– Proceedings –

44th Annual Conference of the Operations Research Society of South Africa

13–16 September 2015
Pecan Manor, Hartbeespoort, South Africa

ISBN: 978-1-86822-666-5

Editorial

I am pleased to present to you the proceedings of the 44th Annual Conference of the Operations Research Society of South Africa (ORSSA). The proceedings contain a collection of carefully selected technical research papers of operations researchers and cover a wide and interesting spectrum of projects. 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 submitted papers, thirteen were ultimately, after consideration and incorporation of reviewer comments, judged to be suitable for inclusion in the proceedings - an acceptance rate of 72%. The proceedings will also be published at:

http://www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

As in the past, the proceedings are not something that was produced in isolation. I would like to thank the authors for submitting their work to our conference - all submissions were of an outstanding quality. Papers cannot be selected without reviewers and I would also like to thank the reviewers who gave generously of their valuable time and expertise to assist with the important task of reviewing papers. The editorial team did a great job in producing the final proceedings - many thanks to Hans Ittmann, Fanie Terblanche and Danie Lotter who provide professional help and guidance during the whole process of reviewing and producing the final proceedings. If I may, I would like to single out Fanie Terblanche who was once again a tremendous help and who played a key role during the whole process. 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 sincerely hope that everyone will find the conference proceedings to be enriching.

Reviewers

The editorial would like to thank the following reviewers

R Bennetto	OPSI Systems, South Africa
I Campbell	University of the Witwatersrand, South Africa
S Das	Council for Scientific and Industrial Research, South Africa
A De Villiers	wiGroup, South Africa
I Durbach	University of Cape Town, South Africa
J Du Toit	wiGroup, South Africa
T Du Toit	North-West University, South Africa
D Evans	Private Consultant, South Africa
P Fatti	University of the Witwatersrand, South Africa
M Fisher	Northwestern University, United States
H Ittmann	University of Johannesburg, South Africa
M Kidd	Technical University of Denmark, Denmark
R Koen	Council for Scientific and Industrial Research, South Africa
H Kruger	North-West University, South Africa
H Nel	Stellenbosch University, South Africa
W Pelser	Armscor, 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
J Van Vuuren	Stellenbosch University, South Africa
S Visagie	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 2015
Operations Research Society of South Africa

Table of contents

Proceedings of the 2015 ORSSA Annual Conference
pp. 1–11
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

An agent-based simulation model describing the *lek* mating process of *Eldana saccharina* Walker

BJ van Vuuren* L Potgieter† JH van Vuuren‡

Abstract

Eldana saccharina Walker (Lepidoptera: Pyralidae) is a serious stalk borer pest which continues to plague the sugar producing industry in South Africa. Various control methods have been proposed to suppress the pest and decrease its detrimental impact on the industry. These solution methods are, however, often difficult and costly to test, implement and develop. In an attempt to better understand the behaviour and population dynamics of *E. saccharina*, an agent-based simulation model is currently being developed. The model simulates the stalk borer's biology, feeding habits, mating behaviour, and dispersal patterns along with the natural variation in its habitat. It is envisaged that the model may facilitate the development and testing of certain pest control strategies prior to in-field implementation. *E. saccharina*'s complex mating process requires particularly careful consideration and structural implementation in the model as it plays a primary role in the continued prevalence of the pest. As part of the above-mentioned simulation development, a framework for a novel, agent-based simulation submodel of the mating process of *E. saccharina* is presented in this paper. Output from the model is compared with existing literature on *E. saccharina*, thereby establishing the degree to which the model replicates the pest's mating cycle as it has been observed in real life.

Keywords: *Eldana saccharina* Walker, Sugarcane pest infestation, Agent-based simulation, *Lek* mating, Integrated pest management.

1 Introduction

In order to develop an effective *integrated pest management* (IPM) system to assist in diminishing the infestation levels of *E. saccharina* in sugarcane, several important behavioural aspects of the pest must be considered as opportunities where pest control

*Stellenbosch Unit for Operations Research and Engineering, 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)**, Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

measures may affect the propagation of the stalk borer in its natural environment. One of these aspects is *E. saccharina*'s complex mating process. A biologically accurate, representative mating sequence of the pest is required for implementation in the agent-based model proposed by van Vuuren *et al.* [9] for the purposes of testing and developing strategies to control the pest at this point in its life cycle. A review of what is known about *E. saccharina*'s mating process, as documented in the literature, is presented in §2, whereafter the manner in which this mating process may be implemented in the simulation model mentioned above, along with relevant assumptions made, is discussed. A general overview of the modelling approach followed to implement the mating process is given in §3.1, and this is followed by a detailed description of the AnyLogic mating process model implementation in §3.2. The paper then closes in §3.3 with a verification of the model against the documented mating process and model assumptions in order to determine the accuracy with which reconstruction has been achieved, and some ideas with respect to possible future work in §4.

2 Biological review of the mating process of *E. saccharina*

The complex mating process followed by *E. saccharina* has been documented by Atkinson [1]. Mating of *E. saccharina* occurs after sunset and is weather-dependent. Males typically emerge from pupae slightly before females and climb up the stalk of the plant to its canopy. Rain and even light winds inhibit mating; moths have been witnessed to wait beneath leaves in the canopy until the conditions become favourable for mating. Approximately 40 minutes after emergence, the male begins to display. During this time, the male also releases an attracting pheromone into the air.

The aim of the display is to attract females for mating purposes. Males remain static for the duration of the display which, in the absence of females, may continue until dawn. Although males can display singly, they far more frequently display in groups of 3–6 males (called *leks*) on the same plant. Mating usually occurs on the first night of emergence, after which females begin oviposition — typically for the remainder of their lives.

During the mating process, a female is attracted to a *lek* of males by their display and the combined scent of the pheromone released. Once a female has located the *lek*, one male is selected to mate with, after which the pair are typically not disturbed by other moths. It is generally understood that the most dominant male (that is, the male with the strongest pheromone or display) is selected by the female [3]. Almost all females mate only once during their lifetime owing to the large size and durability of the spermatophore, as well as the relatively short lifespan of the adult moth [2]. Walton [10] has, however, reported that, under laboratory conditions, females have been noted to mate up to three times. In the same study, males were shown to have the ability to mate up to six times, but more often seem to do so approximately three times over their lifespan. The mating process lasts up to three hours [1] and, about 24 hours after mating, a female will begin oviposition [2].

3 Simulating *E. saccharina*'s mating process

An agent-based model of *E. saccharina* has been designed and developed in the ANYLOGIC UNIVERSITY 7.2.1 software suite. A framework encapsulating the most notable biological attributes of the pest which must be included in a simulation capable of mimicing its behaviour and distribution accurately has previously been presented and discussed in detail by van Vuuren *et al.* [9]. Four fundamental building blocks are included in the model. These are the simulation environment, the life cycle of *E. saccharina* and associated influence of temperature on the pest, the complex mating process of *E. saccharina* and, finally, the in-field spatial distribution of the stalk borer within its habitat. The temporal interaction between these building blocks is described in detail in [9].

In all previous models of *E. saccharina's* population growth and dispersal, an aggregation approach was followed which excluded small-scale intricacies such as the *lek* mating and mate selection processes [4, 5, 8]. Furthermore, in these aforementioned models, the growth and spread rate of the pest, as well as the direction in which the population migrates, is required to be specified during model construction. Our agent-based model, however, facilitates simulation of a mating process which accurately mimics the in-field observations made by Atkinson [1] and Carnegie [2], allowing individual agents to select mating sites and, in turn, naturally facilitate the spread of the population. Furthermore, simulating the competitive nature of the mate selection process incorporates the possibility of stronger offspring which may migrate over longer distances, as well as the prevalence of unmated females which are more likely to migrate if they remain unmated by the end of their first or second night after emergence [1]. From an IPM perspective, the *sterile insect technique (SIT)* is a topical control measure currently under investigation at the *South African Sugarcane Research Institute (SASRI)* [7, 8, 10]. SIT has a direct impact on the competitivity of males and females, which may change their behaviour during the *lek* mating process. The rate at which sterility spreads within a population is directly affected by the selective, competitive mating process within the *lek*. Including the *lek* mating procedure in the simulation model described in [9] is therefore important to accurately describe the imposition of SIT on an *E. saccharina* population.

3.1 General description of implemented process

The logical progression followed and assumptions made to effectively implement *E. saccharina*'s mating process in the simulation model are summarised in Figure 1.

Once a male moth has reached adulthood and the simulated time of day is appropriate for mating activity to commence, it attempts to create a *lek* together with males in its immediate vicinity. This time window is assumed to be 18h00–23h00 [1]. If there are 3–6 males in close enough proximity, a *lek* will be formed and the mating process will progress to the next stage[1]. If a male is drawn into joining a *lek* which was initiated by another male nearby, it looses the ability to initiate its 'own' *lek* until such time as the *lek* to which it belongs ceases to exist. This may occur due to a lack of sufficient males nearby, meaning that the *lek* never actually materialises, or it may materialise, but then dissipate once a female has chosen and approached her preferred partner from the *lek*.

[1] A realistic level of proximity will be determined during model calibration.

When a *lek* is successfully initiated and a group of males become members of it, its location is assumed to be specified as the centroid of the males included in the *lek*. If n males with Cartesian coordinate positions $(x_1, y_1), \ldots, (x_n, y_n)$ participate in the *lek* (when viewed from above), this point is (L_x, L_y), where $L_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $L_y = \frac{1}{n}\sum_{i=1}^{n} y_i$. Once the *lek* has been located at this point, it becomes discoverable to females which are ready to mate and may potentially be attracted to it.

In a full-scale simulation, several *leks* are likely to materialise in the vicinity of a female at a given point in time. It is assumed that the female is aware of and attracted to all *leks* in her vicinity, but the *lek* which has the strongest attraction, based on the total pheromone strength of all the males in the *lek*, as well as its distance from her, has the highest probability of being selected.

It is possible that two or more females in the simulation may be attracted to and begin to approach the same *lek*. In accordance with the standard assumptions of a Poisson arrival process, no two females will arrive at a *lek* at precisely the same instant during the simulation. When a female arrives at a *lek*, the most competitive male has the highest

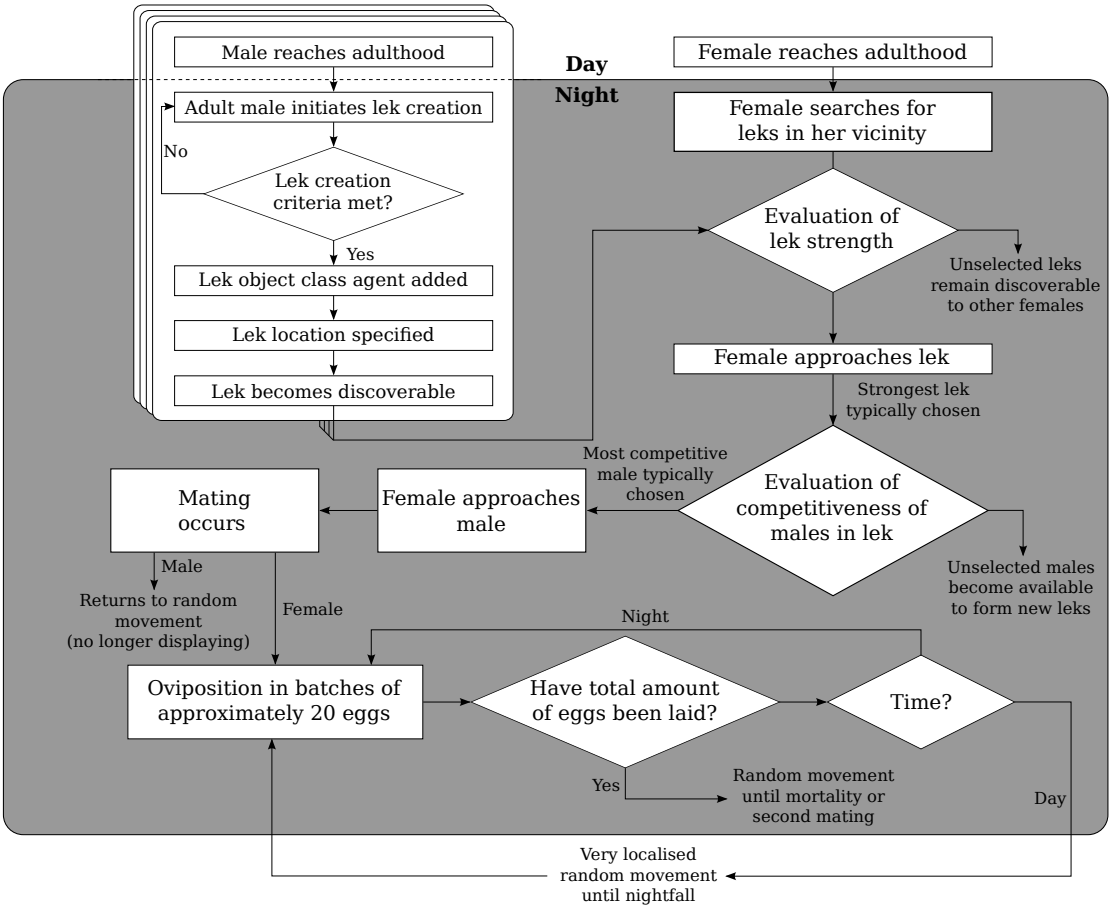


Figure 1: Flowchart of the general mating process of *E. saccharina*.

probability of being chosen for mating[2]. The female occupies the chosen male for the entire mating process and, as a result, the male no longer participates in the *lek* [3]. A weaker *lek* is formed nearby by the remaining members of the original *lek*, provided that at least three unmated males remain. As a result of its close proximity, other females who also approached the original *lek* now approach this weaker *lek* in order to select a mate. This may occur several times until fewer than three males remain as participants of the remaining *lek*, at which point it dissipates — no longer attracting females [3, 6].

