsystems, because such an approach will not enable one to assess whether the entire TEWA system will function as expected. To effectively test such a TEWA system, a *system-of-systems* (SoS) approach is required. Proper SoS engineering entails the allocation of functionality to components as well as inter-component interactions, rather than only focussing on the workings of individual components. SoS engineering is very powerful in terms of exploiting synergies between subsystems and in identifying capabilities that no standalone system testing can provide [12].

The preferred method of ensuring that a TEWA system is fully functional is through conducting full-scale flight tests. However, the complex nature of a TEWA system and the high cost of such an approach, makes it intractable to run flight tests for the purposes of system evaluation. In addition, flight tests alone do not provide insight into scenarios that were not actually tested. Because of the confidential nature of this research area, historical data on flight tests, from which system designers can gain insight into the performance characteristics of existing TEWA systems, are very rare. System designers are therefore forced to utilise modelling and simulation tools to evaluate the performance of TEWA systems.

4 Possible Performance Evaluation Metrics

In this section we propose the use of four performance metrics when evaluating the performance of a TEWA DSS within a simulation modelling paradigm. These metrics may

serve as both *absolute* and *comparative* evaluation measures in the sense that the value of a metric may quantify the suitability of assignments proposed by a TEWA DSS in a specific scenario in absolute terms, but may also be used to identify limitations present in its constituent algorithms by comparing the metric values for different scenarios in a relative manner. By evaluating the metric values of different scenarios, it should be possible to determine the conditions under which the algorithms behave poorly. Finally, these metrics can be used to perform a sensitivity analysis with respect to the implemented algorithms, thereby providing valuable insight into the functioning and limitations of the TEWA system as a whole.

There is often a misconception about the term *performance metric*. A metric is a standard definition of any measurable quantity, while a performance metric goes further by gauging some aspect of a system's performance. For a performance metric to be successfully utilized, it must adhere to certain requirements — evaluating a performance metric should be achievable in a reliable, repeatable and consistent manner, independently of the pressure to drive performance.

The main role of the WSs in an air-defence scenario is to protect the DAs. Therefore, *survivability* is an important criterion for measuring the performance of a TEWA system. Johansson [5] suggested the use of the survivability metric

$$S = \frac{\sum_{j=1}^{D} \omega_j u_j}{\sum_{j=1}^{D} \omega_j}, \tag{1}$$

where D is the number of DAs in a simulation performance evaluation environment, ω_j is the importance value associated with DA j, and u_j is a binary variable which assumes the value 1 if DA j survives, and the value 0 if it is destroyed by an aerial threat. Hence, the survivability S is the ratio between the protection value of surviving assets to the total protection value of all the assets.

The metric in (1) does not penalise the engagement of superfluous aircraft as unnecessary engagements. The introduction of an *economy* metric may, however, account for the cost of engagement by each WS — thereby penalising unnecessary engagements. The economy metric

$$M = \sum_{i=1}^{W} \left(C_i \sum_{j=1}^{T} x_{ij} \right) \tag{2}$$

is proposed for this purpose, where C_i denotes the cost of one burst or round of ammunition for WS i, x_{ij} is the number of times WS i engages threat j, and W and T are respectively the number of WSs and threats. Hence, the economy metric M represents the total WS capital expenditure, based on ammunition used, associated with an engagement strategy.

It is preferable to destroy high-value aerial threats, especially in an ongoing conflict. In this context the value of an aerial threat may be interpreted as its ability to cause considerable damage to important classes of DAs. Our next metric, *engagement effectiveness*, is designed to reward the successful engagement of high-value threats. The value of a specific aerial threat may be determined during the pre-deployment phase of a mission and programmed into the TE subsystem. As stated above, it is desirable from an economic point of view not to engage superfluous targets. During an ongoing conflict, however, it

may be beneficial to destroy these high-value threats even if they do not pose an imminent danger, in a bid to ensure that these threats do not return to attack DAs in the future.

Furthermore, a critical performance-related problem is the engagement of friendly and/or civilian aircraft (see §1). The engagement effectiveness metric may also be used to penalise friendly engagements by assigning a large negative importance value to friendly and commercial aircraft. In this way, friendly engagements can be heavily penalised. The engagement effectiveness metric is given by

$$E = \frac{\sum_{j=1}^{T} \nu_j e_j}{\sum_{j=1}^{T} \nu_j}, \tag{3}$$

where ν_j denotes the importance value associated with destroying threat j and e_j is a binary variable assuming the value 1 if threat j is destroyed, or the value 0 if the threat survives. The value of ν_j may be interpreted as the perceived value that the enemy is most likely to assign to an aircraft — the more important the aircraft, the higher the value. The engagement effectiveness E is the ratio of the importance value of destroyed threats to the total importance value associated with all threats encountered throughout the engagement.

Our final metric attempts to quantify the *adaptability* of a specific engagement strategy. This metric is given by

$$A = \min_i \left\{ A_i - \sum_{j=1}^{T} x_{ij} \right\}, \tag{4}$$

where A_i denotes the initial amount of ammunition available to WS i and the rest of the parameters have the same meanings as before.

The metric A is designed to measure the propensity of an engagement strategy to maximize the number of times that a WS is available for re-engagement after the proposed assignment, thereby ensuring that as many WSs as possible are reusable in future engagements. By using ammunition more effectively during an engagement, WSs on the ground will be more adaptable to changing conditions, such as responding to newly detected threats and performing follow-up engagements.

In addition to the above metrics, certain other parameters also need to be considered to ensure a successful TEWA system. These include the time required to generate WA allocation suggestions and the memory storage requirements of the internal algorithms of the system. Indeed, it is important to ensure that the FCO has enough time to utilize the results generated by the WA subsystem. Also, depending on the environment in which the TEWA system is implemented, there might be memory restrictions. There is often a trade-off between the time complexity and memory complexity of an internal TEWA algorithm — increased memory consumption normally leads to faster execution times, and *vice versa*.

5 Further Work

As stated in §1, the work detailed in this paper is a report on work in progress. The performance prerequisites detailed here will be used during the planned evaluation of a

developed TEWA DSS as part of a larger, ongoing research project. Numerous scenarios will be generated by using commercially available simulation software, in order to test the performance of the system's internal algorithms under a variety of conditions. The metric values (1)–(4) will be calculated for each of these engagements in a bid to quantify the relative and absolute performance of the algorithms in each scenario. The results will make it possible to gain insight into the operation of the TE and WA processes and make it possible to identify limitations and internal conflicts, and to clarify possible improvements to the system.

References

[1] Cohen MS, Freeman JT & Thompson BB, 1997, *Integrated critical thinking training and decision support for tactical anti-air warfare*, Proceedings of the 1997 Command and Control research technology symposium, Arlington (VA).

[2] Danzig B, 2010, *Systems engineering handbook: A guide for system life cycle processes and activities*, 3rd Edition, International Council on Systems Engineering (INCOSE), San Diego (CA).

[3] Directorate of Air Staff, 2004, *Aircraft accident to royal air force tornado GR MK4A ZG710*, (Unpublished), Technical Report, Ministry of Defence, London.

[4] Heyns AM, 2008, *Measuring the threat value of fixed wing aircraft in a ground-based air defence environment*, MSc Thesis, Stellenbosch University, Stellenbosch.

[5] Johansson F, 2010, *Evaluating the performance of TEWA systems*, PhD Dissertation, Orebro University, Kållered.

[6] Liebhaber MJ & Feher B, 2002, *Air threat assessment: Research, model and display guidelines*, (Unpublished), Technical Report, DTIC Document.

[7] Lötter DP, Nieuwoudt I & van Vuuren JH, 2013, *A multiobjective approach towards weapon assignment in a ground-based air defence environment*, ORiON, **29(1)**, pp. 31–54.

[8] Lötter DP & van Vuuren JH, 2013, *Weapon assignment decision support in a surface-based air defence environment*, Military Operations Research, Submitted.

[9] Morrison JG, Kelly RT, Moore RA & Hutchins SG, 2012, *Tactical Decision Making Under Stress (TADMUS) Decision support system*, Proceedings of the 1997 IRIS National Symposium on Sensor and Data Fusion, MIT Lincoln Laboratory, Lexington (MI), p. 17.

[10] Roux JN & van Vuuren JH, 2008, *Real-time threat evaluation in a ground-based air defence environment*, ORiON, **24(1)**, pp. 75–101.

[11] Roux JN & van Vuuren JH, 2007, *Threat evaluation and weapon assignment decision support: A review of the state of the art*, ORiON, **23(2)**, pp. 151–187.

[12] Sommerer S, Guevara MD, Landis MA, Rizzuto JM, Sheppard JM & Grant CJ, 2012, *System-of-systems engineering in air and missile defence*, Johns Hopkins APL Technical Digest, **31(1)**, pp. 5–20.

[13] Sparrius A, 2014, Private Owner: Ad Sparrius System Engineering and Management (Pty) Ltd., [Personal Communication], Contactable at `ad_sparr@iafrica.com`.

Proceedings of the 2014 ORSSA Annual Conference
pp. 62–70
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Prerequisites for the design of an agent-based model for simulating the population dynamics of *Eldana saccharina* Walker

BJ van Vuuren*, L Potgieter† & JH van Vuuren‡

Abstract

Although South Africa boasts one of the most distinguished sugar producing industries in the world, the commercial success of the *South African Sugar Association* (SASA) continues to suffer as a result of the damage caused by a variety of pest species such as *Eldana saccharina* Walker (Lepidoptera: Pyralidae). This stalk borer pest feeds on the internal tissue of the sugarcane stalks, causing yield losses in sucrose. Owing to the fact that *E. saccharina* has specific preferences in terms of egg-laying sites, it has been suggested that suitably heterogeneous crop layouts in sugarcane may be an effective means to localize infestation effects and contribute towards pest suppression. In order to determine good sugarcane field layouts to aid in pest suppression in this manner, it is proposed that a simulation model of *E. saccharina* spatial behavioural patterns be developed and tested for differing ages and varieties of sugarcane. To facilitate the design of such an agent-based simulation model which simulates *E. saccharina* biology, various characteristics of the pest, such as its feeding habits, mating behaviour and dispersal patterns, must be incorporated into the model framework. These characteristics, as well as their impact and incorporation in the aforementioned model, are discussed in this paper. Furthermore, a suitable simulation modelling framework is suggested based on these considerations.

Key words: Sugarcane pest infestation, Eldana saccharina Walker, agent-based simulation.

1 Introduction

Sugarcane cultivation in South Africa dates as far back as the early 1600s. In 1635, Portuguese explorers shipwrecked near the mouth of the Umzimkulu River, where they discovered that sugarcane was one of the crops grown by local inhabitants [23]. The sugar industry expanded following the establishment of the agricultural society in 1848 and,

*Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 16057651@sun.ac.za

†Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: lpotgieter@sun.ac.za

‡**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

today, about 428 000 hectares of cane produce, on average, 2.5 million tons of sugar per year [22].

The stalk borer, *Eldana saccharina* Walker, was first recorded in 1939 as a pest infiltrating sugarcane when a severe infestation occurred in the Umfolozi area. This infestation, however, did not spread and eventually died out in 1950 [13]. It then resurfaced in the Hluhluwe area during the early 1970s [6] and has since spread to much of the local sugarcane industry, but for inland crop areas where the insect is limited by the cooler temperature [1]. Evidence suggests that *E. saccharina* infests sugarcane by virtue of its suitability for egg-laying in dead leaf material [1, 20]. This infestation results in decreased cane quality (measured as a decrease in sucrose yield) and has a negative impact on total plant biomass [11]. An early study showed that, on average, mature infested cane exhibits a percentage loss which is comparable to that of the percentage of internodes damaged[1] [17].

In light of this, *E. saccharina* remains a topical concern in the sugar industry and the *South African Sugar Research Industry* (SASRI) continues to spearhead research efforts towards finding a means of effectively controlling the pest. A number of control measures currently exist, achieving only limited success. These include cane variety development [10, 16], chemical control [13], biological control, habitat management [8, 9, 10] and the *sterile insect technique* (SIT) [18].

The most limiting factors hindering roll-out of more recent methods, such as biological control, habitat management and SIT, include economic constraints on continuous, incremental development of optimal implementation techniques, as well as the lack of a means for practical evaluation of their relative impact when applied to infested sugarcane [9].