Mating ensues for 2–3 hours. Due to the limited time available per night for mating, it is assumed that males will only mate once per night. After mating, the female departs in order to prepare for oviposition. Consistent with assumptions made in the literature, the female only oviposits batches of eggs during the early hours of the nights post the 24 hour gestation period, and, during the day remains relatively static, moving randomly on a very small scale (resembling walking), until such time as she can oviposit again. Once the female has laid all of her eggs, she typically perishes — although the model does incorporate a small probability of some females mating a second or even third time.

3.2 Description of implemented process in AnyLogic environment

In order to translate the description in §3.1 into executable code in the AnyLogic simulation environment, several components were employed to form the basic structure of the simulation model, as shown in Table 1.

Name of component	**Use in simulation**	AnyLogic **icon**
Object class	Represents a separate agent who possesses its own internal structure governing its actions and decisions	
Parameter	Represents agent characteristics and behaviour, and only changes when the behaviour of the agent is changed	
Variable	Stores results of model simulation or model object characteristics, changing over time	
Function	Executes a portion of code or returns the value of an expression each time it is called in the simulation	
Link	Creates links between agents along which messages can be sent for direct control and relationships between linked agents	

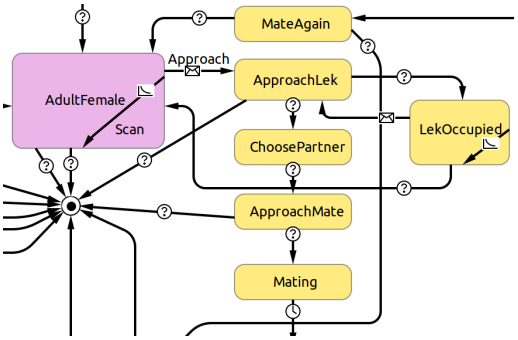
Table 1: Different components employed in the simulation model.

In the simulation, male and female object classes exist to facilitate separate behaviour control of both sexes of *E. saccharina*. An additional, 'theoretical' object class, called Leks, was also incorporated to assist in the creation and control of *leks* in the simulation model. As a result of using continuous space in the simulation, it is not possible for females to search in defined, discrete locations in an attempt to identify *leks* for mating. Furthermore, for the purposes of *lek* selection by females, *leks* are required to possess specific attributes, such as a pheromone strength and positional coordinates. In the case of displaying males, an entity is required to which they can 'belong' so as to ensure that they do not form part of multiple *leks* at any given time within the simulation,

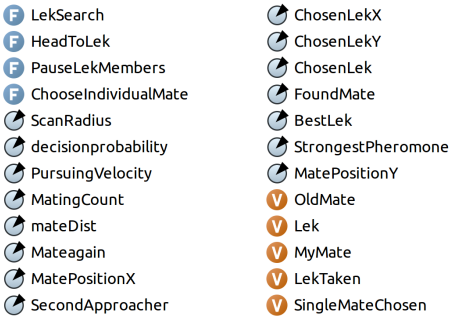
[2]This probability will be determined during model calibration.

which would result in a misrepresentation of the number of females which can feasibly mate per night. In light of these requirements, defining Leks as a separate object class accommodates superior control and efficiency within the model. The individual Leks objects will, however, be hidden in the final simulation and only contribute to control on the back-end of the simulation model.

Upon reaching adulthood, individual males are responsible for initiating *leks* during night-time displays. In the simulation, only males which do not currently belong to *leks* possess the ability to initiate *leks*. Importantly, the male agents are responsible for the initial addition of an agent to the Leks object class only; thereafter, they do not operate or control the *lek* behaviour in any way. A *lek* is controlled by the inner working of its own statechart, since it is a member of an independent object class.



(a) Screenshot of mating control division female statechart.



(b) The AnyLogic tools used for implementing the mating sequence.

Figure 2: Inner working of the female behaviour control.

As may be seen in Figure 2(a), the female performs a scan function upon reaching the AdultFemale state. This scan is performed, on average, five times per simulated hour. This discrete number of search instances was chosen to reduce computational expense when there are a large number of females present in the simulation. This scan frequency was deemed appropriate since the random movement by male agents is small-scale and, as a result, their arrangement (and, therefore, the possibility of new *leks* being formed) does not change significantly over small time intervals. The transition triggers execution of a function, as seen in Figure 2(b), called LekSearch. The logic sequence of this function begins by first determining whether it is an appropriate time for mating to ensue. If so, a test is performed in respect of each *lek* to determine whether enough males are part of the *lek*, as well as whether or not the *lek* is sufficiently close to the female. If both of these conditions are met, the pheromone strength of the *lek* is scaled as a function of its distance from the female[3]. A test is then performed to determine whether this scaled value is the strongest *lek* which has been found so far. If it is the currently strongest *lek* found and the probability threshold is met, its coordinates are saved, and the female disengages from the previously chosen *lek* and engages with the new *lek* via the LekLink connection.

[3]The scaling of the *lek* strength value accounts for dissipation of the pheromone strength with distance from the female. This is caused by physical barriers (such as plants), air resistance and wind between the female and the *lek*, and may result in selection of a weaker *lek* by the female which is closer to her, instead of a stronger *lek* further away [3].

If a *lek* is successfully found by the female, the function concludes by triggering the Approach transition to move the female to the ApproachLek state. If not, the female remains in the AdultFemale state, moving randomly on a small scale until the Scan transition is triggered again to retest for *leks* in her vicinity. In the ApproachLek state, the HeadToLek function is called which sets the flight speed of the female and instructs her to move from her current location to the *lek* chosen in the previous step.

In the event that more than one female approaches the same *lek*, the first arriving female will 'occupy' the *lek* and transition to the ChoosePartner state. Since the Scan function is called at discrete time intervals, females will detect *leks* at different times and be different distances from these *leks*. In light of this, the first female that detects the *lek* will not necessarily arrive first. When the first female 'occupies' the *lek*, the connection between all other approaching females and that particular *lek* is broken, and these females transition to the LekOccupied state. In this state, a female continues on her original flight path to arrive at the location of the *lek* originally identified. Here the LekSearch function is called in rapid succession (within a smaller search radius) to force the female to identify the new *lek* which will materialise from the remaining males[4]. The female will then approach this 'new' *lek* by returning to the ApproachLek state as before. A control parameter, called SecondApproacher, is incorporated to inform the female in the event that an insufficient number of males remain from the original *lek* she approached. This triggers a transition back from the LekOccupied state to the original AdultFemale state. When a female is first to arrive at a *lek* and 'occupies' it, a conditional test is performed to ensure that her coordinates match those of the *lek*. If the match is successful, she transitions to the ChoosePartner state.

During the simulation, adult male moths move randomly within their local area. This random movement means that they have the ability to leave or join *leks* during the mating time window. Once part of a *lek*, males are assumed to be primarily static and the random movement in the simulation subsides. When a *lek* dissipates, the males return to a state of small-scale random movement and, during the appropriate time window, attempt to initiate or join new *leks*.

In the ChoosePartner state, two functions are called consecutively, namely ConnecttoLekMembers, followed by ChooseIndividualMate. ConnecttoLekMembers is responsible for connecting the female to the group of static males which compose the selected *lek*. The logic sequence of the function begins by searching through all male agents in the simulation and, for each male, determining whether it is in the adult state and ascertaining that it belongs to a *lek*. If these two conditions are met, the female then tests whether or not the male belongs to the chosen *lek*. If the male is found to form part of her chosen *lek*, it is associated with the female via the MembersofLek connection.

Once all the males comprising the chosen *lek* have been associated with the female, ChooseIndividualMate is called to determine the most competitive male in the *lek* for mating purposes. This function follows a similar logic to the LekSearch function but, in this case, it searches through all male agents in the simulation, determining

[4]This 'new' *lek* functions fundamentally as the same *lek* that was originally approached. The implementation allows for the necessary amount of control over the remaining agents and excludes the male chosen for mating from the previous *lek*.

whether they are in the appropriate life stage for mating and, if so, testing whether or not they belong to the *lek* chosen by the female. If these conditions are met, the strength of each male's pheromone is evaluated and compared to that of the other males comprising the *lek*. If the pheromone strength of the male is stronger than the previously strongest pheromone encountered by the female and a probability threshold is exceeded, she breaks the connection from her previous best partner, connects to this male via the Partner connection and saves the partner's positional coordinates.

Once the strengths of all of the males comprising the *lek* have been evaluated, the female approaches her chosen mate and disengages from all the other males forming part of the *lek* by breaking the MembersofLek connection. This allows them to form a new *lek*, or return to small-scale random movement. The female then approaches the chosen male and mating ensues for a randomly sampled time period of between two and three hours before proceeding to the gestation, oviposition and consequent dispersal division of the simulation.

3.3 Model functionality assessment

A comparison was performed to determine the accuracy with which the important mating aspects described in §2, as well as the assumptions mentioned in §3.1, were incorporated into and implemented in the model. The sequence of a typical mating instance in the simulation is illustrated by a collection of images in Figure 3. In the figure, male agents are represented as blue squares, whilst female agents are coloured pink[5]. The pheromone strength of a male moth is shown as a number attached to the corresponding agent (these are simply representative values and not attributed any specific unit). The centroid location of *leks* which are currently discoverable by females in the simulation are shown as black squares and the 'strength' of the *lek* by virtue of its composing agents is shown as a bold sum value above the *lek* element. The female's search radius boundary is included as a wide-spaced dotted black arc and the LekLink, MembersofLek and Partner links are shown as a red dashed line, a fine-spaced black dotted line and a solid black line, respectively. The females in this test run are instructed always to choose the strongest male.

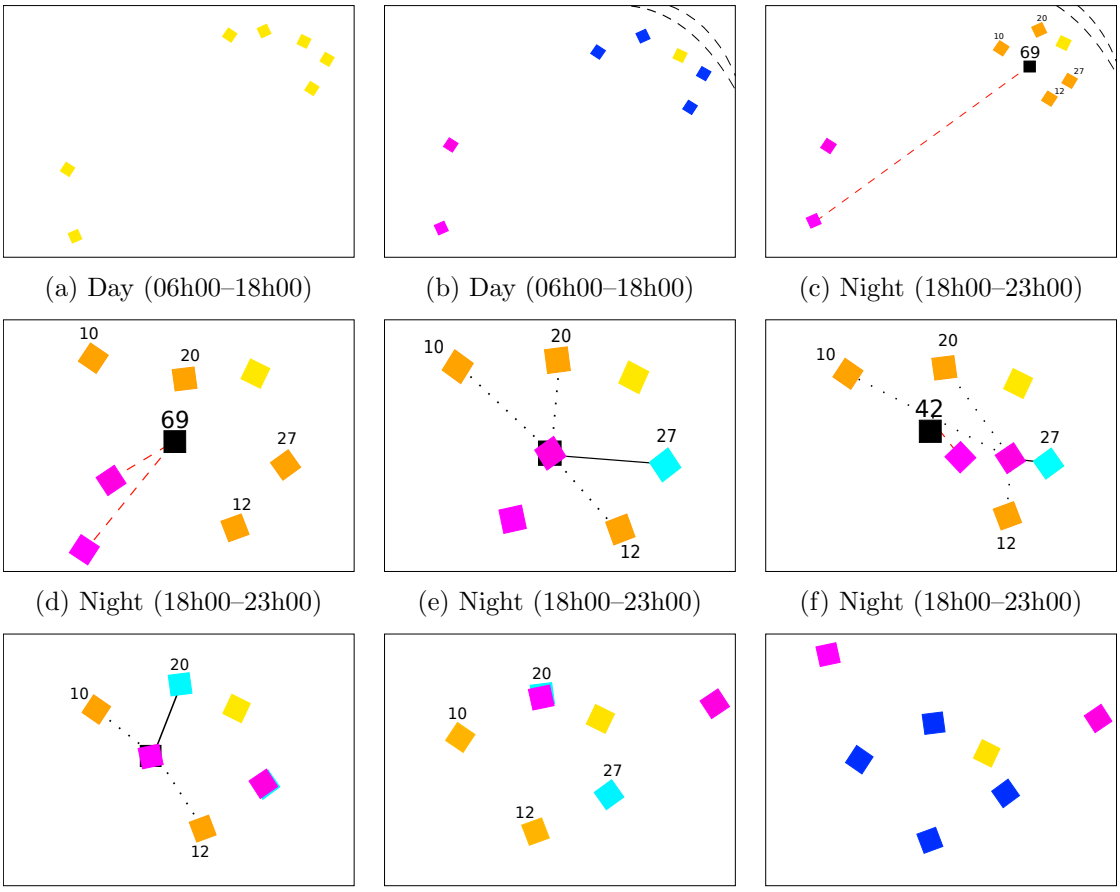
Figure 3 can be interpreted as follows:

(a) All of the agents are in the pupae stage and, as a result, remain static in their respective locations.