The purpose of this paper is to identify a number of prerequisites for the design of an agent-based simulation model which can be used to assess *E. saccharina* control methods before costly in-field testing is conducted. This will aid in acquiring a measure of insight into their predicted effectiveness and optimal implementation of the techniques in a time-efficient, economically viable manner.

This paper is a report on a work in progress within a larger study currently in progress at Stellenbosch University. It is the intention of the study to incrementally progress towards the development of a fully fledged agent-based simulation of *E. saccharina* by laying solid model foundations which will allow future researchers to periodically add elements discussed in this paper with a view to increase the model realism and complexity.

Various possibilities for such a simulation model are outlined in this paper, which is organised as follows. A brief description of the development of an *integrated pest management* (IPM) system approach is provided in §2, whereafter, a review of existing behavioural models and their shortcomings is presented is §3. This is followed by a discussion of the general characteristics of an agent-based simulation model and relevant biological attributes of *E. saccharina* in §5. Thereafter, a model framework is proposed in §6, based on the considerations described in §5. The paper then closes with a conclusion and discussion on the possible future development of this topic in §7.

[1] An internode is the softer part of the cane located between two adjacent growth rings.

2 Integrated pest management system approach

IPM systems aim to combine a series of control methods in order to achieve better overall pest management whilst decreasing the use of pesticides, thereby decreasing associated environmental problems. As a result, IPM is considered more sustainable in the long run [18]. Ideally, a number of the techniques mentioned in §1 should be applied in conjunction with one another to a sugarcane crop in order to achieve improved control of *E. saccharina*.

When applied together, research shows that the existing *E. saccharina* control methods can endorse one another [10]. In view of this knowledge, it is important to identify the shortcomings in the individual control methods, as well as the role of possible interactions between them. Determining these intricacies on either individual or interacting control methods using in-field testing is both costly and labour-intensive, and there is not always sufficient time available to conduct large-scale in-field tests in order to observe these effects. For this reason, simulation is becoming an increasingly attractive alternative to assist in the design and development of integrated control techniques [9].

3 Existing *E. saccharina* simulation models

A number of working simulations have been developed in the literature for evaluating the estimated effect of particular pest control methods imposed on sugarcane fields.

The first work of this nature was performed by Van Coller [24] and Hearne *et al.* [12] who employed a system of differential equations to model the change in population growth in the various stages of the life cycle of *E. saccharina*. This model primarily provided insight into the biological control of the pest through parasitoids, but did not explicitly take into account the spatial spread of an *E. saccharina* population. Specifically, the model enabled policies for the timing, frequency and magnitude of parasitoid releases to be tested for their relative effectiveness in the biological control of *E. saccharina*.

Horton *et al.* [14] later designed a model which was aimed at investigating the effects of insecticides and early cutting as control measures for *E. saccharina*. This model also explicitly included the effects of temperature on the pest's life cycle, but again assumed a homogeneous spread of the population over the spatial habitat. The function of this model was to produce a damage index which measures the extent of cane damage under different temperature patterns to assist in determining the time for harvesting.

Most recently, Potgieter *et al.* [19, 20] developed a discretised reaction-diffusion model of the growth and spread of *E. saccharina* population over time and space. An accompanying SIT simulation tool was also developed whereby the effectiveness of SIT could be investigated in different scenarios. Notably, these tools could be applied in the context of heterogeneous spatial domains — including realistic sugarcane field layouts — and, together, form a framework which can be extended for use in an IPM programme.

Despite the advancement in understanding of *E. saccharina* population growth and relative success achieved by the aforementioned simulation models, each model is founded upon approximations of the pest on a population level. Local interactions of individual moths are not simulated and incorporated explicitly into the population dynamics; in-

stead, approximate changes are executed in the simulations at each discrete time step. Furthermore, the models focus on single control measures, limiting their development and flexibility in the context of IPM systems. The resulting analyses therefore yield conclusions that do not necessarily reflect the continuous, changing nature of *E. saccharina* on a localized level.

4 Benefits of agent-based simulation

In an attempt to address the shortcomings of the existing simulation models described above, an approach is advocated in which the individual members of a population of *E. saccharina* are simulated, thereby incorporating the effects of local interactions between individual stalk borers.

Agent-based modelling is the computational study of social agents interacting in an autonomous manner as evolving systems. It allows for the study of complex adaptive systems and facilitates investigations into how macro-phenomena develop from micro-level behaviour among heterogeneous sets of interacting agents [15]. By simulating *E. saccharina* moths as individual agents who are governed by their biological preferences and limitations, the resulting relationships between these agents can be used to predict population dynamics of the pest in a more realistic manner over space and time, based on local interactions, than is possible in the simulation models described in §3.

It is anticipated that an autonomous, agent-based simulation model of *E. saccharina* may serve two main functions. Firstly, the model can be used to validate and support existing high-level averaging models of the stalk borer, such as those discussed in §3. This may assist in developing more advanced, accurate mathematical models for further understanding and prediction of the spreading of the pest. Secondly, a simple working model of *E. saccharina* can provide a platform facilitating the testing and evaluation of various control mechanisms, either alone or in combination, before expensive in-field testing is pursued.

Moreover, a computerised agent-based simulation model can be equipped with a visually intuitive interface facilitating the interpretation of model results by parties who do not have the necessary training required to understand complex mathematical models. Such a simulation model of *E. saccharina* can incorporate easily configurable variable factors affecting the spatial spread and temporal growth of the pest population. Users of the model can then investigate the effects of minor changes to the system and use these observed effects to predict the potential success of *E. saccharina* control measures in practice. This directly addresses a current shortcoming in the sugarcane industry where confidence levels cannot easily be associated with the expected outcomes of implementing proposed *E. saccharina* control methods, based on the predictions of existing simulation models. Often, further experiments and methodology have to be developed simply to verify whether a mathematical model or control method will operate as expected. A realistic agent-based model can assist in providing a simple, quick verification mechanism which, in turn, may increase confidence levels and creativity in the design of new *E. saccharina* control methods and IPM systems alike.

5 General considerations of agent-based simulation

In an agent-based model, each agent is initialised using a series of parameters, variables and functions [26]. Parameters keep a fixed record of facets of each agent which control their appearance and activity throughout the simulation. Variables indicate aspects of the agent's behaviour which are incorporated into its decision making process and can change during the course of a simulation. These decisions are made according to functions of both the parameters and variables governing agent behaviour. Simulations can also accommodate so-called *events* such as removal of an agent from the simulation or the introduction of a new agent to the simulation. These events are based on several requirements which, if satisfied, execute the event.

The proposed simulation model of *E. saccharina* aims to link the biological aspects and preferences of the stalk borer to functions and events which govern each individual agent in the simulation. The primary aspects of the simulation and the corresponding biological characteristics which must be incorporated are discussed below.

5.1 Different types of agents

Primarily, the simulation should incorporate both male and female *E. saccharina* moths as separate populations. The members of these populations should conform to typical behaviour of the stalk borer, unless otherwise required. Furthermore, all moths may initially be deemed fertile, unless the effects of SIT are incorporated into the simulation, in which case, a different type of agent may be introduced.

5.2 The introduction of new agents

In order to simulate the stochastic nature of an *E. saccharina* population size, the simulation requires the incorporation of an event whereby a new agent can be introduced into the simulation environment. Such introductions should result from mating of adult *E. saccharina* moths.

The life-cycle of *E. saccharina* is typical of that of insects. Eggs hatch into larvae, after which pupation occurs and, finally, adult moths emerge [2]. Carnegie [6] and Way [25] investigated the duration of each stage of the life cycle and the factors which affect moth development. These findings should be incorporated into the simulation to allow for the realistic introduction of new agents into the model and realistic timings between mating and moth emergence.

5.3 The removal of existing agents

Since the simulation will be of living organisms, an event must also be incorporated whereby the agents perish and are removed from the simulation. Mortality of the moth occurs haphazardly during its life cycle and natural enemies are not considered a notable factor in the mortality of *E. saccharina* [18]. Realistic mortality rates, as proposed by Van Coller [24], should be consulted in terms of natural removal of agents from the simulation.

5.4 The spatial movement of agents

Although *E. saccharina* is considered a relatively weak flier, Atkinson [3] describes instances where females may travel for mating purposes. Furthermore, Carnegie [6] notes cases where mated females migrate in search of oviposition sites. *E. saccharina* may also migrate in search of more mature sugarcane which is its preferred habitat [1]. Studies related to other moth species indicate that mobile and sedentary genotypes exist in populations that display dispersal capacity in the field [21]. The observations of Atkinson and Carnegie [5] correlate with these studies. These kinds of spatial movements should be included in the simulation.

5.5 The interaction between agents

The *lek mating* process followed by *E. saccharina* is described by Atkinson [3] and should be incorporated into the simulation preceding the introduction of a new agent, as described in §5.2. The frequency of mating between agents, as well as the number of times they mate, should adhere to the observations of Carnegie [6].

5.6 External influences on the agents

Other characteristics of the environment which affect *E. saccharina* behaviour and survival should also be incorporated into the model.

The effect of temperature on mortality and maturation rates of *E. saccharina* should, for example, be incorporated into the simulation, perhaps employing the polynomial fit to corresponding stage mortality and maturation data proposed by Potgieter [18]. These effects should give rise to seasonal cycles which affect the dominant life stages of *E. saccharina* significantly. The annual May milling season should also be captured in the simulation as it is considered the largest eradicator of the pest [4].

6 Suggested simulation modelling framework

We propose the basic model structure depicted in Figure 1 for the order and flow of events in an agent-based simulation of *E. saccharina*, taking into account the referenced biological considerations. It is suggested that each agent in the simulation follow this basic life cycle which should be programmed within a stochastic framework according to the biological factors discussed in §5.

In order to achieve effective visualisation of the population dynamics and behaviour of *E. saccharina* in the simulation model, it is proposed that male and female agents should be denoted by different colours within a graphical display of the sugarcane field layout. Furthermore, each stage of the life cycle should also be displayed in a different shade of the above-mentioned colours in order to indicate the number of eggs, larvae and pupating moths of each sex which exist within the population.

It is suggested that the time step adopted in the simulation model should be measured in days in order to be able to gauge individual activities of the moths, while simultaneously

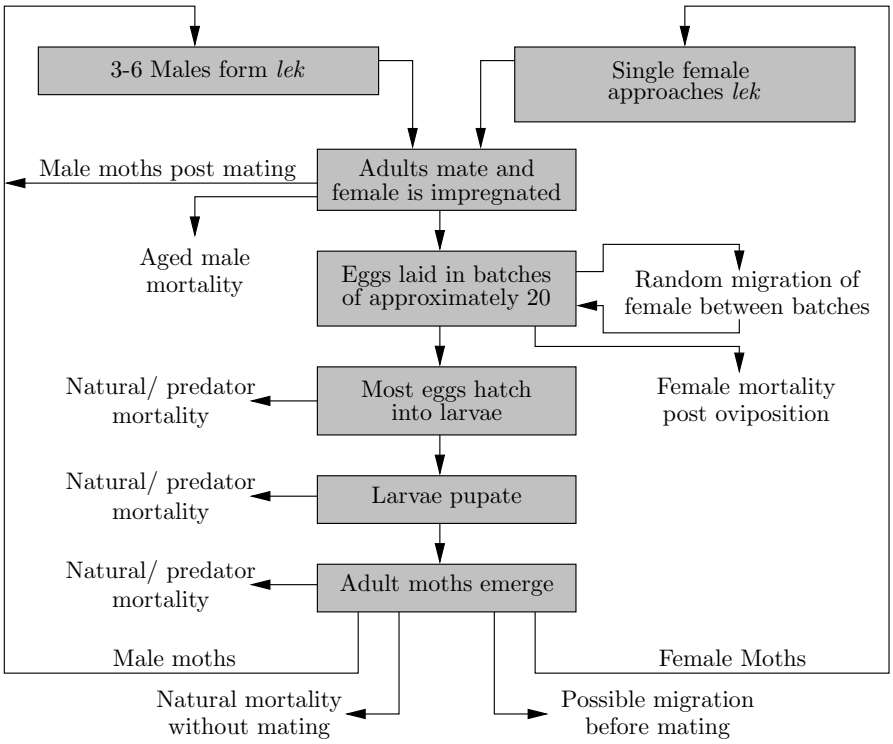


Figure 1: *Flow diagram of the life cycle followed by individual agents in the proposed agent-based simulation model of the growth and spread of an E. saccharina population.*

facilitating the ability to perform an assessment of the long-term development of the population. The model should, thus, possess the ability to be sped up and slowed down for the purposes of observation.