(b) Four of the males and both females have reached adulthood and begin moving around, but no mating occurs since it is still daytime. The boundaries of the females' search radii may be seen in the top right-hand corner of the figure.

(c) Night falls and the four mature males begin displaying in close enough proximity to form a *lek*. Following the process described in §2, one of the females has already discovered this *lek*.

[5]For illustrative purposes, male agents change colour depending on their respective states to facilitate observation of their associations during model execution. During the displaying stage, males are coloured orange and the male selected from the *lek* as a mate by the female turns light blue. These colour changes will not be included in the final model.



(a) Day (06h00–18h00) (b) Day (06h00–18h00) (c) Night (18h00–23h00)

(d) Night (18h00–23h00) (e) Night (18h00–23h00) (f) Night (18h00–23h00)

(g) Night (18h00–23h00) (h) Night (18h00–23h00) (i) Night (23h00–06h00)

Figure 3: A mating sequence executed in AnyLogic simulation.

(d) Both females have discovered the *lek* and are heading towards it. It is evident that the female who was closer to the *lek* initially will arrive first.

(e) As in the process described by Atkinson [1], the first female to arrive at the *lek* has occupied it, identified the most competitive male in the *lek* and connected to him as her mate. The male has a pheromone strength of 27. The second female has now moved to the LekOccupied state and is searching for a *lek* which may result from the remaining males of the original *lek*.

(f) Now that the first female is approaching her mate, she has 'released' the other males once forming part of the *lek* she occupied. Owing to their close proximity, a second *lek* immediately materialises nearby. This functionality was verified by Conlong [3]. The second *lek* has an inferior strength to that of the original *lek* as one male is no longer contributing to the display.

(g) The first female is in the process of mating and, in the meantime, the 'new' *lek* location is being approached by the second female. She will then choose the most competitive remaining male and engage it as her partner.

(h) The first female has completed the mating process and has now moved to the 24 hour gestation period before she begins oviposition. The second female is in the process of mating at this point. As can be seen, no further *lek* has formed since there remain insufficient displaying, unmated males to contribute towards continuing the *lek*.

(i) The time window for mating has expired and the males are no longer displaying. All agents return to random movement and females do not lay eggs until the following evening, in the meantime moving only very small distances.

4 Conclusion and further work

The novel manner in which the complex mating process of *E. saccharina* has been implemented within an agent-based simulation model of the pest's population dynamics was described in this paper. The solution made use of an independent object class in order to control and simulate the formation of *leks* between male agents which precede mating. The modelling approach and underlying assumptions were tested and found to mimic the mating process of *E. saccharina* as detailed in the literature by Atkinson [1] and Carnegie [2].

The portion of the simulation model described here forms part of a larger on-going project at Stellenbosch University aimed at building a comprehensive agent-based model describing the most important behavioural aspects of *E. saccharina* in the presence of sugarcane. The model will eventually facilitate the design and testing of pest control strategies aimed at minimising the damage caused by the stalk borer pest. All aspects of the simulation (including the mating process) require comprehensive calibration and validation in terms of parameter and rate selection for biological processes and activities modelled in the simulation. This will be done using data from the literature, expert opinion and video footage of the pest interaction under laboratory conditions once all aspects of the model have been implemented.

References

[1] Atkinson PR, 1981, *Mating behaviour and activity patterns of Eldana saccharina Walker (Lepidoptera: Pyralidae)*, Journal of the Entomological Society of Southern Africa, **44(2)**, pp. 265–280.

[2] Carnegie AJM, 1974, *A recrudescence of the borer Eldana saccharina Walker (Lepidoptera: Pyralididae)*, Proceedings of the South African Sugar Technologists Association, **48**, pp. 107–110.

[3] Conlong DE, 2014–2015, Senior Entomologist at *the South African Sugarcane Research Institute, Mount Edgecombe*, [Personal Communication], Contactable at `Des.Conlong@sugar.org.za`.

[4] Hearne JW, Van Coller LM & Conlong DE, 1994, *Determining strategies for the biological control of a sugarcane stalk borer*, Ecological Modelling, **73(1)**, pp. 117–133.

[5] 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(2)**, pp. 271–286.

[6] Mudavanhu P, 2015, Weed Biocontrol Researcher at *ARC–PPRI, Vredenburg Campus, Stellenbosch*, [Personal Communication], Contactable at `Mudavanhup@arc.agric.za`.

[7] Mudavanhu P, Conlong DE & Addison P, 2012, *Impact of mass-rearing and gamma radiation on the thermal tolerance of Eldana saccharina Walker (Lepidoptera: Pyralidae)*, Proceedings of the South African Sugar Technologists Association, pp. 139–143.

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

[9] Van Vuuren BJ, Potgieter L & Van Vuuren JH, 2014, *Prerequisites for the design of an agent-based model for simulating the population dynamics of Eldana saccharina Walker*, Proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa, pp. 62–70.

[10] Walton AJ, 2011, *Radiation biology of Eldana saccharina Walker (Lepidoptera: Pyralidae)*, MSc Thesis, Stellenbosch University, Stellenbosch.

Proceedings of the 2015 ORSSA Annual Conference
pp. 12–22
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Application of artificial neural networks for predicting core parameters for the SAFARI-1 nuclear research reactor

EB Schlünz*†‡ PM Bokov* JH van Vuuren§

Abstract

The *in-core fuel management optimisation* (ICFMO) problem is a nonlinear assignment problem in which an optimal fuel reload configuration for a nuclear reactor core is sought. Function evaluations for the problem are performed by a reactor core calculation code and are deemed computationally expensive. This computational cost severely hinders efforts toward the investigation of appropriate techniques for solving the ICFMO problem. It is possible, however, to reduce this cost by replacing the reactor code with a computationally cheaper surrogate model. In this paper, *artificial neural network* (ANN) surrogate models are constructed for the prediction of core parameters for the SAFARI-1 nuclear research reactor. The parameters correspond to possible ICFMO objectives and constraints. The Neural Network Toolbox within the Matlab software suite is utilised for the construction. A description of the neural networks is given, along with details regarding the training process. The accuracy of the neural networks is verified on different sets of data points by comparing the predicted values to their actual values. The results indicate that the ANNs may be employed to significantly reduce the computational time required within an investigation of appropriate techniques for solving SAFARI-1 ICFMO problems. Furthermore, possibilities exist for using neural networks (provided that sufficient generalisation can be attained) in areas other than ICFMO *e.g.* within reactor codes.

Key words: Artificial neural network, In-core fuel management optimisation, Nuclear reactor.

1 Introduction

Nuclear reactors typically undergo a fuel reloading process at regular intervals (*e.g.* between operational cycles) in order to replenish their fuel. As part of this process, the

*Radiation and Reactor Theory, The South African Nuclear Energy Corporation SOC Ltd (Necsa), Building 1900, PO Box 582, Pretoria, 0001, South Africa, emails: bernard.schlunz@necsa.co.za and pavel.bokov@necsa.co.za

†Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

‡Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

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

loading configuration of fuel assemblies in the reactor core may be changed in order to satisfy prescribed operational and safety requirements. The problem, then, of finding an optimal fuel reload configuration for a nuclear reactor core is called the *in-core fuel management optimisation* (ICFMO) problem.

As part of our current research, we are investigating the appropriateness of several multiobjective metaheuristic techniques for solving multiobjective instances of the ICFMO problem for the SAFARI-1 nuclear research reactor at Pelindaba, South Africa. In order to evaluate the suitability of any configuration in terms of its objective and constraint function values, a reactor core calculation code is usually employed. Many codes utilise a deterministic approach for this purpose in which the evaluation of a reload configuration involves the numerical solution of a complicated partial differential equation (or an approximation thereof). Therefore, the function evaluations of the ICFMO problem are deemed computationally expensive. The computational cost thus introduced by a reactor code severely hinders efforts toward designing appropriate techniques for solving the ICFMO problem due to the several thousands of reload configuration evaluations typically required.

It is possible, however, to reduce the computational cost by replacing the reactor code with a computationally cheaper surrogate model. As demonstrated in the literature [8, 11], an *artificial neural network* (ANN) surrogate model can predict reactor core parameters with sufficient accuracy at only a fraction of the computation time. ANNs have also successfully been used in conjunction with a variety of solution techniques for solving ICFMO problems [3, 9]. In order to aid our investigation, ANN surrogate models are constructed in this paper for the prediction of SAFARI-1 core parameters corresponding to possible ICFMO objectives and constraints. The Neural Network Toolbox [1] within the Matlab software suite [14] is utilised for this construction.

The paper is organised as follows. Section 2 contains a general introduction to ANNs while Section 3 contains a description of the SAFARI-1 reactor. In Section 4, we present the steps that were followed in constructing the neural networks for predicting SAFARI-1 core parameters. Results pertaining to the training and application of these ANNs are presented in Sections 5 and 6. The conclusions of the paper are presented in Section 7.

2 Artificial neural networks

ANNs can be used for various applications, such as classification, clustering, function approximation and optimisation [10]. According to Fausett [4], *an artificial neural network is an information-processing system that has certain performance characteristics in common with biological neural networks.*

As an information-processing system, a neural network consists of several simple processing elements called *neurons*. The neurons are connected to one another by means of directed communication links, and the pattern (or topology) of these connections is called the *architecture* of the network. Each neuron in an ANN is able to receive input signals over the links, to process these signals by means of an associated *activation function*, and to send an output signal over the links to other neurons. Furthermore, each communication

link is associated with a *weight*, which, in a typical neural network, scales the signal being sent. The weights of a neural network are determined according to some specific method, called the *training algorithm*.

2.1 Multilayer feedforward neural networks

One of the most popular types of ANNs in use is the class of *multilayer feedforward* neural networks [13]. These networks consist of neurons that are partitioned into subsets, called *layers*. Neurons within a particular layer are only connected to neurons in the next layer, and signals are therefore sent in a forward direction over the network. Neurons in the *input layer* typically perform no computations and simply transmit external input signals onwards. The last layer of the network is called the *output layer*, while all layers in between are called *hidden layers*. An illustration of a multilayer feedforward neural network with one hidden layer is presented in Figure 1.

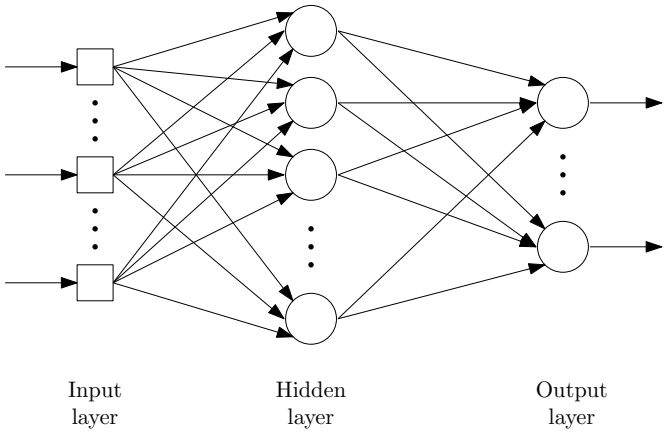


Figure 1: *A multilayer feedforward neural network with one hidden layer.*

The popularity of these networks is due to their power of being universal approximators, *i.e.* multilayer feedforward network architectures can approximate virtually any function of interest to an arbitrary degree of accuracy, given that a sufficient number of hidden neurons are available [7].