7 Conclusion and discussion

A number of prerequisites have been outlined in this paper for consideration when designing an agent-based simulation model of the population dynamics of *E. saccharina*. This model is expected to aid primarily in investigating possible mechanisms for the control or eradication of *E. saccharina* and, in doing so, diminish its negative effects on sugarcane production.

The next phase in this ongoing research project involves a detailed development of the proposed agent-based simulation which incorporates biological and behavioural attributes of the stalk borer discussed in §5 within the framework suggested in §6.

References

[1] Atkinson PR, 1979, *Distribution and natural hosts of Eldana saccharina Walker in Natal, its oviposition sites and feeding patterns*, Proceedings of the South African Sugar Technologists Association,

53, pp. 111–115.

[2] Atkinson PR, 1980, *On the biology, distribution and natural host-plants of Eldana saccharina Walker*, Journal of the Entomological Society of South Africa, **43**, pp. 171–194.

[3] Atkinson PR, 1981, *Mating behaviour and activity patterns of Eldana saccharina Walker (Lepidoptera: Pyralidae)*, Journal of the Entomological Society of South Africa, **44(2)**, pp. 265–281.

[4] Atkinson PR, 1984, *Seasonal cycles of Eldana borer in relation to available control measures*, Proceedings of the South African Sugar Technologists Association, **58**, pp. 165–167.

[5] Atkinson PR & Carnegie AJM, 1989, *Population dynamics of the sugarcane borer, Eldana saccharina Walker (Lepidoptera: Pyralidae), in Natal, South Africa*, Bulletin of Entomological Research, **79**, pp. 61–80.

[6] Carnegie AJM, 1974, *A recrudescence of the borer Eldana saccharina Walker (Lepidoptera: Pyralididae)*, Proceedings of the South African Sugar Technologists Association, **55**, pp. 116–119.

[7] Carnegie AJM, 1981, *Combating Eldana saccharina Walker: A progress report*, Proceedings of the South African Sugar Technologists Association, **55**, pp. 116–119.

[8] Conlong DE, 1990, *A study of pest-parasitoid relationships in natural habitats: An aid towards the biological control of Eldana saccharina (Lepidoptera: Pyralidae) in sugarcane*, Proceedings of the South African Sugar Technologists Association, **64**, pp. 111–115.

[9] Conlong DE, 2014, *Senior Entomologist at the South African Sugarcane Research Institute*, Mount Edgecombe, [Personal Communication], Contactable at `Des.Conlong@sugar.org.za`

[10] Conlong DE & Rutherford RS, 2009, *Conventional and new biological and habitat interventions for integrated pest management systems: Review and case studies using Eldana saccharina Walker (Lepidoptera: Pyralidae)*, pp. 241–261 in Peshin R & Dhawan AK (Eds), *Integrated pest management: Innovation-development process*, Springer, New York (NY).

[11] Goebel FR & Way MJ, 2003, *Investigation of the impact of Eldana saccharina (Lepidoptera: Pyralidae) on sugarcane yield in field trials in Zululand*, Proceedings of the South African Sugar Technologists Association, **26**, pp. 805–814.

[12] Hearne JW, van Coller LM & Conlong DE, 1991, *Determining strategies for the biological control of a sugarcane stalk borer*, Ecological Modelling, **73**, pp. 117–133.

[13] Heathcote RJ, 1984, *Insecticide testing against Eldana saccharina Walker*, Proceedings of the South African Sugar Technologists Association, **58**, pp. 154–158.

[14] Horton PM, Hearne JW, Apaloo J, Conlong DE, Way MJ & Uys P, 2002, *Investigating strategies for minimising damage caused by the sugarcane pest Eldana saccharina*, Agricultural Systems, **74**, pp. 271–286.

[15] Janssen MA, 2005, *Agent-based modelling*, prepared for the Internet Encyclopaedia of Ecological Economics, [Online], [Cited April 4th, 2014] Available from `http://isecoeco.org/pdf/agent_based_modeling.pdf`

[16] Keeping MG & Rutherford RS, 2004, *Resistance mechanisms of South African sugarcane to the stalk borer Eldana saccharina (Lepidoptera: Pyralidae): A review*, Proceedings of the South African Sugar Technologists Association, **78**, pp. 307–312.

[17] King AG, 1989, *An assessment of the loss in sucrose yield caused by the stalk borer, Eldana saccharina, in Swaziland*, Proceedings of the South African Sugar Technologists Association, **63**, pp. 197–201.

[18] Potgieter L, 2013, *A mathematical model for the control of Eldana saccharina Walker using the sterile insect technique*, PhD Dissertation, Stellenbosch University, Stellenbosch.

[19] Potgieter L, van Vuuren JH & Conlong DE, 2011, *Modelling the effects of the sterile insect technique applied to Eldana saccharina Walker in sugarcane*, ORiON, **28(2)**, pp. 59-84.

[20] Potgieter L, van Vuuren JH & Conlong DE, 2012, *A reaction-diffusion model for the control of Eldana saccharina Walker in sugarcane using the sterile insect technique*, Ecological Modelling, **250(2013)**, pp. 319–328.

[21] Schumacher P, Weyeneth A, Webe DC & Dorn S, 1997, *Long flights in Cydia pomonella L.(Lepidoptera: Tortricidae) measured by a flight mill: influence of sex, mated status and age*, Physiological Entomology, **22**, pp. 149–160.

[22] South African Sugar Association, 2013, *SASA: Facts and figures*, [Online], [Cited April 4th, 2014], Available from `http://www.sasa.org.za/sugar_industry`

[23] Snyman SJ, Baker C, Hucket BI, McFarlane SA, van Antwerp T, Berry S, Omarjee J, Rutherford RS & Watt DA, 2008, *South African Sugar Research Institute: Embracing biotechnology for crop improvement research*, Sugartech, **10(1)**, pp. 1–13.

[24] van Coller LM, 1992, *Optimum biological control strategies for a problem in the sugar industry — A mathematical modelling approach*, MSc Thesis, University of Natal, Pietermaritzburg.

[25] Way MJ, 1995, *Developmental biology of the immature stages of Eldana saccharina Walker (Lepidoptera: Pyralidae)*, Proceedings of the South African Sugar Technologists Association, **69**, pp. 83–86.

[26] XJ Technologies Company, 2007, *AnyLogic 6 Users Guide*, [Online], [Cited April 2nd, 2014], Available from `https://www7.informatik.uni-erlangen.de`

Proceedings of the 2014 ORSSA Annual Conference
pp. 71–78
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Semi-automated maritime vessel activity detection using hidden Markov models

J du Toit* JH van Vuuren†

Abstract

Maritime surveillance systems make use of a dearth of sensor data which often include spatio-temporal vessel updates provided by vessels fitted with onboard self-reporting Automatic Identification Systems. These spatio-temporal updates supply low-level information to an operator tasked with observing the surveillance scene and identifying threatening or undesirable behaviour. In this situation, the operator is thus required to interpret the updates by attaching semantic or high-level information to these data.
To this end, automatic activity detection is pursued in this paper as a means to describe vessel motion patterns within the surveillance scene. In particular, the activity of vessels travelling along a well-established route is investigated. Spatial regions of interest are extracted from historical data using a simple spatial clustering technique. The resulting data set is further reduced by removing outliers subject to chosen features before the remaining patterns are clustered. With the assistance of an operator, who may attribute activities to clusters that have some geographical or behavioural meaning, this approach may contribute to a rudimentary understanding of the scene. The motion patterns within these clusters provide the training data for hidden Markov models which are tasked with classifying newly observed motion patterns that engage in the suggested activity. This process of enriching the vessel updates with semantics is expected to lead to more effective decision making on the part of a maritime surveillance operator who may thus direct cognitive resources towards unknown activities.

Key words: Maritime surveillance, motion patterns, activity detection, hidden Markov models, DBSCAN, dynamic time warping, partitioning around medoids.

1 Introduction

The act of surveillance is the systematic observation of regions, entities or objects by visual, aural, electronic or other means [7]. Integrated systems designed for such tasks have been deployed over a broad spectrum of scenarios, typically with the common purpose of providing security. Areas of application include traffic monitoring on motorways, the

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 13421530@sun.ac.za

†**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

monitoring of public spaces and quality control of industrial processes [6, 8, 14]. As many of these systems mature, their focus invariably expands from observation and data collection to automated decision support and intelligent action[1]. In the video surveillance community, these systems are therefore referred to as second generation systems [20]. This shift to automation assists surveillance operators by alleviating their workload and potentially directing their attention to activities of interest.

These notions are also of import in maritime surveillance, which is driven by the societal demand for improved security. The large volumes of trade passing through ports[2] along with the responsibility of coastal states to protect their natural resources, their citizens, sea-faring vessels and the personal safety of mariners, necessitates such measures. Any maritime entities which endanger these notions are considered *threats* or are considered to be exhibiting threatening behaviour. For example, in a vessel traffic control scenario, the threat of collision is ever present and acts of poaching and pollution are threatening to the environment. The timely identification and effective response to threats is of great importance in mitigating their impact or neutralising them entirely.

In this paper a data-driven approach is taken to identifying the activity of vessels travelling along a well-established route. A similar approach was applied with success in traffic surveillance [13] and its applicability to *Automatic Identification System* (AIS) data is explored in this paper, where the training data are determined using data mining.

2 Related work and discussion

Algorithms capable of identifying activities in a maritime surveillance picture rely in part on spatio-temporal sensory data[3]. The participation in an activity by an entity may be determined from a single datum or from a sub-sequence of its recorded trajectory data (an example of the former is the event that a vessel enters a no-go area, while an example of the latter is the *Williamson turn* manoeuvre which brings a vessel about in the case of a man overboard). However, a significant amount of information is captured in the sequential and temporal nature of the data, requiring time and resource consuming spatio-temporal analyses by sub-components of such a surveillance system. A fundamental problem in time-series analysis and storage is the question of representation. Various approaches towards time series-representation, along with series indexing and similarity, have been researched extensively by the data mining community [4]. Time-series matching has been investigated from both the perspectives of entire sequence matching and sub-sequence matching, using methods such as *dynamic time warping* (DTW) [19] and the *longest common sub-sequence* model [21].

[1] *Closed circuit television* camera systems bear testament to this evolution. As technology has advanced and larger quantities of data can be collected, it has become necessary to develop systems capable of automatically detecting events [20].

[2] Approximately 95% of all trade in the Southern African Development Community passes through South African and East African ports [16].

[3] Cooperative reporting mechanisms employed by authorities include the requirement that a vessel docking at a South African port notifies the Maritime Rescue Coordination Centre ninety six hours before arrival at port. The self-reporting *automatic identification system* is another example of a cooperative data source, whereas radars are non-cooperative data sources which collect vessel kinematic data.

The non-trivial task of pattern discovery in temporal data has been approached using distance-based clustering techniques and the statistical model-based technique of *hidden Markov models* (HMMs) [15]. HMMs are also ideally suited as classifiers of temporal sequences where the parameters of an HMM are estimated from training data. This approach has been applied successfully in various contexts where trained HMMs act as representatives of the class on which they were trained [13].

In order for a surveillance system to be able to assist operators in the decision making process of identifying threats at sea, mechanisms are required by which semantics may be added to low-level information associated with vessels, such as raw kinematic data. De Vries *et al.* [2] achieve this by enriching their vessel trajectories with geographical domain knowledge whilst Makris *et al.* [13] identify spatial regions and model the dynamics of objects moving between them using HMMs in a traffic surveillance setting. Although the former method is concerned with identifying high-level behaviours in the maritime domain, many researchers direct their efforts towards the identification of anomalous behaviour in vessel traffic [9, 10, 18]. These approaches are predominantly concerned with instantaneous updates and attempt to build a picture of normality based on historical data. One such approach is the discretization of the surveillance picture into cells on which *kernel density estimation* (KDE) is performed [11]. The data points within those cells are assumed to represent normal vessel behaviour and vessels that deviate significantly from this estimated cell-model are flagged as anomalous. All of these approaches rely on vast quantities of data, and although methods such as KDE are unsupervised, classification techniques such as HMMs additionally require labelled training data. It should also be noted that illicit and threatening behaviour is very seldom observed, thus making their modelling via data-driven approaches more difficult. Anomaly detection thus assumes that all anomalous behaviour is to be considered threatening. As a counterpoint to this approach, rule-based approaches are typically employed to identify specific threats, such as vessels in pursuit of one another or vessels meeting at sea [1]. However, a disadvantage of this approach is that each threatening scenario must then be determined and integrated into the system beforehand.

3 Methodology

A filtering approach is pursued in this paper in combination with clustering in an effort to extract trajectories that are relatively compact in space. Contrary to traffic surveillance applications, vessel trajectories are geographically less constrained. It is expected that trajectories along established routes should be well represented in origin and destination clusters[4]. The method of *density-based spatial clustering of applications with noise*, referred to as DBSCAN, is used as the spatial clustering technique for which it is sufficient to consider simplified incarnations of the trajectories. The poly-line simplification technique of *Douglas-Peucker* (DP) [5] reduces the number of points in a piecewise linear path while retaining its shape. The simplified trajectories are clustered by DBSCAN (using the

[4] A decision support system would likely apply viewport clipping or have particular models in operation over certain areas of interest. These viewports would induce origin and destination regions at their boundaries.

same threshold used in DP).

These regions are used to further reduce the data set by discarding trajectories that do not have them in common and trajectories that begin and end in the same region. This reduced data set is expected to contain trajectories that differ markedly from the majority of retained trajectories, and outlier removal with respect to the derived attribute of *sinuosity* is pursued.

Lastly, a final clustering is obtained via the *partitioning around medoids* (PAM) method [3] which utilises *dynamic time warping* (DTW) as a measure of the similarity between vessel trajectories with respect to their positional data. This alignment technique allows contractions or dilations of the temporal axis whilst determining an optimal alignment between two sequences subject to monotonicity and step continuity constraints [3].

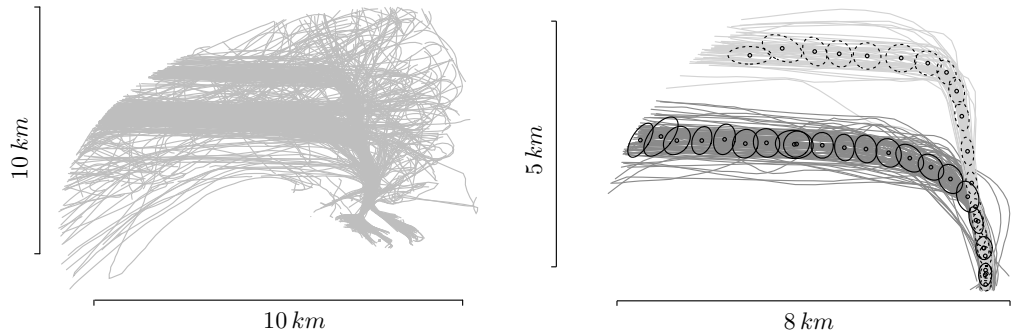
These clusters provide the training data for as many HMMs. Sequential data may be represented by an HMM in which it is assumed that the underlying Markov process is hidden, but that the process emits observable quantities. In this setting, the aforementioned features are observable, but the dynamical process leading to their observation is considered unobservable. Suppose the k-th observable feature in a set of M features is denoted by v_k. Then an HMM [17] is specified by a set $S = \{S_1, S_2, \ldots, S_N\}$ of hidden states and an associated $N \times N$ transition probability matrix $A = [a_{ij}]$ that describes the probability of transitioning from state S_i to S_j, *i.e.* $a_{ij} = p(q_{t+1} = S_j | q_t = S_i)$ for all $1 \leq i, j \leq N$, where $a_{ij} \geq 0$ and q_t denotes the state occuring at time t. Furthermore, a probability distribution describing the probability that v_k may be observed when the system is in state S_j is captured in an $N \times M$ *emission* matrix $B = [b_j(k)]$, where $b_j(k) = p(v_k \text{ at time } t | q_t = S_j)$ for all $1 \leq j \leq N$ and all $1 \leq k \leq M$. Finally, the initial state probability distribution is represented by $\pi = [\pi_i]$, where $\pi_i = p(q_1 = S_i)$ for all $1 \leq i \leq N$.

Once the HMMs have been trained, it is possible to determine whether an observed trajectory was generated by a particular HMM, by determining whether the log-likelihood of the observation, given the HMM model, is greater than a specified threshold.

4 Vessel data

A data set comprising AIS reports was obtained from vessels in the region of Cape Town harbour for use in this paper (see Figure 1(a)). Only the kinematic quantities of the vessels were considered, *i.e.* each vessel's position (reported in geographical coordinates), heading and speed, together with a time-stamp associated with the report. Reports emanating from within a nine-kilometre radius of a chosen reference point at Cape Town harbour over a two-month period were used in the analysis (the reference point is indicated by the filled black circle in Figure 1(a)). Spurious updates were discarded[5] and individual trajectories were partitioned into *stop* and *move* segments. A vessel is deemed to have come to a stop if the reported speed falls below a particular threshold for a sufficient amount of time.

[5]Reports have been found to contain reported speeds of hundreds of knots whilst some reported positions would require a vessel to achieve impossible speeds in order to reach that location in the provided time.



(a) AIS data within a 9 km radius of the reference point.

(b) The sample means and sample covariances for the bivariate positional data.

Figure 1: The data used for training and testing.

A clustering of vessel data into two clusters is shown in Figure 1(b) in which the the upper cluster is composed of vessel trajectories resulting from vessels departing from Cape Town harbour whilst the lower cluster comprises trajectories entering the scene and travelling to the harbour. The number of vessel trajectories utilised in this paper are presented in Table 1.

	Total trajectories	Training set	Validation set	Test set
Entry cluster	97	67	10	20
Exit cluster	53	30	8	15
Unclustered	280			

Table 1: The number of vessel trajectories used in this paper.

5 HMM Structure and Training

A *left-to-right* HMM structure was utilised in capturing the sequential nature of the vessel reports. These reports are described by a state-dependent distribution $p(X_t|S_t)$, where $X_t = (x_t, y_t, u_t, v_t)$ is an observation vector comprising positional (x_t, y_t) and velocity (u_t, v_t) data at time t. This distribution was modelled as a multivariate Gaussian distribution, in keeping with the approach of [13]. Furthermore, the *Bayesian information criterion*[6] was used to inform the choice, in favour of mixture distributions.

The initial transition probability matrix A was specified as an upper triangular matrix ($a_{ij} = 0$ for all $i > j$) so as to ensure that the left-to-right structure is maintained during parameter estimation and to lessen the number of free parameters. Similarly, for each HMM, the initial state distribution was taken as $\pi_0 = (1, 0, \ldots, 0)$, thus ensuring that each trajectory begins in the first state. The number of states were arbitrarily chosen in this instance, but they may also have been determined via model selection methods

[6]The *Bayesian information criterion* is a model selection criterion which is expressed as $BIC = -2\log L + p\log T$, where L is the log-likelihood of the fitted model, p is the number of parameters of the model and T is the number of observations [12]. This measure typically favours models with fewer parameters.

(such as the aforementioned BIC). An HMM was fitted to the training data using the *Baum-Welch* method[7], the convergence of which is assisted by the initialisation of the parameters of the state-dependent distributions. The medoids determined by PAM in the data mining step were linearly interpolated at intervals that produce the desired number of states and all points within a radius of this interval length were used to compute the sample means and covariances for each state-dependent distribution (the bivariate case is illustrated in Figure 1(b)).

6 Activity Classification

The estimated HMMs were considered to represent an activity in the scene and they may be labelled as such. For instance, a great deal of vessel traffic is observed travelling to and from Cape Town harbour that may successfully be classified by an HMM corresponding to the route that they have taken. Thresholds for class membership were determined from a validation set $\mathbb{V}$ (a portion of the test set was withheld during training) by selecting the minimum log likelihood value attained by members of each class with respect to its corresponding HMM. A vessel trajectory $X = X_0, \ldots, X_n$ was deemed to be a member of a class $\mathcal{C}_i$ if the probability of the sequence of observations, given the corresponding HMM λ_i, is less than the selected threshold for the i-th class. That is, $\log p(X|\lambda_i) > \min_{Y \in \mathbb{V}} \log(p(Y|\lambda_i))$. Trajectories that do not feature a displacement of more than five kilometres were not considered for classification.

7 Results

Two HMMs were estimated from the clustered data, corresponding to entry and exit trajectories. The test sets were classified once the thresholds had been chosen for each HMM. The resulting number of *true positive* (TP), *false positive* (FP), *true negative* (TN) and *false negative* (FN) classifications for the test sets are shown in Table 2. The results of calculating the membership of the trajectories in the unclustered data set are also reported in this table.

	Test sets				Unclustered set	
	TP	TN	FP	FN	TN	FP
Entry HMM	22	15	0	0	241	39
Exit HMM	13	22	0	2	278	2

Table 2: The classification results of the entry and exit test sets for each HMM, as well as classification results for the unclustered set.

Closer inspection of the data revealed that vessels approaching the harbour often reduce speed or perform loop manoeuvres in the entry channel. The latter trajectories were removed from the training data set in the data mining phase as they are considered to be outliers with regard to the sinuosity measure, whilst the former are regarded to have

[7]The Baum-Welch method is a special case of the *expectation-maximisation* algorithm and is not guaranteed to find a global optimum [12].

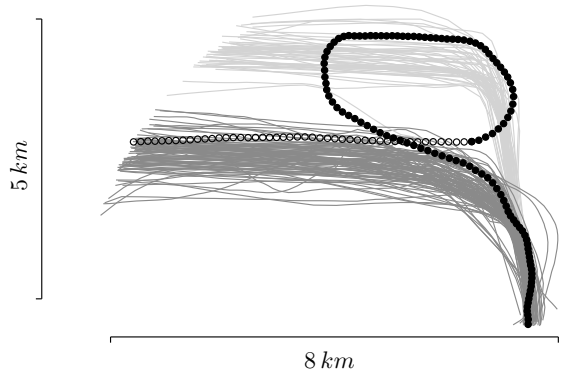


Figure 2: The filled circles indicate the updates for which the entry HMM successfully classified the trajectory.

come to a stop and the trajectories are therefore subdivided into two segments. The data mining process assigned these trajectories to the unclustered set, but the entry HMM still successfully classified them.

Lastly, the responsiveness to a change in behaviour whilst a vessel is underway was investigated. The vessel in Figure 2 changes lanes by travelling along the entry route and then switching to the exit route. The updates along its trajectory for which it was classified by the entry HMM are indicated by open circles. The updates for which the trajectory is no longer described by either of the HMMs, are indicated by solid circles.

8 Conclusion

Selecting position and velocity as the features for training HMMs results in a satisfactory classification of trajectories on the limited AIS test data. However, the use of a single classifier to describe vessels travelling towards the port is insufficient as there are vessels that slow down on their approach and vessels that do not.

The performance of the classification models was found to rely heavily on the training data extracted via the filtering and clustering method. The use of the filters and features of sinuosity were demonstrated to be useful in extracting spatially compact trajectories and the resulting data were shown to lend themselves to HMM classification and training.

References

[1] Brax C & Niklasson L, 2009, *Enhanced situational awareness in the maritime domain: An agent-based approach for situation management*, Proceedings of the SPIE: Intelligent Sensing, Situation Management, Impact Assessment, and Cyber-Sensing, Orlando (FL), pp. 3–13.

[2] De Vries G, Van Hage W & Van Someren M, 2010, *Comparing vessel trajectories using geographical domain knowledge and alignments*, Proceedings of the 10^{th} IEEE International Conference on Data Mining Workshops, Sydney, pp. 209–216.