2.2 Training and other considerations

Multilayer feedforward neural networks for function approximation follow a training process called *supervised training*. According to this process, training is performed by presenting the network with a set of known input-output pairs. A training algorithm then adjusts the weights of the network in such a way that the predicted and known outputs are close to one another. Each iteration of adjusting the weights of the network using the complete training set is called an *epoch*.

The method known as *backpropagation of errors* is very popular for training these networks, and is essentially a gradient descent method for mimimising the total or mean squared error of the calculated output of the network [4]. An activation function for a

backpropagation neural network should be continuous, differentiable, and monotonically non-decreasing [4]. The typical choice of an activation function is a sigmoid function, which satisfies these characteristics.

A neural network should have the ability of responding well to the training data, as well as the ability of responding reasonably well to new (unseen) input data that are similar to, but different from, the training data. This ability to make good predictions for new input data is known as *generalisation*. If a neural network model is too simple or too complicated, then the network will achieve poor generalisation, analogous to polynomial curve fitting in which a polynomial with too small or too large a degree will yield poor predictions for new data points [2]. Several techniques may be applied to obtain good generalisation for a network, one of which is called *regularisation* [1, 2]. With regularisation, the complexity of the network model is controlled by adding a penalty term to the network's original error performance measure. Another important factor related to generalisation involves the training data used — the training points must be a representative subset of all the data points to which one wishes to generalise [13].

Critical aspects that arise during the construction of ANNs include the required number of hidden layers, as well as the required number of neurons in each layer. Given the universal approximator property of multilayer feedforward neural networks, one hidden layer should be sufficient for almost all function approximation applications [4]. The required number of neurons depends on the number of training samples available, the amount of noise therein, the complexity of the function to be approximated, and the method used to obtain good generalisation [13]. Essentially, the required number of neurons has to be determined empirically. It is, however, worth noting that networks of smaller size are preferred over larger networks [10]. A large network may be able to memorise the training data and thus potentially exhibit poor generalisation.

3 The SAFARI-1 nuclear research reactor

As stated in Section 1, our study relates to the multiobjective ICFMO problem for the SAFARI-1 nuclear research reactor, operated by the South African Nuclear Energy Corporation SOC Ltd. The reactor is utilised for nuclear materials research and commercial activities, such as irradiation services for isotope production and silicon transmutation doping. The core layout of SAFARI-1 is presented in Figure 2. The SAFARI-1 core consists of a 9×8 lattice which houses twenty-six fuel assemblies, six control rods, seven dedicated *molybdenum-99* (^{99}Mo) production rig facilities, two general *isotope production rig* (IPR) facilities, as well as other core components which we do not specify in detail. An ex-core facility is utilised for the silicon doping.

The OSCAR (*Overall System for the CAlculation of Reactors*) code is used as the primary reactor core calculation code for performing SAFARI-1 reload evaluation simulations. In general, ICFMO objectives and constraints are translated into core parameters returned by a reactor code *e.g.* OSCAR-4. The aim in this paper is to construct neural network surrogate models for the prediction of SAFARI-1 core parameters corresponding to possible ICFMO objectives and constraints. The computational cost incurred by using OSCAR-4 for reload evaluations may then be reduced by using the neural networks instead.

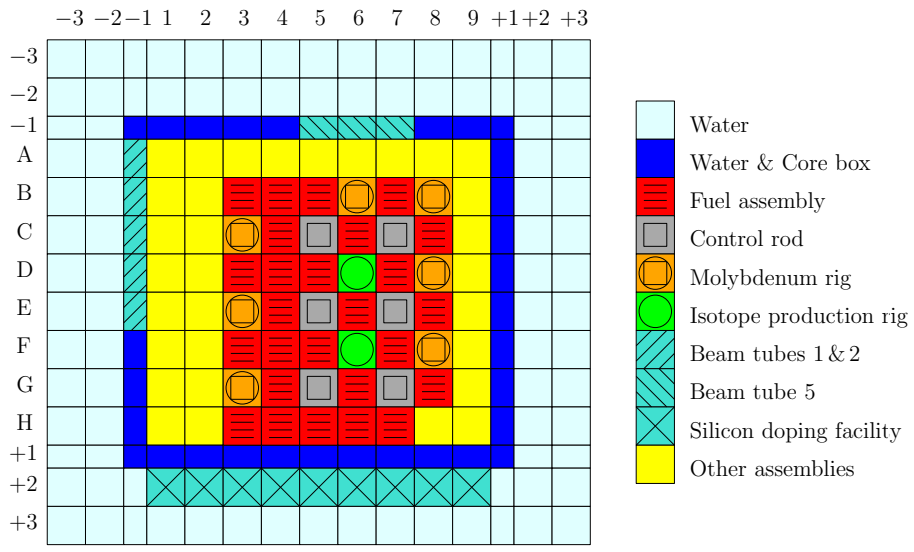


Figure 2: *Top view of the core layout of the SAFARI-1 model used in OSCAR-4.*

There are several potential objectives and constraints that may be considered within an ICFMO problem instance for SAFARI-1. The core parameters that we consider for neural network modelling are listed in Table 1. They comprise a number of operational and safety parameters for SAFARI-1. A detailed description of these parameters fall beyond the scope of this paper but may, however, be found in [12].

Parameter label	Parameter description	Parameter type
Beams 1 & 2	neutron flux in beam tubes 1 & 2	operational
Beam 5	neutron flux in beam tube 5	operational
Silicon	neutron flux in the silicon doping facility	operational
IPR-1	neutron flux in the IPR at position D6 in Figure 2	operational
IPR-2	neutron flux in the IPR at position F6 in Figure 2	operational
^{99}Mo total	power levels in all molybdenum rigs	operational
^{99}Mo min	power level of the molybdenum rig with minimum power	operational
CBW	*control bank worth*	safety
SM	*shutdown margin*	safety
ER	*excess reactivity*	operational
Abs PP	absolute *power peak*	safety
Rel PP	relative *power peak*	safety

Table 1: *SAFARI-1 core parameters considered for neural network modelling.*

4 The construction of neural networks for SAFARI-1

We utilised the Neural Network Toolbox [1] within the Matlab software suite [14] for the construction of our ANNs. The steps that were followed during their construction are presented in this section.

4.1 Training data

We considered an actual SAFARI-1 operational cycle during the year 2012 for the construction and training of our networks. In our previous studies, we evaluated numerous reload configurations for that cycle as part of solving ICFMO problem instances, each comprising different combinations of core parameters (*i.e.* objectives and constraints). From this collection of evaluated configurations, we selected a subset containing approximately 5 000 configurations achieving the largest and smallest values for each core parameter listed in Table 1. By including these configurations in our set, we attempt to cover the extremes of the core parameter space in which we wish to predict. Furthermore, we randomly generated new reload configurations for that cycle (*i.e.* random permutations of the twenty-six fuel assemblies in their loading positions, sampled according to a uniform distribution), evaluated them using OSCAR-4, and added them to our set. By doing so, we attempt to achieve diversity in both the configuration space and core parameter space. A set of 20 000 fuel reload configurations thus constructed forms our representative subset. For each network, this set was randomly partitioned into a *training set* of 17 000 configurations and a *test set* of 3 000 configurations.

4.2 Architecture and input specification

Twelve multilayer feedforward neural networks were constructed, one each for predicting the core parameters listed in Table 1. As such, each network contains only one neuron in its output layer. Although we could theoretically have constructed a single neural network for predicting all twelve parameters, preliminary testing indicated that the prediction errors thus incurred would be too large in practice. The input layer for each network contains twenty-six neurons that correspond to the fuel loading positions in the SAFARI-1 core. Furthermore, the inputs to the network were chosen as the *uranium-235* (^{235}U) mass of each fuel assembly assigned to each loading position in a reload configuration. Each network contains only one hidden layer which should be sufficient for our purposes, as mentioned in Section 2.2. The numbers of hidden neurons were determined empirically in each case, as described later in this section.

The activation functions suggested in the Toolbox were adopted in each network, namely a hyperbolic tangent sigmoidal function for neurons in the hidden layer, and a linear function for neurons in the output layer. In addition, the Toolbox also performs pre-processing on the network input and output data which normalise the values to a range of $[-1, 1]$. This can improve the efficiency of the network training process.

4.3 Training

Several training algorithms are available for use in the Toolbox. The Levenberg-Marquardt backpropagation algorithm [6] is very fast and generally recommended as a first-choice algorithm to adopt when training feedforward networks [1]. Given our intended purpose of using the neural networks for predicting thousands upon thousands of different reload configurations during ICFMO studies, however, we opted to place more emphasis on the generalisation of our networks, rather than the speed of training them. Therefore, we chose to use the Bayesian regularisation backpropagation algorithm within the Toolbox.

As demonstrated by Foresee & Hagan [5], networks trained using Bayesian regularisation typically achieve excellent generalisation. Furthermore, one of the features of the algorithm is that it provides a measure of how many network weights are effectively used in reducing a network's error performance measure. This effective number of weights can aid in deciding whether a network contains an appropriate number of hidden neurons [5]. If the effective and total number of weights are very close to each other, then more hidden neurons should be added to the network.

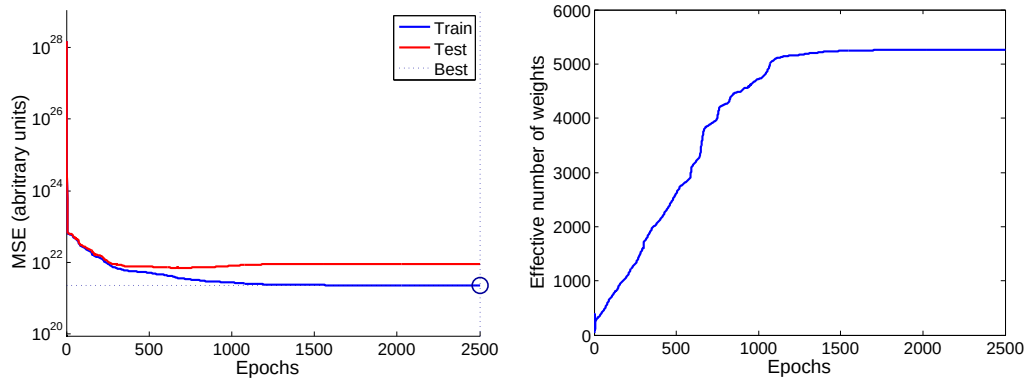
As mentioned earlier, we determined the number of hidden neurons included in each network by an empirical study. In it, we incrementally increased the total number of hidden neurons in the network architecture from 100 up to a satisfactory number (in increments of 50), and performed the training using the default stopping criteria provided in the Toolbox. Upon termination of the training process, we determined from the results whether the training algorithm had, in fact, converged. This can be concluded if the *mean squared error* (MSE) of the training and test sets, and the effective number of weights remain relatively constant over several epochs [1]. If the training process did not converge, we increased the number of epochs performed and continued the training process. When the training process did converge, we verified that the effective and total number of weights were not too close to each other, and that the absolute relative errors for the training and tests sets were acceptable. If either verification failed, we increased the number of hidden neurons and restarted the training process. Following this approach, we constructed our twelve final neural networks for predicting the SAFARI-1 core parameters listed in Table 1. The computation time required for training each final network varied between six and thirty-seven hours, depending on their hidden layer sizes and the number of epochs required.

5 Training results and the application of the networks

By using the neural networks instead of OSCAR-4, the computation time required for the evaluation of a reload configuration may be reduced by four orders of magnitude.