[3] Du Toit J & Van Vuuren JH, 2012, *Towards coastal threat evaluation decision support*, Proceedings of the 41st Annual Conference of the Operations Research Society of South Africa, pp. 31–39.

[4] Fu T, 2011, *A review on time series data mining*, Engineering Applications of Artificial Intelligence, **24(1)**, pp. 164–181.

[5] Heckbert PS & Garland M, 1997, *Survey of polygonal surface simplification algorithms*, (Unpublished) Technical Report, Carnegie-Mellon University, Pittsburgh (PA).

[6] Hu W, Tan T, Wang L & Maybank S, 2004, *A survey on visual surveillance of object motion and behaviors*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, **34(3)**, pp. 334–352.

[7] Ince AN, Topuz E & Panayirci E, 2000, *Principles of integrated maritime surveillance systems*, Kluwer Academic, Norwell (MA).

[8] Kumar A, 2008, *Computer-vision-based fabric defect detection: A survey*, IEEE Transactions on Industrial Electronics, **55(1)**, pp. 348–363.

[9] Lane R, Nevell D, Hayward S & Beaney T, 2010, *Maritime anomaly detection and threat assessment*, Proceedings of the 13th International Conference on Information Fusion (FUSION'10), pp. 1–8.

[10] Laxhammar R, 2008, *Anomaly detection for sea surveillance*, Proceedings of the 11th International Conference on Information Fusion (FUSION'08), pp. 55–62.

[11] Laxhammar R, Falkman G & Sviestins E, 2009, *Anomaly detection in sea traffic—A comparison of the Gaussian mixture model and the kernel density estimator*, Proceedings of the 12th International Conference on Information Fusion (FUSION'09), pp. 756–763.

[12] MacDonald IL & Zucchini W, 2009, *Hidden Markov and other models for discrete valued time series*, CRC Press, Boca Raton (FL).

[13] Makris D & Ellis T, 2005, *Learning semantic scene models from observing activity in visual surveillance*, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, **35(3)**, pp. 397–408.

[14] McCahill M & Norris C, 2002, *CCTV in Britain*, (Unpublished) Technical Report, Centre for Criminology and Criminal Justice, University of Hull, Hull.

[15] Oates T, Firoiu L & Cohen P, 1999, *Clustering time series with hidden Markov models and dynamic time warping*, Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, pp. 17–21.

[16] *Ports and ships*, [Online], [Cited June 10th, 2013], Available from `http://ports.co.za/index.php`.

[17] Rabiner LR, 1989, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, **77(2)**, pp. 257–286.

[18] Ristic B, La Scala B, Morelande M & Gordon N, 2008, *Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction*, Proceedings of the 11th International Conference on Information Fusion (FUSION'08), pp. 40–46.

[19] Salvador S & Chan P, 2007, *Toward accurate dynamic time warping in linear time and space*, Intelligent Data Analysis, **11(5)**, pp. 561–580.

[20] Valera M & Velastin S, 2005, *Intelligent distributed surveillance systems: A review*, Proceedings of the 2th IEEE Conference on Vision, Image and Signal Processing, pp. 192–204.

[21] Vlachos M, Kollios G & Gunopulos D, 2002, *Discovering similar multidimensional trajectories*, Proceedings of the 18th International Conference on Data Engineering, San Jose (CA), pp. 673–684.

Proceedings of the 2014 ORSSA Annual Conference
pp. 79–87
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Solution representation for a maritime law enforcement response selection problem

A Colmant* & JH van Vuuren†

Abstract

Designing a *maritime law enforcement* (MLE) response selection decision support system requires, *inter alia*, an optimization methodology component in which solution search methods are used to provide the decision maker with a set of high-quality solution alternatives to a particular problem instance. In order to facilitate this process, solutions should be encoded in very specific data formats which allow for effective application of local search operations, easy evaluation of objective function values and tests for solution feasibility. The various complex dynamic features associated with this problem, however, make it difficult to standardise these data formats to be used as part of a neighbourhood search process. Consequently, the aim in this paper is to propose an effective solution data encoding scheme that can be incorporated into a real-time MLE response selection decision support system.

1 Introduction

Based on the detection and evaluation of potentially threatening *vessels of interest* (VOIs) at sea, a *maritime law enforcement* (MLE) response selection *decision support system* (DSS) aims to assist human operators in solving the so-called *MLE response selection problem* — allocating and routing of MLE resources, such as patrol vessels, military vessels and armed helicopters, for the purpose of intercepting VOIs.

The proposed DSS consists of a *model generation* subsystem and a *solution search* subsystem. The former subsystem comprises a *model selection* component, which includes the construction and storage of fundamental mathematical structures and fixed modelling components (such as objective functions, routing constraints and MLE resource parameters), and a *model management* component, which consists of dynamic features that are used to model the problem on a temporal basis. The latter subsystem is concerned with finding and presenting a set of non-dominated solutions in multiple objective space to the

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 15427498@sun.ac.za

†**(Fellow of the Operations Research Society of South Africa)**, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

MLE response selection operator for every problem instance formulated in the model generation subsystem. These subsystems share input data *via* a centralised MLE database, as illustrated schematically in Figure 1.

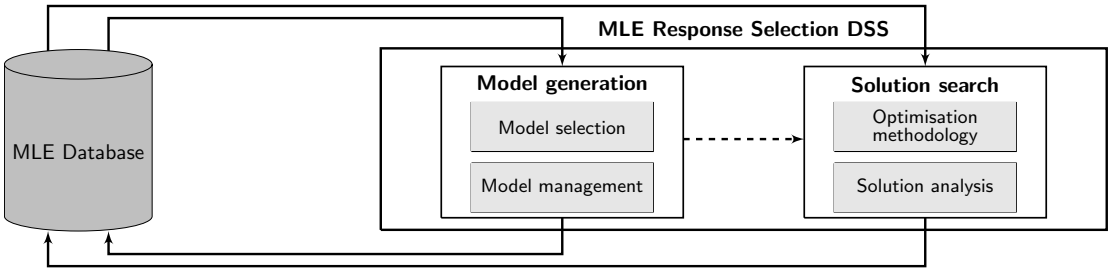


Figure 1: *The MLE response selection DSS with its two subsystems and their components.*

The focus of this paper is on the solution search subsystem of this DSS and the objective is to put forward a flexible solution encoding scheme facilitating the use of a wide range of local search transformation operators and taking into account the crucial end-of-route decisions and other model management aspects associated with the MLE response selection problem.

This paper is structured as follows. Since the MLE response selection problem can be modelled as a multi-depot *vehicle routing problem* (VRP) with a heterogeneous vehicle fleet, a brief literature review on this class of problems is presented in §2, and this is followed by a generic graphical representation of MLE response selection operations in §3. In §4, a discussion is conducted on the end-of-route assignment of MLE resources. Certain features of the model management component, demonstrating the dynamism of the model generation subsystem, are then presented in §5. A method for encoding routing solutions in an MLE response selection DSS is proposed in §6, after which the paper closes with some concluding remarks in §7.

2 Literature review

The multi-depot VRP is known in the literature to be a variant of the capacitated VRP in which routes are simultaneously sought for several vehicles originating from multiple depots, serving a fixed set of customers and then returning to their original depots. Compared to other capacitated VRP variants, a smaller volume of research has been done on the multi-depot VRP, but a number of solution representation techniques have nevertheless appeared in the literature for this problem. Most of these techniques tend to break down an instance into a series of single depot sub-instances and/or only solve it for a single objective and a homogeneous fleet of vehicles [2, 3, 5, 6]. Furthermore, because vehicles always start and finish their routes at the same depots, it is easy to merge the set of customer vertices with that of the depot vertices in such simplified model formulations.

The innovative work of Salhi *et al.* [4], on the other hand, provides a complete mixed integer linear formulation for a generic single-objective multi-depot VRP for a heterogeneous vehicle fleet. In particular, they offer formulation variants for alternative multi-depot

VRP scenarios that are relevant in the MLE response selection routing problem. For instance, they investigate multi-depot VRP variants in which the number of vehicles of a given type is known; some types of vehicles cannot be accommodated at certain depots; or a vehicle is not required to return to the same depot from whence it originated. They then solve the problem using a variable neighbourhood search applied to the notion of *borderline customers*, using six different local search operators.

As described in [1], an MLE response selection problem instance may be modelled as a special type of VRP in which the depots represent the bases from whence MLE resources are dispatched, the fleet of vehicles represents the fleet of MLE resources and the customers represent the VOIs tracked at sea within the territorial waters of the coastal nation. A list of differences between this particular VRP and typical capacitated VRPs encountered in the literature was also given in [1]. The notion of *time stages* was then incorporated into the model formulation in order to accommodate the dynamic nature of the problem: an MLE response selection environment is subjected to so-called *disturbances*, which are threshold phenomena occurring stochastically over time that may cause the currently implemented solution to suffer significantly in terms of quality, hence triggering a new time stage during which the current situation is re-evaluated (*i.e.* the instance is re-solved under the original information combined with the data update which brought along the disturbance).

3 A graphical representation of MLE response selection

A generic graph structure is used to represent the interaction amongst the entities in an MLE response selection environment. The vertex set of this graph is an extension of the vertex set presented in [1], where only the VOIs and MLE resources previously formed part of vertex set for a single depot.

For any given time stage, the vertices in an MLE response selection environment may be partitioned into four sets: VOIs, MLE resources (both active and idle[1]), patrol circuits and bases. While the set of VOIs is typically updated at the start of every time stage due to its high level of dynamism[2], the other three sets remain somewhat more fixed and are updated independently from time stages[3].

Henceforth, let $V^e(\tau) = \{1, \ldots, n(\tau)\}$ represent the set of VOIs at the beginning of time stage τ, let $V^r = \{1, \ldots, m\}$ be the set of MLE resources with respective initial spatial locations $V_0^r(\tau) = \{1_0, \ldots, m_0\}$, let $V^b = \{1, \ldots, |V^b|\}$ denote the set of bases and let $V^p = \{1, \ldots, |V^p|\}$ represent a set of pre-determined patrol circuits. Additionally, let $V(\tau) = V^e(\tau) \cup V^r \cup V^b \cup V^p$. These vertex subsets, along with the arcs inter-linking them, form the directed graph $G(V(\tau), E(\tau))$ depicted in Figure 2, where $E(\tau)$ is the

[1]MLE resources are generally either allocated for the purpose of intercepting VOIs at sea (these MLE resources are then defined as being in a so-called *active* state) or are otherwise strategically allocated to patrol certain areas at sea until needed for law enforcement purposes (these MLE resources are then defined as being in a so-called *idle* state).

[2]Changes in input data linked to a VOI also cause the whole set to be updated.

[3]With occasional exceptions, such as a disturbance caused by an active MLE resource breaking down at sea.

set of pre-calculated arcs linking the vertices and where all pairs of vertices in $V^e(\tau)$ are reachable from one another. Finally, let $V_k^e(\tau) \subseteq V^e(\tau)$ be the set of VOIs scheduled to be investigated by MLE resource k during time stage τ. It is therefore assumed that active MLE resources in V^r are assigned to investigate subsets of VOIs in $V^e(\tau)$ during time stage τ, after which they are assigned to either travel back to a base in V^b or to a designated patrol circuit in V^p.

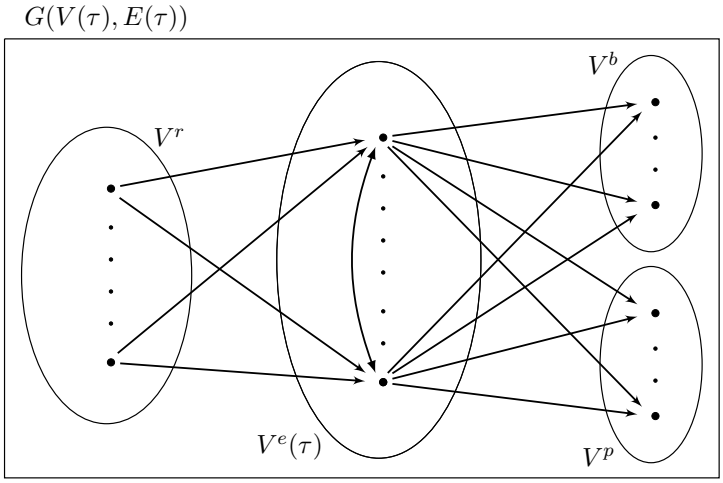


Figure 2: *Graph representation of decisions in an MLE response selection environment.*

4 On end-of-route assignments

This section is devoted to a discussion on the last arc of every route, where active MLE resources transfer from an active state to an idle state in a sub-process defined as *end-of-route* or *post-mission* assignments. This feature of the MLE response selection problem, located in the model management component of the MLE response selection DSS, requires a certain form of input from the idle MLE resources DSS[4]. As discussed in the previous section, after investigating the VOIs assigned to it (*i.e.* completing its mission), the idle MLE resources management operator may either assign an MLE resource to travel to a designated base or to join a designated patrol circuit. Because the MLE response selection process determines where MLE resources will be located in space after completing their missions, it is crucial to consider the impact of this final route arc with respect to operating costs and distance-constrained feasibility.

Due to possible distance-constrained feasibility issues, and because the idle MLE resources management operator cannot know *a priori* where active MLE resources will be located in space after investigating the last VOIs on their routes, so-called *autonomy thresholds* are incorporated into the model formulation. These thresholds ensure that an MLE resource is only allowed to join a patrol circuit after completing its mission provided that the

[4]The DSS in support of the allocation of idle MLE resources in both time and space, which is external to the MLE response selection DSS.

travel distance to the circuit is within a certain autonomy level[5]. Ultimately, a route may only be classified as distance-constrained feasible if there exists at least one approved base that is at most as far away as the autonomy level threshold associated with the MLE resource after having investigated the last VOI on its route. In any case, the idle MLE resources management operator has the power to allocate the resources as soon as they complete their missions, as they then transit into an idle state. Consequently, these idle MLE resources do not have to follow the pre-scheduled end-of-route assignment as dictated by the MLE response selection solution uncovered during the search process at the beginning of the time stage since end-of-route preferences may in the meantime have evolved differently over time.

5 Other aspects of model management

The necessity of establishing a strong format of solution encoding originates from the highly dynamic nature of the MLE response selection problem, where subjective requirements are dictated by the MLE response selection operators on a temporal basis. More specifically, following a disturbance, or simply from a subjective preference point of view, the operators may wish to input additional information to a problem instance in order to accommodate a variety of special requests into the model prior to launching the solution search process for the next time stage. Certain features of the model management component are briefly presented in this section so as to demonstrate examples of dynamic elements introduced during the model generation process. The three features considered in this section are *VOI inclusion sets*, *VOI exclusion sets* and *end-of-route base exclusion sets*.

5.1 Inclusion sets for imposed VOI assignments

In this paper, only the most basic case of inclusion sets is considered, namely *unordered* VOI inclusion sets. These sets are used to force certain VOIs to be intercepted by certain MLE resources, but with no particular degree of urgency in respect of the order in which these VOIs are visited within their routes. Define the unordered inclusion set $I_k(\tau)$ to contain the VOIs forced to be included in the visitation route of MLE resource k during time stage τ. It is assumed that $I_k(\tau) \cap I_\ell(\tau) = \emptyset$ for all $\tau \in \mathbb{N}$ and all $k, \ell \in V^r$, with $k \neq \ell$. Furthermore, $I_k(\tau) \subseteq V_k^e(\tau)$ for all $\tau \in \mathbb{N}$ and $k \in V^r$. To incorporate the above-mentioned visitation requirements into the model formulation in [1], the set of constraints $\sum_{j \in V(\tau) \setminus V^r} x_{ijk}(\tau) = 1,\ i \in I_k(\tau),\ k \in V^r$, may (temporally) be included at the beginning of time stage τ, where $x_{ijk}(\tau)$ is a binary variable which assumes the value 1 if MLE resource k is scheduled to traverse arc (i, j) of the graph in Figure 2 during time stage τ.

[5]The autonomy level of an MLE resource with respect to distance, expressed as a function of time, measures the maximum distance that it may travel at sea before having to return to a designated base.

5.2 Exclusion sets for forbidden VOI assignments

Contrary to inclusion set requirements, the conditions imposed by VOI exclusion sets are met as long as the respective MLE resources are *not* scheduled to intercept VOIs specified in these sets. Define the exclusion set $E_k(\tau)$ to contain the VOIs forbidden to be included in the visitation route of MLE resource k during time stage τ. Then, the set of constraints $\sum_{j \in V(\tau) \setminus V^r} x_{ijk}(\tau) = 0,\ i \in E_k(\tau),\ k \in V^r$, may (temporally) be included in the formulation at the beginning of time stage τ.

5.3 Exclusion sets for forbidden end-of-route base assignments

As discussed in §4, a major dynamic aspect of the MLE response selection problem involves end-of-route assignment decisions, which consists of deciding where certain MLE resources should or should not be sent after completing their missions. For example, the idle management operator may want to control the distribution of idle MLE resources amongst the bases by managing their spread and strategic placement. This information is communicated to the response selection DSS at the beginning of every time stage.

Dictating end-of-route assignments may be achieved in a manner similar to forcing or prohibiting VOI visitations by certain MLE resources. Prohibiting these assignments, for instance, can be achieved by defining $B_k(\tau)$ to contain the bases forbidden to be scheduled for visitation by MLE resource k at the end of its route during time stage τ. Then, the set of constraints $\sum_{i \in V^e(\tau)} x_{ibk}(\tau) = 0,\ b \in B_k(\tau),\ k \in V^r$, may (temporally) be included in the formulation at the beginning of time stage τ.

6 Proposed solution representation scheme

In this section, a suitable method of encoding solutions to the MLE response selection problem is proposed. This encoding scheme is illustrated for a hypothetical problem instance with the following parameters: $V^b = \{B_1, B_2\}$, $V^r = \{a, b, c\}$, $V_0^r(\tau) = \{0_a, 0_b, 0_c\}$, $V^e = \{1, 2, 3, 4, 5, 6, 7\}$ and $V^p = \{P_1, P_2, P_3, P_4\}$.

6.1 Solution strings

In the literature, solutions to a VRP instance are typically encoded as *strings* which comprise substrings representing routes consisting of a subset of customers scheduled to be visited by a particular vehicle. The order in which customers are entered in such a substring is also the order in which the assigned vehicle visits them along its route.

An example of such a solution string for the above hypothetical MLE response selection problem instance is *String 1* of Table 1. In the first route (substring), for instance, MLE resource a is scheduled to first visit VOI 2, then VOI 5, after which it is scheduled to relocate to base B_1. In terms of the decision variables of the combinatorial optimization model proposed in [1], this part of the solution associated with MLE resource a may be written as $x_{0_a 2a} = 1$, $x_{25a} = 1$, $x_{5B_1 a} = 1$, and $x_{ija} = 0$ otherwise.

Because the initial and end-of-route cells in the above solution encoding are typically not involved in solution transformations, the string may be simplified before attempting to generate a neighbouring solution during the search process. In particular, the initial and end-of-route cells may be removed and a dummy cell, indicated by the zero element in *String 2* of Table 1 may be placed between routes of different MLE resources.

6.2 String configuration for VOI inclusion and exclusion sets

Part of configuring a solution string involves accommodating the various complexities associated with the dynamic features of the problem and neighbourhood search techniques that may be employed to solve the problem. The use of inclusion sets as proposed in §5.1, for example, imply that any solution transformation resulting in removing one or more VOIs belonging to inclusion sets from their respective routes will generate infeasible neighbouring solutions. One way of eliminating this shortcoming is to remove the VOIs belonging to inclusion sets from the solution string, carry out the solution transformation process with respect to the reduced string, and strategically reinsert these VOIs into feasible substrings[6] after completing the transformation process. Returning to our example, suppose that $I_a(\tau) = \{2\}$, $I_b(\tau) = \emptyset$ and $I_c(\tau) = \{4\}$. The VOIs belonging to any of these sets are then temporally removed from *String 2* to arrive at *String 3* of Table 1.

A random reduced neighbouring string may then be generated from the current reduced string, as shown in *String 4* of Table 1. Here, an inter-route transformation is performed, where VOI 3 and VOI 6 are removed from the second substring (route) and reverse-inserted into the first substring, while VOI 5 is removed from the first substring and inserted into the second substring (a popular neighbourhood move operator in the literature). Following this transformation, the VOIs that were temporally removed are placed back at random positions within their substrings, as shown in *String 5* in Table 1, so as to maintain feasibility.

String 1	$\{\{0_a, 2, 5, B_1\}; \{0_b, 3, 6, B_2\}; \{0_c, 4, 1, P_4\}\}$
String 2	$\{2, 5, 0, 3, 6, 0, 4, 1\}$
String 3	$\{5, 0, 3, 6, 0, 1\}$
String 4	$\{6, 3, 0, 5, 0, 1\}$
String 5	$\{6, 2, 3, 0, 5, 0, 4, 1\}$
String 6	$\{\{0_a, 6, 2, 3, B_2\}; \{0_b, 5, B_1\}; \{0_c, 4, 1, P_3\}\}$

Table 1: *String variants throughout a solution transformation process.*

6.3 String configuration for end-of-route assignments

The next step required to complete the solution string transformation process is to determine where MLE resources are scheduled to end their routes after investigating the VOIs assigned to them. In §4, it was proposed that the idle MLE resource management operator

[6]This does not mean that the generated neighbouring solution is feasible, as there are other routing feasibility criteria that have to be assessed.

must provide some form of input to the MLE response selection operator with respect to end-of-route assignment preferences. Such input may be configured as a set of preferred destinations associated with each active MLE resource, represented as sets containing one or more elements from the patrol circuit and MLE resource base sets. Given the reduced neighbouring solution obtained, as described in §6.2, and the autonomy thresholds, as described in §4, this input may furthermore be filtered by accounting for the positions of the MLE resources after completing their missions[7]. Because the lengths of the last arcs on every route only impact travelling costs in the objective space, the approved (and feasible) end-of-route vertices closest to the last VOI on their respective routes should therefore be configured as the vertices to be visited by the active MLE resources after completing their missions[8].

Let $\overline{B_k}(\tau)$ contain the set of bases that MLE resource k is allowed to travel to after completing its mission during time stage τ, noting that $\overline{B_k}(\tau) \subseteq V^b \backslash B_k(\tau)$. Similarly, let $\overline{P_k}(\tau)$ contain the set of patrol circuits that MLE resource k is allowed to join after its mission during time stage τ. Without loss of generality, it is assumed that an MLE will always be scheduled to join a patrol circuit at the end of its route provided that there is at least such a circuit to join; else it will relocate to one of the approved bases (if possible). Considering our example again, suppose that $\overline{B_a}(\tau) = \{B_2\}$, $\overline{P_a}(\tau) = \emptyset$, $\overline{B_b}(\tau) = \{B_1, B_2\}$, $\overline{P_b}(\tau) = \emptyset$, $\overline{B_c}(\tau) = \{B_1\}$ and $\overline{P_c}(\tau) = \{P_3, P_4\}$. Furthermore, suppose that B_1 is spatially closer to VOI 5 than B_2 is and suppose that P_3 is spatially closer to VOI 1 than P_4 is. Then, after removing the dummy cells and inserting bases or patrol circuits from the sets presented above at the end of their respective routes, the neighbouring string in *String 6* of Table 1 of the solution depicted in *String 1* of the same table results.

7 Conclusion

In this paper, a method of encoding solutions to the MLE response selection problem was proposed. This encoding scheme facilitates independent end-of-route assignments and has been designed for effective use in conjunction with most multi-objective local search techniques. The next step in designing the optimization methodology component is to build a metaheuristic engine that operates, *inter alia*, on the solution encoding scheme proposed (in particular with regard to crossover operators in genetic algorithms), which may be subjected to further adaptations, as well as to develop efficient re-insertion techniques for VOIs that are temporally removed from their routes during the process of generating neighbouring solutions.

References

[1] Colmant A & van Vuuren JH, 2013, *Prerequisites for the design of a maritime law enforcement resource assignment decision support system*, Proceedings of the 42^{nd} Annual Conference of the Operations Research Society of South Africa, pp. 90–101.