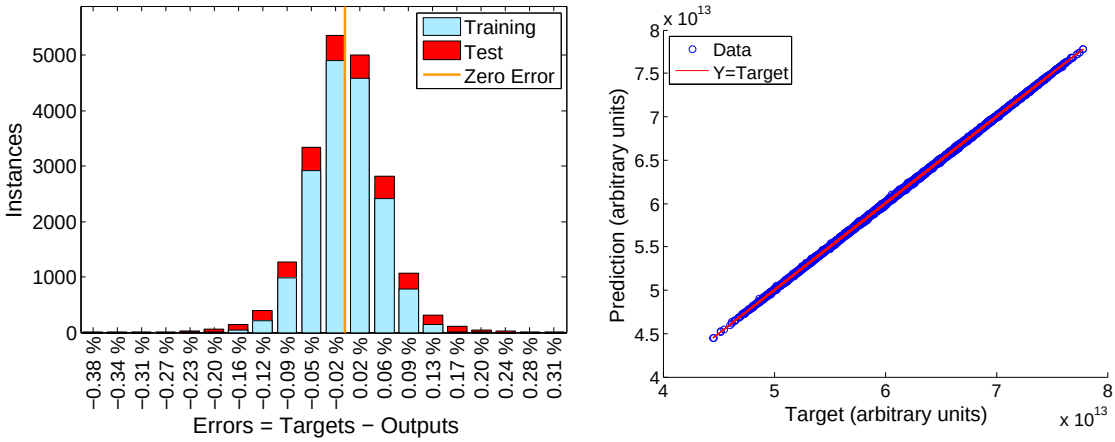
Due to space limitations, we present only the Matlab graphical results that were obtained for one of our constructed neural networks, namely that of Beam 5 with 200 hidden neurons. In Figure 3, we first present the convergence graphs of the network training process. We observe in Figure 3(a) that the MSE of the training and test sets remains approximately constant for several hundred epochs, while similar behaviour is observed in Figure 3(b) for the effective number of weights.

Next, we present results relating to the network's predictive capabilities in Figure 4. An error histogram in Figure 4(a) illustrates the distribution of the training and test set errors. We observe that the distribution of errors visually resembles a normal distribution, and it was determined that approximately 73 % of the errors fall within one standard deviation away from the mean (and approximately 95 % within two standard deviations). A scatter graph of the actual (target) values versus the predicted values is presented in Figure 4(b). From the graph we observe the exceptionally good fit of our network predictions to their target values for the combined training and test sets.



(a) MSE of the training and test set. **(b)** Effective number of weights.

Figure 3: *Matlab graphical results for the Beam 5 neural network convergence.*



(a) Error histogram with 20 bins. **(b)** Target versus predicted values.

Figure 4: *Matlab graphical results for the Beam 5 neural network predictions.*

Since the test set contains only 3 000 reload configurations, we wished to further test the accuracy and generalisation of our networks. We therefore created a larger set of new (unseen) reload configurations, namely a *verification set* of 30 000 random (uniform) reload configurations, evaluated by OSCAR-4. The neural networks were then applied to the verification set in order to predict the corresponding SAFARI-1 core parameters. A summary of all twelve networks' performances in respect of their prediction errors is presented in Table 2. The table contains the average and maximum absolute relative prediction errors for the training, test and verification sets.

The networks produce good predictions on average, with ten of the twelve networks producing an average error of less than 1 % on the test and verification sets. Furthermore, the maximum errors for networks that predict flux levels are all also less than 1 % on the test and verification sets. The larger maximum errors of the Abs PP and Rel PP networks (approximately 8 % and 10 % on the test and verification sets, respectively) are still of an acceptable accuracy when compared to an error of 14 % found in the literature [8].

	Training set		Test set		Verification set	
	Average	Maximum	Average	Maximum	Average	Maximum
Beams 1 & 2	0.07 %	0.35 %	0.13 %	0.75 %	0.13 %	0.88 %
Beam 5	0.06 %	0.38 %	0.12 %	0.73 %	0.13 %	0.82 %
Silicon	0.07 %	0.37 %	0.13 %	0.61 %	0.13 %	0.80 %
IPR-1	0.09 %	0.52 %	0.14 %	0.92 %	0.14 %	0.88 %
IPR-2	0.09 %	0.44 %	0.12 %	0.95 %	0.12 %	0.71 %
^{99}Mo total	0.19 %	0.93 %	0.20 %	0.90 %	0.20 %	0.96 %
^{99}Mo min	0.75 %	2.96 %	0.82 %	3.61 %	0.85 %	3.60 %
CBW	0.03 %	0.20 %	0.07 %	0.33 %	0.07 %	0.41 %
SM	0.12 %	0.67 %	0.24 %	1.28 %	0.25 %	1.73 %
ER	0.09 %	0.60 %	0.18 %	0.97 %	0.19 %	1.24 %
Abs PP	0.73 %	4.70 %	1.70 %	8.19 %	1.81 %	10.02 %
Rel PP	0.80 %	5.52 %	1.62 %	8.71 %	1.67 %	10.59 %

Table 2: *Average and maximum absolute relative prediction errors.*

6 Predictions for other operational cycles

Given that our networks receive the ^{235}U mass of a fuel assembly as input, it is possible that the networks might also be used to predict SAFARI-1 core parameters for a different operational cycle than the one they were trained on. In order to test this possibility, we chose two other operational cycles from the SAFARI-1 history which exhibit relatively different fuel distributions than the original cycle. Using the fuel distributions from the two other cycles, we generated a set of 4 500 random (uniform) reload configurations for each cycle, and evaluated them using OSCAR-4. The neural networks were then applied to predict SAFARI-1 core parameters for these additional operational cycles. The average and maximum absolute relative prediction errors are presented in Table 3.

	Additional cycle 1		Additional cycle 2	
	Average	Maximum	Average	Maximum
Beams 1 & 2	1.84 %	7.11 %	2.58 %	9.01 %
Beam 5	2.18 %	9.47 %	4.91 %	17.62 %
Silicon	9.52 %	15.37 %	5.50 %	11.81 %
IPR-1	2.61 %	7.24 %	5.78 %	10.89 %
IPR-2	4.17 %	8.27 %	5 %	8.40 %
^{99}Mo total	0.31 %	1.50 %	3.62 %	4.76 %
^{99}Mo min	3.55 %	8.05 %	7.47 %	12.91 %
CBW	3.25 %	8.64 %	2.86 %	9.45 %
SM	2.87 %	16.17 %	12.48 %	24.87 %
ER	2 %	6.68 %	16.71 %	22.68 %
Abs PP	5.81 %	22.88 %	5.28 %	22.64 %
Rel PP	5.09 %	20.04 %	4.47 %	19.61 %

Table 3: *Average and maximum absolute relative prediction errors (additional cycles).*

Our neural networks yield predictions of unacceptable quality for the additional cycles. The maximum and average errors for these cycles are worse than those of the original cycle by approximately an entire order of magnitude. Hence, we would not be able to use our networks to predict SAFARI-1 core parameters for other operational cycles than the

one we trained them for. However, if we were able to construct networks with sufficient generalisation in order to predict parameters for an arbitrary SAFARI-1 cycle, possibilities exist for using the networks in areas other than ICFMO as well, *e.g.* within reactor codes or simulator training. Such networks would require retraining on a much larger set of configurations which incorporates different fuel distributions. Furthermore, the network architectures would likely require alteration for additional input neurons so as to incorporate more information (*e.g.* axial ^{235}U mass distribution of each assembly and/or control rod positions) for sufficient accuracy.

7 Conclusions

Artificial neural networks were constructed in this paper as surrogate models for predicting core parameters for the SAFARI-1 nuclear research reactor in the context of a specific operational cycle. The Neural Network Toolbox within the Matlab software suite was utilised for the construction of these networks. The ANNs were applied to test and verification sets corresponding to the specific operational cycle. The results obtained in respect of these sets demonstrated the ability of the networks to predict SAFARI-1 core parameters (with acceptable accuracy) much quicker than when using explicit calculations (as in using the OSCAR-4 code). A computation time improvement of four orders of magnitude was achieved. The neural networks can therefore be employed within our investigation into the appropriateness of several techniques for solving multiobjective ICFMO problems for SAFARI-1.

Acknowledgements

The first author was financially supported in part by Necsa via their study assistance scheme, and by the *National Research Foundation* (NRF) of South Africa (Grant 88003). The second and third authors were financially supported by the NRF (Grant 70730 and Grant 70593, respectively). Any opinion, finding, and conclusion or recommendation expressed in this material is that of the author(s) and the NRF does not accept any liability in this regard.

References

[1] Beale MH, Hagan MT & Demuth HB, 2014, *Neural Network Toolbox*, [Online], [Cited September 19, 2014], Available from `http://www.mathworks.com/help/releases/R2014a/pdf_doc/nnet/index.html`

[2] Bishop CM, 1995, *Neural networks for pattern recognition*, Clarendon Press, Oxford.

[3] Erdoğan A & Geçkinli M, 2003, *A PWR reload optimisation code (XCore) using artificial neural networks and genetic algorithms*, Annals of Nuclear Energy, **30**, pp. 35–53.

[4] Fausett LV, 1994, *Fundamentals of neural networks: Architectures, algorithms, and applications*, Prentice-Hall, Englewood Cliffs (NJ).

[5] Foresee FD & Hagan MT, 1997, *Gauss-Newton approximation to Bayesian learning*, Proceedings of the International Conference on Neural Networks, IEEE, Houston (TX), pp. 1930–1935.

[6] Hagan MT & Menhaj MB, 1994, *Training feedforward networks with the Marquardt algorithm*, IEEE Transactions on Neural Networks, **5**(**6**), pp. 989–993.

[7] Hornik K, 1989, *Multilayer feedforward networks are universal approximators*, Neural Networks, **2**, pp. 359–366.

[8] Mazrou H & Hamadouche M, 2004, *Application of artificial neural network for safety core parameters prediction in LWRRs*, Progress in Nuclear Energy, **44**(**3**), pp. 263–275.

[9] Mazrou H & Hamadouche M, 2006, *Development of a supporting tool for optimal fuel management in research reactors using artificial neural networks*, Nuclear Engineering and Design, **236**, pp. 255–266.

[10] Mehrotra K, Mohan CK & Ranka S, 1997, *Elements of artificial neural networks*, MIT Press, Cambridge (MA).

[11] Mirvakili SM, Faghihi F & Khalafi H, 2012, *Developing a computational tool for predicting physical parameters of a typical VVER-1000 core based on artificial neural network*, Annals of Nuclear Energy, **50**, pp. 82–93.

[12] Schlünz EB, Bokov PM, Prinsloo RH & Van Vuuren JH, 2015, *A unified methodology for single- and multiobjective in-core fuel management optimisation*, Annals of Nuclear Energy, Submitted.

[13] Svozil D, Kvasnička V & Pospíchal J, 1997, *Introduction to multi-layer feed-forward neural networks*, Chemometrics and Intelligent Laboratory Systems, **39**, pp. 43–62.

[14] The MathWorks Inc, 2014, *MATLAB R2014a — The language of technical computing*, [Online], [Cited March 26, 2015], Available from `http://www.mathworks.com/products/matlab/`.

Proceedings of the 2015 ORSSA Annual Conference
pp. 23–32
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Decision support for the assignment of real-estate agents to suburbs

J Janse van Rensburg*† JH van Vuuren‡

Abstract

Matching a real-estate agent to a suburb that best fits his or her property marketing expertise is an important (and difficult) decision for a real-estate agency aiming to achieve good turnover. In this respect the agents have to take into consideration the expected income that can be associated with a suburb when searching for new property stock to include in their personal portfolios. Real-estate agencies are also interested in having a presence in as many suburbs of a city as possible so as to be able to attract potential sellers in a large area. A bi-attribute decision support framework is proposed in this paper to aid real-estate agencies in selecting the best suburbs in such a way that both the agency's exposure and its expected income is maximised within a specified city. For this purpose a rich data set is required, containing all listed properties within the city, the agency's corresponding listing agents, as well as the values, locations and characteristics of properties in the market within the city. The working of the decision support framework is illustrated in the context of a special case study involving a real-estate agency in the Somerset West basin of the City of Cape Town, which employs eleven estate agents.

Key words: Real-estate, Agent assignment, Multiple criteria decision support, Suburb selection.

1 Introduction

Decision-making techniques can be separated into two broad categories: *group decision-making techniques* and *individual decision-making techniques*. There is a growing need by both individuals and businesses in a variety of sectors to use software in aid of good decision-making. This is due to increasing complexity associated with decision making as technology develops and data become abundantly available, thereby increasing the need to consider larger sets of stakeholders, criteria, alternatives and other factors that affect decisions.

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: johanjvrens@sunore.co.za

†Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

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

The objective in this paper is to put forward a decision support framework for the assignment of real-estate agents to suburbs based on the available properties for sale as a function of time within the Somerset West basin. The system is aimed at the facilitation of agency-based decision-making in cases where the sets of assignment alternatives are large. The paper is a report on work in progress within a larger and ongoing research project at Stellenbosch University and is structured as follows. After performing a brief review of relevant literature in §2 on assignment problems and decision-making in the context mentioned above, we describe in §3 the problem considered in this paper and we turn our attention in §4 to a discussion on how the required case study real-estate data were captured and cleaned. This is followed, in §5, by a detailed proposal of how the algorithms associated with the multi-criteria assignment decision may be utilised to facilitate decision-making in the context of real-estate agent assignment to suburbs. An overview of the results obtained when applying the methodology to the data described in §4 is presented in §6. The paper closes in §7 with a brief conclusion and some pointers to related further work are given in §8.