[7] If no such input exists for at least one of the routes, then, clearly, the entire neighbouring solution is classified as infeasible.

[8] The same reasoning cannot be used in the reinsertion process of §6.2, as more than one objectives are simultaneously affected.

[2] Ho W & Ji P, 2003, *Component scheduling for chip shooter machines: A hybrid genetic algorithm approach*, Computers & Operations Research, **30(14)**, pp. 2175–2189.

[3] Ho W, Ho G, Ji P & Lau H, 2007, *A hybrid genetic algorithm for the multi-depot vehicle routing problem*, Engineering Applications of Artificial Intelligence, **21(4)**, pp. 548–557.

[4] Salhi S, Imran A & Wassan NA, 2013, *The multi-depot vehicle routing problem with heterogeneous vehicle fleet: Formulation and a variable neighborhood search implementation* (Working Paper No. 277), [Online Accessed], Available from `http://www.kent.ac.uk/kbs/documents/res/working-papers/2013/mdvfmpaper(May2013)Web.pdf`.

[5] Surekha P & Sumathi S, 2011, *Solution to multi-depot vehicle routing problem using genetic algorithms*, World Applied Programming, **1(3)**, pp. 118–131.

[6] Vidal T, Crainic TG, Gendreau M, Lahrichi N & Rei W, 2012, *A hybrid genetic algorithm for multidepot and periodic vehicle routing problems*, Operations Research, **60(3)**, pp. 611–624.

Proceedings of the 2014 ORSSA Annual Conference
pp. 88–96
www.orssa.org.za/wiki/uploads/Conf/2014ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-656-6

Using agent-based simulation to explore sugarcane supply chain transport complexities at a mill scale

CS Price* D Moodley† CN Bezuidenhout‡

Abstract

The sugarcane supply chain (from sugarcane grower to mill) have particular challenges. One of these is that the growers have to deliver their cane to the mill before its quality degrades. The sugarcane supply chain typically consists of many growers and a mill. Growers deliver their cane daily during the milling season; the amount of cane they deliver depends on their farm size. Growers make decisions about when to harvest the cane, and the number and type of trucks needed to deliver their cane. The mill wants a consistent cane supply over the milling season. Growers are sometimes affected long queue lengths at the mill when they offload their cane.

A preliminary agent-based simulation model was developed to understand this complex system. The model inputs a number of growers, and the amount of cane they are to deliver over the milling season. The number of trucks needed by each grower is determined by the trip, loading and unloading times and the anticipated waiting time at the mill. The anticipated waiting time was varied to determine how many trucks would be needed in the system to deliver the week's cane allocation. As the anticipated waiting time increased, the number of trucks needed also increased, which in turn delayed the trucks when queuing at the mill. The growers' anticipated waiting times never matched the actual waiting times. The research shows the promise of agent-based models as a sense-making approach to understanding systems where there are many individuals who have autonomous behaviour, and whose actions and interactions can result in unexpected system-level behaviour.

Key words: Agent-based simulation, sugarcane supply chain, transport, queue.

1 Introduction

A supply chain is formed when different business entities co-operate to source raw materials, manufacture finished products and deliver these products to the market (Beamon,

*Corresponding author: School of Management, IT and Governance, University of KwaZulu-Natal (UKZN), and member of Centre for Artificial Intelligence Research (CAIR) (UKZN/CSIR), South Africa, Private Bag X54001, Durban, 4000, email: pricec@ukzn.ac.za

†School of Mathematics, Statistics and Computer Science, UKZN, and member of Centre for Artificial Intelligence Research (CAIR) (UKZN/CSIR), South Africa, Private Bag X54001, Durban, 4000, email: moodleyd37@ukzn.ac.za

‡SASRI Research Fellow, School of Engineering, UKZN, South Africa, Private Bag X1, Scottsville, Pietermaritzburg, 3209, email: bezuidenhoutc@ukzn.ac.za

1998). Materials flow forwards in the chain, and information (*e.g.* about how much of the material needs to be sent to the next entity, and production rates) flow backwards up the chain (Beamon, 1998). The demand for materials from the downstream business helps the upstream business to adapt to the rate of flow of materials in the chain as a whole (North & Macal, 2007).

The supply chain environment can be described as complex in terms of detail, with many variables which are interconnected; it is also complex from the point of view of the chain's dynamics, in that the variables are related in a non-linear way, with time delays, which makes cause-effect relationships more difficult to identify (Größler & Schieritz, 2005). For example, a participatory simulation game called the Beer Game (Sterman, 1989) was invented to show students the complexities and adaptive nature of the supply chain environment. In this game, the market demand was kept constant for a number of time periods to enable the participants to get to know the game and develop ordering rules. The demand was then doubled for the remainder of the time periods in the game. After the sudden doubling of the demand, it was found that the rest of the chain struggled to adapt the quantity of materials to send to the next stage in the chain, irrespective of how long the game continued (North & Macal, 2007).

Agricultural supply chains are those in which the raw material is gained as a result of a farming activity (Higgins *et al.*, 2010). They have more challenges than other manufacturing supply chains (Higgins *et al.,* 2007; Higgins *et al.*, 2010). They typically have thousands of participants rather than a few participating firms; in addition to responding to differing market demands, they are subject to uncertainties in weather and climate change (Higgins *et al.*, 2010).

The sugar supply chain is an agricultural supply chain with two parts: the sugarcane supply chain (where sugarcane is taken to the mill to be crushed) and the distribution chain for the stabilised raw sugar crystals (Bezuidenhout *et al.,* 2012). The first part of the sugar supply chain is of particular interest, because of the complexities outlined above in terms of the climatic environment, and because of the large number of role players. For example, one such sugarcane supply chain in KwaZulu-Natal had over 1000 participants (Le Masson, 2007). By contrast, after processing the sugarcane, the raw sugar belongs to one company, which is responsible for distributing it to its markets (Bezuidenhout *et al.,* 2012).

The sugarcane supply chain's role players consist of growers, harvesters, transporters, the mill and the Mill Group Board (Bezuidenhout *et al.,* 2012), which co-ordinates the supply chain and ensures a consistent supply of sugarcane to the mill during the milling season. A number of growers typically supply cane to one mill. Many of the growers transport their own sugar cane to the mill, whereas others use the services of contracted hauliers. During the 38-week long milling season, the growers need to deliver their cane to the mill as fast as possible after it has been harvested, as the sugarcane's quality declines thereafter. To ensure a consistent supply to the mill, growers deliver their cane daily. Among the problems faced by growers delivering their sugarcane to the mill are long queues at the mill area. If the growers have delivery problems, the mill could be starved of cane, which causes processing problems for the mill.

During one milling season, there are many interactions between similar and different types

of role players, and also the environment and the role players (for example, the onset of rainy weather could cause the sugarcane supply chain to work differently from how it works in fine weather). These interactions can have system-wide impacts which are difficult to explain from a local understanding of the role players and their actions. In addition, the relationships (and interactions) between the role players are not static for the whole milling season.

The aim of this research is to create a computational framework for the simulation and modeling of the sugarcane supply chain. The framework approaches the supply chain in a bottom-up way, so that the system-level effects of individuals' actions and interactions (such as queues at the mill, the delay between when the cane is harvested and crushed at the mill, and number of hours in which the mill is starved of cane) can be explored further.

2 Literature review

Owen *et al.* (2010) have identified three main ways of simulating supply chains: using System Dynamics, discrete event simulation models and agent-based modelling. System Dynamics models use a top-down approach to modelling, and the resultant models tend to be "highly aggregated, high-level" representations of the processes and flows in a system (North & Macal, 2007). These types of models assume that the processes do not change over time, and the ways in which the entities in the model relate to each other is static. Discrete event simulation models focus on processes and events which occur during the lifetime of the process (North & Macal, 2007). Like System Dynamics models, discrete event simulation models also assume fixed processes and interrelationships at the start of the simulation period (North & Macal, 2007). In these models, there is a single thread of control (Siebers *et al.*, 2010). Agent-based modelling, on the other hand, uses a bottom-up approach and caters for heterogeneous entities (called agents). This approach also assumes that relationships between the agents are not static in the lifetime of the model (North & Macal, 2007), and that control is decentralised since the agents behave independently of each other (Siebers *et al.*, 2010). Since there are many different types of decision-makers in the sugarcane supply chain, and each decision maker could have a different approach to making decisions, the agent-based modelling and simulation approach is better for modelling the complex dynamics of the supply chain.

Over recent years, several authors have developed simulation models of sugarcane supply chains. In Australia, Thorburn *et al.* (2005) created an agent-based model of regional sugarcane value chains to determine the wider impacts of implementing or increasing the electricity co-generation capacity in the chain. In their model, an agent represented the different supply chain sectors. As a result of the study, in one region, maximising electricity co-generation was abandoned because of the negative agronomic impact of this option. Le Masson (2007) developed a discrete event simulation of a KwaZulu-Natal sugarcane supply chain. This work revolved around testing the impacts of mechanical harvesting on the logistical supply to the mill – to see if the milling season length could be reduced. The simulation worked on a weekly basis (Le Gal *et al.,* 2009). McDonald *et al.* (2008) also developed a discrete event simulation of the same mill area as Le Masson (2007)'s work

Grower no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Distance to mill	3	3	4	5	5	6	7	9	10	13	13	13	13	16	16	16	17

Grower no.	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Distance to mill	19	20	20	20	21	22	23	24	25	27	27	30	33	36	36	37	42

Grower no.	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Distance to mill	43	43	44	47	48	48	48	50	52	54	54	55	55	57	57	58	

Table 1: *Grower distances to mill (in km).*

on an hourly basis, including spiller and bundle cane. Simulation models of the sugarcane supply chain which modelled at time steps of under one hour were not found.

3 Methodology and model development

An initial analysis of the sugarcane supply chain domain was carried out. This included visiting two different South African sugar mills, speaking to domain experts and reviewing published literature (articles and theses). The analysis showed that growers, hauliers (truck drivers) and the mill (including the weighbridge and mill yard) were the main role players and these became agents in the model. The analysis concentrated on what decisions the role players made and how the decisions were made.

An iterative model implementation approach was followed, as is recommended for developing agent-based simulations (North & Macal, 2007). The first iteration implemented a restricted list of features. More features were implemented in subsequent iterations of the model. After each iteration, the model was tested to ensure that the features were working as expected before proceeding to the next iteration. The model was developed in the Repast Symphony platform (North *et al.*, 2013).

The current model respresents the workings of a single sugar mill and the growers and hauliers which supply it. It models the activities of the grower, haulier and mill every minute. In this model, each grower uses his own truck(s) to take the cane to the mill. The model uses similar input data (number of growers, their distance from the mill) to that of a KwaZulu-Natal sugarcane supply chain (Le Masson, 2007; McDonald *et al.*, 2008; Le Gal *et al.*, 2009). There are 50 growers which supply sugarcane to the mill (see Table 1).

The growers and hauliers work in daylight hours (6am to 6pm), Monday to Saturday, whereas the mill works 24 hours per day, seven days per week. Only growers who bundle their cane into 9 tonne bundles before transporting it to the mill are considered in this model. (The growers which transport their cane loose and then spill it onto the infeed conveyer belt at the mill have been ignored in this version.) Growers which are less than 18 km away from the mill use trucks which are configured to transport three bundles per trip, whereas those which are further away are configured to transport four bundles per trip.

In the model, each of the 50 growers has to deliver 38 000 tonnes of cane during the 38-week milling season (*i.e.* they have to deliver a weekly allocation of 1 000 tonnes, or a daily allocation of 166.66 tonnes). The grower calculates how many trips are needed, and

Grower no.	No. of trips needed per week	No. of bundles per trip	No. of bundles to make (weekdays)	No. of bundles to make (Saturday)	Max. tonnes to deliver weekly
1 to 17	3	38	21	9	1 026
18 to 50	4	28	20	12	1 008

Table 2: *Weekly grower delivery targets.*

rounds up the number of bundles to suit his transport configuration (to avoid partially empty loads). Based on the number of bundles to be made, the grower ensures that the cane for the day's trips is bundled and ready for transport by 6am. Table 2 shows these details for the growers.