2 Related literature

The assignment problem is one of the fundamental combinatorial optimisation problems in the operations research literature. It consists of finding a maximum weight matching (or minimum weight perfect matching) in a weighted bipartite graph. In its most general form, the problem may be described as follows. There are a number of agents and a number of tasks. Any agent can be assigned to perform any task, incurring some cost that may vary depending on the agent-task assignment. It is required to perform all tasks by assigning exactly one agent to each task in such a way that the total cost of the assignment is minimised.

If the numbers of agents and tasks are equal and the total cost of the assignment for all tasks is equal to the sum of the costs for each agent (or the sum of the costs for each task), then the problem is called the classical assignment problem [1]. A variety of algorithms have been devised to solve this assignment problem, such as the Hungarian algorithm [2] and the Auction algorithm [3]. There are even algorithms for solving multi-objective versions of the assignment problem [4].

In economics, diminishing returns (also called law-of-diminishing returns or marginal returns) is the decrease in the marginal (incremental) output of a production process as the amount of a single factor of production is incrementally increased, while the amounts of all other factors of production remain constant [6]. This simply means higher investment does not always fetch better returns. The law of diminishing returns states that as one invests more money, effort or time in an investment, the marginal rate of return eventually drops.

3 Problem description

A basic problem faced by real-estate agencies, and the problem considered in this paper, involves the selection of some subset of suburbs in which to invest the time, energy and resources of the agents employed by an agency. This problem is not only limited to the current staff of an agency, as agencies are typically also interested in the potential gains that may realise as a result of expanding their workforce. Finding the most valuable set of suburbs to focus on within a city is based on two criteria. The first criterion is to find a suburb set that maximises the agency's expected income, while the second criterion is to ensure that the agency is exposed to as many suburbs and subsequent properties which fall under its jurisdiction as possible in order to minimise the risk of its properties being withdrawn from the listings.

The following assumptions are made in this paper:

1. *All agents are equally likely to make a sale in any suburb.* This assumption is necessary in order to consider all suburbs as fertile ground for all the agents. It may, however, be relaxed by recording viable suburb sets per agent, based on their respective skill sets, and then limiting the suburb set assignable to each agent.

2. *All agents have similar workloads and are willing to work in any suburb.* This assumption negates the possibility that some agents may be assigned to suburbs containing only a few properties while other agents may be assigned to suburbs with a large collection of properties, in which case their workloads and respective incomes may differ vastly. The impact of this assumption can easily be reduced by assigning more than one agent to a suburb if there are too many properties in a suburb to be handled by a single agent, but this possibility lies beyond the scope of the current paper.

3. *The only agents competing in a specific suburb are the agents with listings in the suburb.* This assumption is made to simplify the problem under consideration, as the total number of agents is in reality typically unknown, but can be estimated by recording all suburbs' sales records per agent in the city so as to construct a suburb list per agent.

4. *All agents within the agency are to be assigned to one suburb only, but many agents may be assigned to any one suburb.*

In the case study considered in this paper, which involves the Somerset West basin (partitioned into 63 suburbs and comprising approximately 19 000 properties), only residential properties are considered. No *commercial properties*, *farms*, *apartments* or *plots* are therefore considered, as the assignment problem for the set of combined property types can be partitioned into smaller disjoint problems specific to each property type.

Of the 44 real-estate agencies operating in the Somerset West basin, one specific real-estate agency is considered as case study agency throughout this paper. This real-estate agency was selected as it already tracks all the information and data necessary to solve the problem at hand.

The problem considered here is different from the classical assignment problem, and even its multi-objective variants, in that the suburb properties remain the same for all agent assignments and that the agents and suburbs are not equal in number.

4 Data capturing and cleaning

Somerset West originally consisted of three separate areas, namely Bakershoogte, Parel Vallei and Somerset West, but today they are all merged to form the Somerset West basin. Due to this legacy divide, care must be taken when attempting to identify properties uniquely when working with their addresses, since an erf number may potentially refer to three different properties.

In what follows, a *property snapshot* refers to a *Microsoft Excel* file containing records of all properties that are in the market by all agencies in Somerset West. Such a data file is produced bi-weekly. The property snapshot data used in this paper were taken on 16 March 2015 and contained 571 properties, with 19 attributes tracked for each of these properties. For the purpose of this paper, only four of these attributes were used. They are the *advertised price* of the property, the *property address* (which includes the suburb in which it resides), the *agent* who has a mandate on the property and the *agency* for which the agent works.

The data set first had to be cleaned by removing all unusual characters and spacings. Critical fields were checked for completeness and cross-checked for accuracy and spelling. All the suburb field data were validated against *Google*'s map *application programming interface* (API). The agent details were cleaned and unique agency data were extracted corresponding to the case study agency mentioned above.

The programming language R was used to read the data files, clean the data and plot the graph included in this paper.

5 Modelling approach

Let m denote the number of agents to be assigned to n suburbs, and let

$$\boldsymbol{a}^{(i)} = \left[\begin{array}{cccccc} a_1^{(i)} & \cdots & a_j^{(i)} & \cdots & a_n^{(i)} \end{array} \right], \quad i = 1, \ldots, \ell$$

denote the i^{th} assignment alternative, where $a_j^{(i)}$ represents the number of agents assigned to suburb j in alternative i.

The i^{th} assignment alternative has the property that

$$\sum_{j=1}^{n} a_j^{(i)} = m, \quad i = 1, \ldots, \ell. \tag{1}$$

An upper bound

$$a_j^{(i)} \leq u, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, n. \tag{2}$$

is placed on the number of agents that may be assigned to any one suburb. The average agent income x_j generated by suburb j is the total property sales value s_j in suburb j divided by the total number of agents α_j (both internal and external to the agency) working in suburb j.

Let y_j denote the total number of properties listed in suburb j and let r denote the agency's commission rate. The evaluation criteria associated with the i^{th} assignment alternative may be expressed mathematically as

$$c_1^{(i)} = r \sum_{j=1}^{n} x_j a_j^{(i)}, \qquad i = 1, \ldots, \ell, \tag{3}$$

and

$$c_2^{(i)} = \sum_{j=1}^{n} y_j a_j^{(i)}, \qquad i = 1, \ldots, \ell. \tag{4}$$

The evaluation matrix

$$\boldsymbol{C} = \begin{bmatrix} c_1^{(1)} & c_2^{(1)} \\ \vdots & \vdots \\ c_1^{(i)} & c_2^{(i)} \\ \vdots & \vdots \\ c_1^{(\ell)} & c_2^{(\ell)} \end{bmatrix}$$

contains all quality evaluations associated with assignment alternatives with respect to the criteria in (3)–(4) and for all $i = 1, \ldots, \ell$.

The expected agency income $E_j^{(i)}$ for the i^{th} assignment alternative is measured as the average income x_j per agent, weighted by the number of agency agents $a_j^{(i)}$ assigned to suburb j, *i.e.*

$$E_j^{(i)} = a_j^{(i)} \times x_j^{(i)}, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, n. \tag{5}$$

In order to forecast the expected return associated with assigning an additional agent to a suburb, the expected return may be calculated as the marginal return the newly assigned agent will add to the expected return for suburb j, taking into account the smaller share the already assigned agents will receive after the assignment.

As a result of the problem dimensions, we designed a heuristic for finding approximate solutions to the case study agent-suburb assignment problem instance. Given a sufficiently large second criterion value, our algorithm selects the best possible alternative $\boldsymbol{a}^{(i)}$ based on the first criterion which satisfies the lower bound threshold on the second criterion. This is achieved by sorting the expected agency income values $E_1^{(i)}, \ldots, E_n^{(i)}$ in decreasing order, selecting the m best possible assignments according to the first criterion and then testing whether the lower bound condition set on the second criterion is met by the assignment. The agent assignment is altered iteratively by replacing the worst assignment with respect to the first criterion with the best assignment for the second criterion that was not in the list, followed by the second best assignment, then the third best assignment, *etc.* until the second criterion's threshold value is reached.

Algorithm 1 may be used to generate an alternative assignment that will meet the lower bound selected for the second criterion.

Algorithm 1 Generate an alternative assignment

sort data desending on criteria one's values
$t \leftarrow$ select the lower bound for the second criterion
$a \leftarrow$ first criterion data
$b \leftarrow$ second criterion data
$x \leftarrow$ assign the top m records of the sorted data
$y \leftarrow$ assign the remaining $n - m$ records
$i = 0$
for i in 1:nrow(x)-1 **do**
    if $sum(x.b) >= t$ **then** break
    end if
    $indexY \leftarrow$ which(y$b == max(y$b))[1] ▷ select second best value criterion
    $indexX \leftarrow nrow(x) - i$
    $xStor \leftarrow x[indexX,]$
    $x \leftarrow rbind(x[-indexX,], y[indexY,])$
    $y \leftarrow rbind(y[-indexY,], xStor)$
end for

In cases where more than one agent is assigned to a suburb, the suburb may appear multiple times in the assignment matrix.

6 Case study results

The results presented in this section all refer to the 16 March 2015 data snapshot of listed properties for the case study agency, as described in §4.

The case study real-estate agency has the following attributes. Currently there are eleven active agents specialising in *residential property* at the agency. Together they currently cover 15 unique suburbs of the total of 63 suburbs; two suburbs are covered twice. These parameters lead to $\ell = \binom{n+m-1}{m} \approx 3.55 \times 10^{12}$ possible assignment alternatives in total. Some agents also focus on more than one suburb, resulting in a total of 17 possible suburb assignments. The current agency configuration of 17 assignments results in an expected income of R 3 139 126 and an exposure to 137 properties across 15 suburbs. These suburbs are shown in Table 1.

1	Audas	3	Bayview Heights	7	Briza	11	Die Wingerd
14	Fernwood Estate	18	Haumannshof	23	Heritage Park	29	La Sandra
31	Links	39	Raithby	43	Schonenberg	47	Somerset Mall
49	Somerset Ridge	53	Stuarts Hill	60	Winery Road		

Table 1: *The case study real-estate agency's 15 suburbs used for their 17 agent assignments, denoted in the text by $\boldsymbol{a}^{(0)}$, as it was on 16 March 2015.*

In the results reported in this section only eleven agent assignments are made in contrast to the case study agency's 17 agent assignments shown in Table 1. In other words, only one

suburb is assigned per agent, while multiple agents may be assigned to a suburb (therefore possibly resulting in fewer than m uniquely covered suburbs).

With only eleven agents, the largest possible value for the first criterion is obtained by the assignment of agents to the suburbs shown in Table 2, which achieves an expected income of R 9 289 561 and an exposure to 91 properties. This assignment was calculated by selecting the 11 best suburb agent assignments based on expected returns.

12	Erinvale Estate	21	Heldervue	43	Schonenberg	50	Spanish Farm
51	Stellenbosch	61	Worlds View				

Table 2: *The assignment alternative $\boldsymbol{a}^{(1)}$ of agents to the eleven suburbs yielding the largest possible value for the first assignment criterion.*

To calculate the upper bound for the second criterion the list of suburbs was sorted descending according to number of properties and the eleven highest suburbs was selected. These suburbs are shown in Table 3. The best assignment of agents to these eleven suburbs achieves an exposure to 269 properties and an expected return of R 6 666 329.

1	Audas	4	Bizweni	12	Erinvale Estate	21	Heldervue
22	Helena Heights	28	La Concorde	33	Montclair	37	Parel Vallei
43	Schonenberg	44	Sir Lowry's Pass	50	Spanish Farm		

Table 3: *The assignment $\boldsymbol{a}^{(2)}$ of agents to the eleven suburbs yielding the largest possible value for the second assignment criterion.*

The recommended assignment of agents to the suburbs in Table 4 was calculated by replacing the lowest expected return in the list of assignments in $\boldsymbol{a}^{(1)}$ with the second, third, *etc.* lowest expected return until the desired value for criterion 1 had been reached. For the case study agency the best assignment of agents to these properties resulted in an exposure to 151 properties and an expected return of R 9 197 771.

12	Erinvale Estate	21	Heldervue	37	Parel Vallei	43	Schonenberg
50	Spanish Farm	51	Stellenbosch	61	Worlds View		

Table 4: *The assignment alternative $\boldsymbol{a}^{(3)}$ achieving the best possible value for the second assignment criterion with a first assignment criterion larger than or equal to that of $\boldsymbol{a}^{(0)}$.*

The assignment alternative for the current agency suburb selection is

$$\boldsymbol{a}^{(0)} = \left[\begin{array}{ccccccc} a_1^{(0)} = 2 & a_3^{(0)} = 1 & a_7^{(0)} = 1 & a_{11}^{(0)} = 1 & a_{14}^{(0)} = 1 & a_{18}^{(0)} = 1 & a_{23}^{(0)} = 1 \\ a_{29}^{(0)} = 1 & a_{31}^{(0)} = 1 & a_{39}^{(0)} = 1 & a_{47}^{(0)} = 1 & a_{49}^{(0)} = 1 & a_{53}^{(0)} = 1 & a_{60}^{(0)} = 2 \end{array} \right]$$

with corresponding evaluation matrix entries

$$\boldsymbol{C}_0 = \left[\begin{array}{cc} c_1^{(0)} = 3\,139\,126 & c_2^{(0)} = 137 \end{array} \right],$$

where we follow the convention that all a-values not listed assume the value zero. The assignment alternative achieving the largest income was, however, found to be

$$\boldsymbol{a}^{(1)} = \left[\begin{array}{ccccc} a_{12}^{(1)} = 5 & a_{21}^{(1)} = 2 & a_{43}^{(1)} = 1 & a_{50}^{(1)} = 1 & a_{51}^{(1)} = 1 & a_{61}^{(1)} = 1 \end{array} \right]$$

with corresponding evaluation matrix entries

$$\boldsymbol{C}_1 = \left[\begin{array}{cc} c_1^{(1)} = 9\,289\,561 & c_2^{(1)} = 91 \end{array} \right],$$

while the assignment alternative achieving the largest exposure is

$$\boldsymbol{a}^{(2)} = \left[\begin{array}{cccccc} a_1^{(2)} = 1 & a_4^{(2)} = 1 & a_{12}^{(2)} = 1 & a_{21}^{(2)} = 1 & a_{22}^{(2)} = 1 & a_{28}^{(2)} = 1 \\ a_{33}^{(2)} = 1 & a_{37}^{(2)} = 1 & a_{43}^{(2)} = 1 & a_{44}^{(2)} = 1 & a_{50}^{(2)} = 1 & \end{array} \right]$$

with corresponding evaluation matrix entries

$$\boldsymbol{C}_2 = \left[\begin{array}{cc} c_1^{(2)} = 6\,666\,329 & c_2^{(2)} = 269 \end{array} \right].$$

The agency was quite happy with their current exposure and sought the best possible expected return with a similar or better exposure. Hence, we recommended the assignment

$$\boldsymbol{a}^{(3)} = \left[\begin{array}{ccccccc} a_{12}^{(3)} = 4 & a_{21}^{(3)} = 1 & a_{37}^{(3)} = 1 & a_{43}^{(3)} = 1 & a_{50}^{(3)} = 2 & a_{51}^{(3)} = 1 & a_{61}^{(3)} = 1 \end{array} \right]$$

with corresponding evaluation matrix entries

$$\boldsymbol{C}_3 = \left[\begin{array}{cc} c_1^{(3)} = 9\,197\,771 & c_2^{(3)} = 151 \end{array} \right],$$

which yields an expected return of 193% of the current expected return with a 10% increase in property exposure.

The results described above are summarised in Table 5 and illustrated graphically in Figure 1.

$\boldsymbol{a}^{(i)}$	Total suburb expected income	Total property exposure	Description	Status
$\boldsymbol{a}^{(0)}$	R 3 139 126	137	current	dominated
$\boldsymbol{a}^{(1)}$	R 9 289 561	91	highest earnings	non-dominated
$\boldsymbol{a}^{(2)}$	R 6 666 329	269	highest exposure	non-dominated
$\boldsymbol{a}^{(3)}$	R 9 197 771	151	recommended	non-dominated

Table 5: *Relative performances of four assignment alternatives.*

7 Conclusion

The president of the case study agency's intuition going into this research endeavour was that 12 Erinvale Estate, 37 Parel Vallei, 43 Schonenberg and 50 Spanish Farm are the suburbs that are worth spending extra time and money on.

The results obtained in §6 were presented to the president and the recommended alternative $\boldsymbol{a}^{(3)}$ aligned with what the agency thought might be the most profitable suburbs. It is interesting to note how similar the recommended assignment is to the intuition of the agency's president. Although the assumptions made in §3 seem to be rather restrictive, the results are quite encouraging.

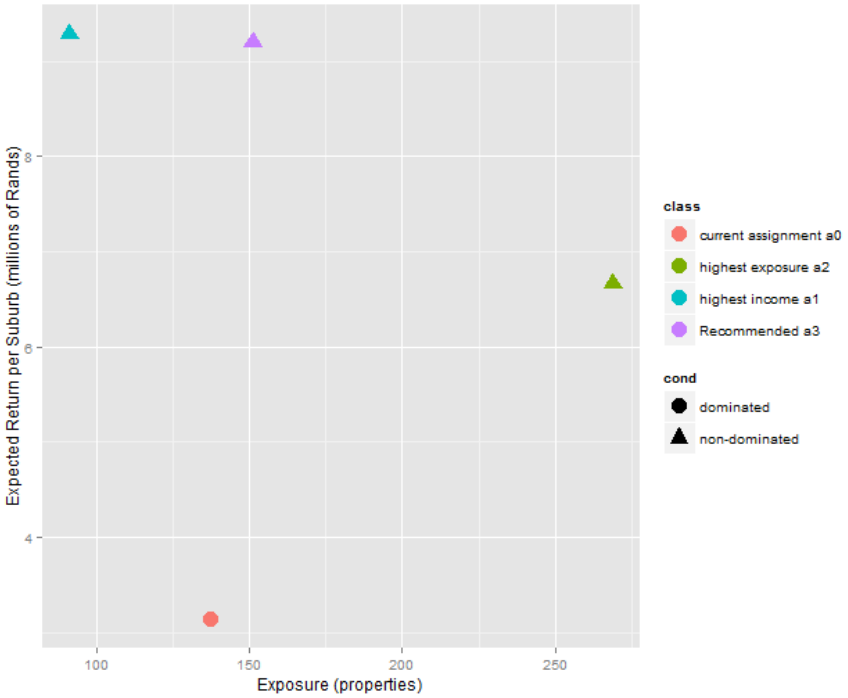


Figure 1: *The real-estate agency's current assignment is suboptimal compared to the approximately Pareto set, shown as triangles. Both the best possible expected return in (3) and the best possible exposure in (4), are shown.*

8 Further work

This paper only touched on the basics of the fundamental problem of assigning agents to suburbs. Much more can, however, be done to refine the modelling approach so as to better reflect real-world exposure and expected return (such as, for example, taking into account the expertise and marketing skills of agents). The results obtained in §6 may also be compared with those returned by other (meta)heuristic and MCDA approaches, such as, for instance, selecting an approximately Pareto-optimal set of solutions generated by an evolutionary algorithm. Historical data may further be used to predict popular suburbs for future assignments.

References

[1] Munkres J, 1957, *Algorithms for the assignment and transportation problems*, Journal of the Society for Industrial and Applied Mathematics, **5(1)**, pp. 32–38.

[2] Kuhn HW, 1991, *On the origin of the Hungarian Method, history of mathematical programming*, North Holland, Amsterdam, pp. 77–81.

[3] Bertsekas DP, 1988, *The auction algorithm: A distributed relaxation method for the assignment problem*, Annals of Operations Research, **14(1)**, pp. 105–123.

[4] De PK & Yadav B, 2011, *An algorithm to solve multi-objective assignment problem using interactive fuzzy goal programming approach*, International Journal of Contemporary Mathematical Sciences, **6(34)**, pp. 1651–1662

[5] Weistroffer HR & Narula SC, 2005, *Multiple criteria decision support software*, Springer, New York (NY).

[6] Shephard RW, 1969, *Proof of the law of diminishing returns*, University of California, Berkeley.

Proceedings of the 2015 ORSSA Annual Conference
pp. 33–43
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Decision support for the selection of water release strategies at open-air irrigation reservoirs

JC van der Walt* JH van Vuuren†

Abstract

Water earmarked for irrigation purposes in the agricultural sector is typically stored in open-air reservoirs. The availability of irrigation water greatly impacts the profitability of this sector and this availability is largely determined by prudent decisions related to water release strategies at open-air reservoirs. The release strategy for an open-air irrigation reservoir is typically decided upon by a board of management at the start of the hydrological year. The selection of such a strategy is difficult, since the objectives which should be pursued are not generally agreed upon and unpredictable weather patterns cause reservoir inflows to vary substantially between hydrological years. A mathematical model is proposed in this paper which may form the basis of a decision support system for the selection of a beneficial water release strategy. Based on historical data, the proposed model generates a probability distribution of the reservoir volume at the end of a hydrological year for a given initial water release strategy and stochastically simulated reservoir inflows. The initial strategy is dictated by irrigation demands and reservoir sluice parameters. This strategy is then adjusted iteratively, with the aim of centring the hydrological year end volume distribution on some target value. Adjustments are made according to user-specified weight factors, which represent the demand satisfaction importance of the various decision periods. The repeatability of a given water release strategy is taken to depend on an estimate of the most likely reservoir volume at the end of the hydrological year as a result of this strategy. The probability of water shortage for a given transition volume may be determined using this model by equating the start and end volumes for simulated hydrological years. This information allows for the computation of acceptable tradeoff decisions between the fulfilment of the current hydrological year's demand and the future repeatability of a release strategy.

1 Introduction

Water is one of the most important resources for the sustenance of life on planet earth. Other than its immediate and most obvious use as drinking water, the applications of this resource in human endeavour are extensive. According to the World Economic Forum

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa, email: 17124891@sun.ac.za

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

2015 [7], water crisis is the number one global risk, based on impact to society. The success of many economic systems depends on the continual availability of water.

The portion of the crop farming industry which makes use of irrigation is one such system. This industry depends on the availability of water for its livelihood. Since precipitation periods and river flows are dynamic, volatile in some cases and do not necessarily overlap with demand periods, water must typically be stored to meet irrigation demands. Open-air water reservoirs are most commonly used for this purpose in South Africa.

Water shortages or flood damage may occur downstream from irrigation reservoirs with disastrous effects for the farmers in the region if reservoir levels are not carefully controlled. Thus, an effective release strategy must be employed for beneficial reservoir level control. The release strategy for an open-air irrigation reservoir is typically decided upon either by a formal board or by an informal group of farmers at the start of the hydrological year.

Keerom Dam is an excellent example of an open-air reservoir with the primary purpose of supplying water for irrigation purposes. It is the second largest privately owned open-air reservoir in South Africa and is situated in the Nuy agricultural irrigation district, north-east of Worcester, in the Western Cape. The reservoir's wall height from dam crest to river bed level is 38 meters and when at maximum storage capacity the water surface area is 92 hectares. Nineteen farmers benefit from its water supply, of which six serve on the Nuy irrigation board. This board determines the release strategy for the reservoir on an annual basis.

The best choice of release strategy is not an obvious one for four reasons. Firstly, the objectives which should be met by such a strategy are not generally agreed upon. Irrigation demands should be met, while the risk of water shortage and/or the risk of flood damage may be minimised, or evaporation losses may be minimised. Secondly, unpredictable weather patterns cause reservoir inflows to vary substantially between hydrological years, thus making planning and water allocation exceedingly difficult. Thirdly, the calculation of irrigation demands is a non-trivial problem, which is influenced by the climate as well as the distribution of crop types under irrigation and various agricultural policies. Finally, the persons responsible for the selection of a release strategy may differ vastly in their attitude toward risk, which plays a critical role in the selection of a consensus strategy.

In this paper, a mathematical model is proposed which may form the basis of a decision support system for the selection of a beneficial water release strategy. The risk related to a given strategy is quantified in the model, so as to accommodate tradeoff decisions between the fulfilment of the current hydrological year's demand and future repeatability of good strategies.

This paper is organised as follows. First a brief literature review pertaining to existing models for reservoir operation is given in §2, after which the assumptions made in the development of the model proposed in this paper are listed and motivated in §3. A framework of our modelling approach, depicting the required inputs, processes and information flows is supplied in §4, after which the relevant processes are described in more detail. More specifically, the simulation of inflows is described in §5, after which the method proposed for generating period volume distributions is discussed in §6. A method for the quantification of risk is described next in §7, before concluding remarks are made in §8.

2 Literature Review

In this section, previous models built for irrigation reservoir operation decision support are described in general, after which the focus shifts to a description and evaluation of the models which have previously been implemented at Keerom Dam specifically.