At the beginning of the simulation, each grower calculates the cycle time (time to make a round trip) to determine the number of trucks needed to transport the week's allocation. The cycle time (in minutes) is calculated by

$$\begin{aligned} Cycle\ time\ =& loading\ time + trip\ time\ to\ mill + weighbridge\ time + waiting\ time\ at\ mill \\ &+ unloading\ time + \ trip\ time\ to\ farm \end{aligned}$$

where *loading time* = 12 mins; *trip time to mill* = travel time at 40km/hr; *weighbridge time* = 2 mins; *waiting time at mill* = growers' anticipated wait at the mill (this value was varied for each simulation run); *unloading time* = 4 mins; *trip time to farm* = travel time at 45km/hr.

The actual waiting time at the mill depends on the number of trucks in the queue. The cycle time is used to determine how many trucks each grower needs to deliver the week's cane allocation.

The simulation model was initially run for a two week period with the grower distances to the mill as shown in Table 1, but with only one truck per grower. In this run, there was an ample supply of cane. Each grower had to deliver 1 000 tonnes of cane weekly. This output gave a baseline of the maximum tonnes which a grower can deliver with one truck. All growers underdelivered, and could not complete the required number of trips (see section 4 for the results of this run). The model was then changed to allow growers to have as many trucks as were needed to deliver the week's allocation. The grower also bundled the daily allocation of cane at 6am. The cycle time for each grower was then calculated. If a grower had more than one truck, the trucks would start to load cane and then leave the farm in a staggered fashion described by

$$Time\ for\ i^{th}\ truck\ to\ start\ loading = 6am + (cycle\ time) * \frac{(i-1)}{no.\ of\ trucks}$$

Thirteen two-week long runs were performed to analyse the effect of changing the *waiting time at mill* component of the cycle time formula. The *waiting time at mill* was incremented in steps of 30 mins from 0 mins to 360 mins. The results for the second of the two weeks were analysed, since the hauliers were still learning the average trip time in the first week.

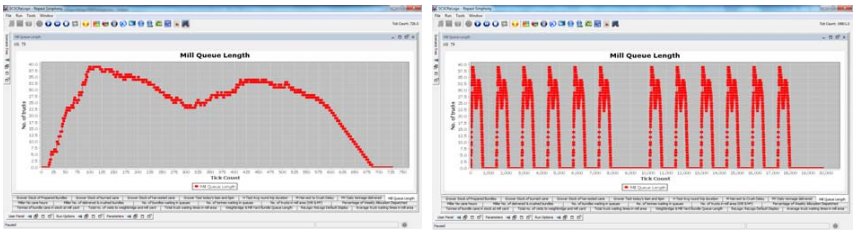


Figure 1: *Model output: queue length at the mill for Monday of the first week (left) and for two weeks (right). The growers deliver cane six days a week, Monday to Saturday.*

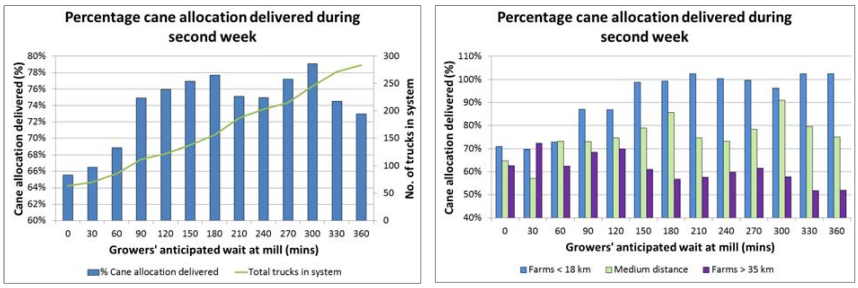


Figure 2: *Percentage cane allocation delivered during the week: for all the growers supplying the mill (left) and broken down by distance category (right).*

4 Results

The model first simulated the supply chain for two weeks, with ample cane supply. Each grower only had one truck with which to deliver the cane to the mill. None of the growers were able to supply the full 1 000 tonnes required for the week. The three growers within a 4 km radius of the mill were able to deliver 97.2% of their weekly allocation, while the growers between 47 and 58 km from the mill were only able to deliver 43.2% of their weekly allocation. On average only 63.9% of the week's cane allocation was delivered.

For this run, each day's cumulative deliveries was compared for weeks 1 and 2. The results were identical except for growers 35 and 36 (both 43 km from the mill). For these growers, the order of joining the queue made a difference to the number of trips they could make that day.

Figure 1 shows model output for this run for the queue at the mill for the first day (Monday) and for two weeks, starting on Monday. As can be seen, the queue length at the mill is very similar each day. Those familiar with truck queues at the mill have remarked that it is common to have a double hump in the mill queue, as shown here. The first truck arrives at 6:15. The first hump has its maximum between 7:39 and 8:00. The second hump is between 12:52 and 13:44. The last truck leaves the mill area at 17:25. The model was then changed so that growers would prepare enough cane for the day's trips at 6am, and each grower had as many trucks as he needed, based on the cycle time. The *waiting time at mill* component of the cycle time was varied from 0 to 360 mins in steps of 30 mins.

The results (Figure 2) show that even though more trucks were added, the growers were unable to deliver the full cane allocation for the week for any of the 13 anticipated waiting

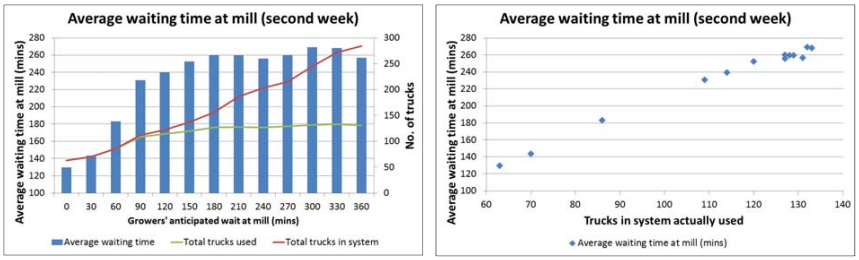


Figure 3: *Average waiting time at mill for each anticipated wait at mill run (left), and plotted by actual trucks used (right).*

times. At 300 minutes anticipated waiting time, the maximum of 79.1% of the total week's allocation was delivered. At 180 minutes, the next highest delivery rate of 77.7% was achieved. Figure 2 also shows the percentage allocation delivered for farms close to the mill (under 18 km), those far from the mill (greater than 35 km) and those farms lying in between. As the growers' anticipated waiting time increased, the farms close by and a medium distance away were better able to deliver.

Figure 3 (left) shows the average truck waiting time at the mill for each of the growers' anticipated waiting times. It shows that the growers were unable to anticipate the actual waiting time at the mill, since the other growers were also adding trucks to the system, causing the queues (and therefore the actual waiting time) to be longer. It was found that although the growers were adding trucks to the system, for anticipated waiting times of 90 minutes and longer, not all of them were being used. This is because the staggered leaving time formula used prevented some of the trucks from leaving earlier: when it was their time to leave, they found they could not get back to the farm by 6pm, so did not leave at all. When plotting the actual waiting time at the mill by the number of trucks actually used to deliver cane (Figure 3 right), the graph shows a linearly increasing trend.

5 Discussion

When growers used only one truck each, only 63.9% of the cane was delivered; using more than one truck increased growers' ability to deliver (they delivered between 66.6% and 79.1% of the week's allocation). Alternatives for the "truck leaving time" rule need to be investigated.

Le Gal *et al.* (2009, pg 171) report that with 195 trucks (considered "severely over-fleeted"), the wait at the mill can be 2 hours. In addition, between 2001 and 2006, the mill only managed to crush 76% of the allocation due to cane supply shortages and mill breakdowns. Our model shows that at 2 hours, the allocation delivered would also be 76%, but with fewer trucks (122 were created and 114 were used). This may be due to the fact that in our model, growers have exactly the same allocation, whereas in reality, their allocation would vary depending on the farm size and growth rate *etc.* Our model also does not take into account spiller cane deliveries. McDonald *et al.* (2008) also report that the waiting time in the mill yard is about 2 hours for 105 trucks. It should be noted

that these two studies simulated a more realisitc grower allocation distribution, and they also modelled both bundle and spiller cane deliveries.

The agents in this simulation (growers and hauliers) all followed the same rules of behaviour. Even so, the results are a useful indication of how the behaviours of individual autonomous agents (role players) can affect the system's behaviour.

6 Conclusions and future work

This model gives a starting point for being able to explore sugarcane supply chain complexities. For example, using a non-agent-based approach, it is difficult to determine at the individual level (grower) the effect of adding more trucks to the system (represented by the mill queue) because of the system feedback: adding more trucks increases the waiting time in the queue at the mill. This model has shown the benefit of the bottom-up approach to modelling the mill queuing time as an emergent system effect, rather than using it as an average model input.

Using the agent-based modelling approach opens opportunities for varying agents' behaviour, modelling groups of growers, changing weather conditions, *etc.* to investigate the system-wide effects. In future, the model will be used to investigate other truck leaving time rules. The model's functionality will also be extended to include spiller growers and expand the decision-making behaviour of the agents.

7 References

Beamon, B., 1998. Supply chain design and analysis: models and methods. *International Journal of Production Economics,* 553, pp. 281-294.

Bezuidenhout, C., Bodhanya, S. & Brenchley, L., 2012. An analysis of collaboration in a sugarcane production and processing supply chain. *British Food Journal,* 1146, pp. 880-895.

Größler, A. & Schieritz, N., 2005. Of stocks, flows, agents and rules - "strategic" simulations in supply chain research. In: H. Kotzab, S. Seuring, M. Müller & G. Reiner (Eds). *Research methodologies in supply chain management.* Heidelberg: Physica-Verlag, pp. 445-460.

Higgins, A. *et al.*, 2010. Challenges of operations research practice in agricultural value chains. *Journal of the Operational Research Society,* 616, pp. 964-973.

Higgins, A., Thorburn, P., Archer, A. & Jakku, E., 2007. Opportunities for value chain research in sugar industries. *Agricultural Systems,* 943, pp. 611-621.

Le Gal, P.-Y., Le Masson, J., Bezuidenhout, C. & Lagrange, L., 2009. Coupled modelling of sugarcane supply planning and logistics as a management tool. *Computers and Electronics in Agriculture,* 682, pp. 168-177.

Le Masson, J., 2007. *Articulation de modèles de planification logistique et d'approvisionnement d'une sucrerie: application à la mécanisation de la récolte de canne dans le bassin de Noodsberg (Afrique du Sud),* Masters thesis, Montpellier, France: AgroParisTech - Cirad.

McDonald, B., Dube, E. & Bezuidenhout, C., 2008. *Modelling and simulation for analysis of sugarcane transport systems.* Proceedings of the Second IASTED Africa Conference Modelling and Simulation (AfricaMS 2008), pp. 247-253.

North, M. *et al.*, 2013. Complex adaptive systems modeling with Repast Simphony. *Complex Adaptive Systems Modeling,* http://www.casmodeling.com/content/1/1/3 [accessed 15-5-2014].

North, M. & Macal, C., 2007. *Managing business complexity: discovering strategic solutions with agent-based modelling and simulation.* New York: Oxford University Press, Inc.

Owen, M., Albores, P., Greasley, A. & Love, D., 2010. *Simulation in the supply chain context: matching the simulation tool to the problem.* Proceedings of the 2010 Operational Research Society Simulation Workshop (SW10), pp. 229-242.

Siebers, P. *et al.,* 2010. Discrete-event simulation is dead, long live agent-based simulation!. *Journal of Simulation,* 43, pp. 204-210.

Sterman, J., 1989. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science,* 353, pp. 321-339.

Thorburn, P. *et al.*, 2005. *Integrated value chain scenarios for enhanced mill region profitability. Final report to the Sugar Research and Development Corporation on SDRC Project CSE010.,* Sugar Research and Development Corporation, Australian Government.

8 Acknowledgements

CSP and DM would like to thank CAIR for the funding which contributed to this research; CNB would like to thank SASRI for research funding. CSP would also like to thank Drs J. Sam, J.-C. Chappelier and V. Lepetit of École Polytechnique Fédérale de Lausanne for their two MOOCs on Java programming, offered via Coursera, which helped to lay the foundation for the coding of the simulation model. The authors thank the reviewers for their helpful comments.