Mathematical models which have been implemented in support of the formulation of good water release strategies for open-air reservoirs can be partitioned into the classes of deterministic optimisation models on the one hand and stochastic simulation and optimisation models on the other.

Several stochastic modelling approaches have been implemented in the context of reservoir operation management. A stochastic linear programming model was, for example, developed by Loucks [2]. According to Yeh [11], this formulation suffers from the problem of dimensionality since its constraints may easily exceed several thousands in real-life applications.

Dynamic programming models are very popular for analysing complex water resource problems because of their ability to incorporate the non-linear and stochastic aspects of these problems into the formulation [11]. Butcher [1] successfully applied a stochastic dynamic programming approach to a multi-purpose reservoir.

More recently, non-linear multi-objective models have been applied in the context of reservoir management problems and these models have been solved using evolutionary algorithms. Reddy & Kumar [4] used this approach to obtain tradeoff release strategies for the multi-purpose Bhadra reservoir system in India, which is used for irrigation and hydro electricity generation. The point by point search approach utilised by traditional optimisation methods are inappropriate for multi-objective optimisation, since these methods produce a single optimal solution. Open-air reservoir operation necessarily involves tradeoff decisions; it may therefore be beneficial to supply more than one solution option to the managers of such reservoirs.

Artificial Neural Networks (ANNs) have been utilised for the simulation of reservoir inflows and for determining release strategies [3]. ANNs is a class of artificial intelligence techniques which mimic the behaviour of neuron connections in the brain. An interconnected set of nodes, called neurons, are organised in input, hidden and output layers. The connections between nodes each has a certain weight, representing the strength of the connection. An ANN learns, much like a biological brain, through the strengthening of connections between selected neurons, by experience. Historical, verified data are fed into the input neurons of the network, after which the output is observed and the weights of the neuron connections are adjusted with the aim of decreasing the deviation of the network output from the historically observed values. This approach works well in the context of reservoir management, even for a limited amount of data in the presence of irregular seasonal variation, is robust and is faster than conventional approaches [3].

Previous models applied to Keerom Dam include a deterministic linear programming model developed for incorporation into a flexible decision support system called ORMADSS (an acronym for *Optimal Reservoir Management Active Decision Support*), by Van Vuuren & Gründlingh [10]. In their model the objective is to obtain an optimal

release strategy for average years, serving as input to the DSS, by

$$\text{minimising } \sum_{i\in\mathcal{T}} E_i(V_i, V_t), \quad t \equiv i-1 \,(\text{mod}\, T), \quad \mathcal{T} = \{0, 1, \ldots, T-1\}$$

subject to

$$\begin{aligned}
\left(1 - \frac{ke_i}{2}\right) V_t - \left(1 + \frac{ke_i}{2}\right) &\geq q_i - I_i + e_i c, && i \in \mathcal{T},\ t \equiv i-1 \,(\text{mod}\, T),\\
\left(1 - \frac{ke_i}{2}\right) V_t - \left(1 + \frac{ke_i}{2}\right) &\leq Q_i - I_i + e_i c, && i \in \mathcal{T},\ t \equiv i-1 \,(\text{mod}\, T),\\
V_i &\geq rV_{\max} && i \in \mathcal{T},\\
V_i &\leq V_{\max} - V_i && i \in \mathcal{T},
\end{aligned}$$

where E_i denotes an evaporation loss function of the reservoir volume V_i, I_i denotes the net volume inflow, e_i denotes the evaporation rate, and q_i and Q_i denote respectively lower and upper bounds on the expected amount of water during period i. Furthermore, k is a constant of proportionality between reservoir volume and water surface area, c is an offset constant, r denotes a safety risk factor between zero and one, and $V_{\max}$ denotes the maximum reservoir storage capacity, before overflow occurs. According to Cheng and Rezicek [5], deterministic models which are based on average or mean stream flows may, however, result in overly optimistic release policies.

Strauss [8] criticised the simplistic manner in which Van Vuuren & Grundlingh [10] accommodated risk by only including a minimum reserve for the operating level of the reservoir, while not directly allowing for the possibility of unmet demand. Strauss implemented a very similar model, together with the additional risk-related constraint

$$\alpha D_i \leq B_i \leq \min\{(1+\alpha)D_i, B^{\max}\}, \quad i \in \mathcal{T},$$

where D_i and B_i denote respectively the demand and release during period i, to ensure that demand is met to within a certain variation parameter α. It is important to note that risk was, however, not quantified in the model of Strauss.

Quantifying the representation of risk is a crucial element lacking in the two models reviewed above. Furthermore, models which provide a single optimal solution may be insufficient, since reservoir operation commonly involves tradeoff decision options, as mentioned above. As a response, multi-objective models have often been implemented in the context of reservoir management, as mentioned. This works well for complex, multi-purpose reservoir systems with multiple decision variables and complex benefit functions, whereas for single-purpose reservoirs, only two directly conflicting objectives exist, and the decision options depend on a single variable, namely sluice control.

3 Modelling Assumptions

The farmers who benefit from Keerom Dam are in agreement that the release of more water (up to maximum sluice capacity, thus not including floods) is more beneficial than

the release of less water. It may generally be assumed that the benefit function for normal operation of an irrigation reservoir is a strictly increasing function. This assumption makes the problem of determining a suitable release strategy fairly simple: release the maximum amount of water, keeping in mind the risk of not being able to achieve repeatability of the strategy over successive hydrological years. The focus of a reservoir release model should therefore be on quantifying the risk related to a given release strategy, rather than merely searching for an optimal strategy or set of strategies.

In order to develop a mathematical model for irrigation reservoir operation which incorporates a quantification of risk, the following assumptions are made:

I *Irrigation reservoir.* The model will be designed for use in the context of open-air reservoirs used for irrigation purposes only. Release strategy formulation will not be considered for reservoirs designed for other uses (such as for the storage of drinking water or the generation of hydro-electricity).

II *Time continuum discretization.* The scheduling horizon over which a release strategy is to be determined will be discretized into a number of time intervals, called *decision periods*, which are typically weeks or fortnights.

III *Evaporation rate.* The evaporation rate during a given decision period will be considered to be directly proportional to the average exposed water surface area of the reservoir and thus a function of the average reservoir volume during that period. The coefficient of proportionality will be taken to depend on the historical meteorological conditions of the time interval in question. For relatively short decision periods (*e.g.* weeks) this is a realistic assumption. The South African Department of Water Affairs and Forestry keeps a database of all water reservoirs exceeding a certain minimum storage capacity, which includes historical daily evaporation losses. For new reservoirs, the evaporation rates of older reservoirs in the vicinity may be used as an initial estimate.

IV *Repeatability.* The repeatability of a given water release strategy will be taken to depend on an estimate of the most likely reservoir volume at the end of the hydrological year as a result of this strategy. Thus, the reservoir volume during the transition between hydrological years is used as the measure of future repeatability.

V *Silting rate.* The silting rate is the rate at which the reservoir's capacity is reduced by the accumulation of sediments in the reservoir. This rate will be considered negligibly small. We consider the formulation of release strategies over planning horizons not exceeding one year in this paper. The notion of repeatability is incorporated to ensure that a reasonable starting volume is available at the start of the following year, when planning will commence for that year. Changes in reservoir volume due to silting can thus be taken into account on an annual basis. For such scheduling windows this is a realistic assumption.

VI *Seepage rate.* Seepage is the escape of water into the reservoir floor. Because it is very difficult to measure this kind of water loss rate separately, it is assumed that the seepage rate is included in the reservoir's net influx.

VII *Demand.* The water demand during a specific decision period is assumed to be constant. For relatively short decision periods (*e.g.* weeks) this is a realistic assumption.

VIII *Conservation law.* It is assumed that the change in water volume during a given decision period equals the net influx (including all the reservoir's water sources, precipitation into the reservoir and its catchment area, as well as seepage losses), less evaporation losses and all reservoir outflows, including controlled sluice outflow and overflow.

4 Modelling Framework

The framework depicted in Figure 1 partitions the inputs of the model into historical data, as well as various user-inputs and reservoir-related parameters. Historical data refer to past inflows, to be utilised in the simulation of future inflows, and past evaporation losses used to estimate the coefficient of proportionality of evaporation during any given decision period.

The required user-inputs are the decision period length (typically weekly or biweekly), the number of remaining decision periods in the hydrological year, the current reservoir volume and some target end-of-hydrological-year volume (a target volume corresponding to a certain level of risk should be computed and suggested to the user, but this value should be user-adjustable). The weekly demand profile, which may be computed using standard irrigation decision support software, such as *CROPWAT* [9], is also considered a user-input.

Reservoir-related parameters include the maximum sluice release capacity, the minimum allowed release volume per decision period according to legal requirements, the reservoir's storage capacity and its shape characteristic, which relates the water level, stored water volume and exposed water surface area of the reservoir. The processes and information flows in Figure 1 are described in some detail in the following sections. The process to be conducted first, before the model execution, is the simulation of inflows.

5 Monte Carlo Simulation of Inflows

The historical inflows will be used to obtain the cumulative inflow distribution for each simulation period in the year. The simulation period length may be daily, weekly, biweekly or monthly, but must be shorter than or equal to the decision period length. Visualising the cumulative distribution plot, historical net inflow for a given period may be placed in bins on a horizontal axis, with each bin containing the number of historical inflows less than the bin's upper limit, and where the vertical axis then denotes the number of inflow data points. The vertical axis may be normalised, to represent portion of total inflows.

The South African Department of Water Affairs and Forestry's database includes daily inflows, typically resulting in ample historical data. This allows for accurate distributions to be obtained. In other words, it is usually not required to fit a theoretical distribution to the data — the empirically obtained distributions may be utilised directly. For a

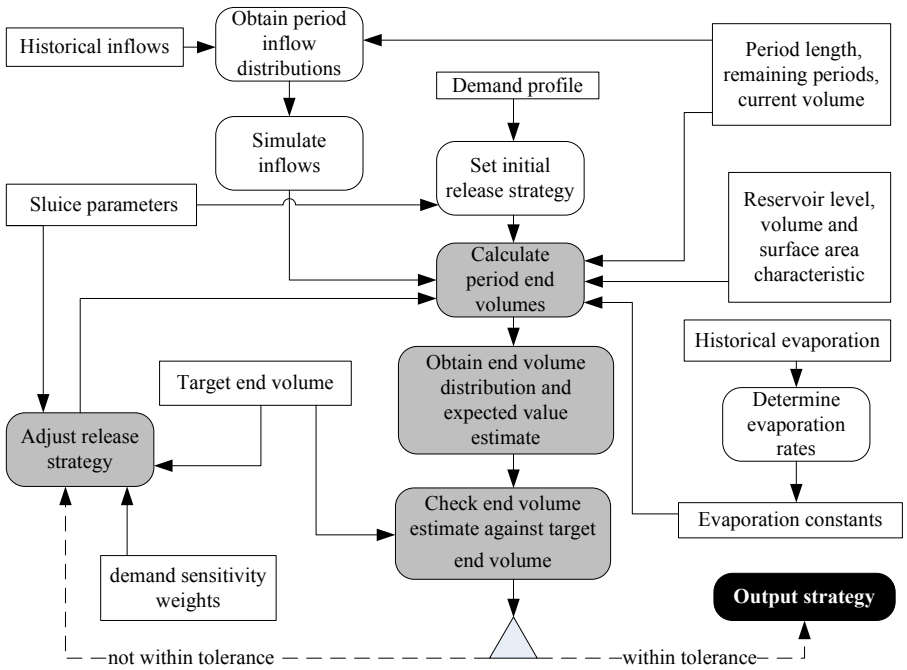


Figure 1: *A framework depicting the processes and flows of information in the proposed model.*

newly built reservoir the prediction of future inflows, without any historical data, would constitute a challenging separate research project.

Based on these distributions, the Monte Carlo method may be utilised in conjunction with the inverse transform method, to simulate inflows. The inverse transform method relies on the probability integral transformation, as described by Rizzo [6].

Let I_t be the net reservoir inflow during decision period t and let U be a uniform random variable on the interval $[0, 1]$. If I_t has the cumulative distribution function F_{I_t}, then $F_{I_t}^{-1}(U)$ has the same distribution as I_t. An instance of I_t may therefore be simulated according to the inverse transform method by generating a uniform $[0, 1]$ variate u and recording the value $F_{I_t}^{-1}(u)$. Once this has been done for each simulation period, the inflows of a single hydrological year have been simulated. A large number of parallel years (one thousand, for example) may thus be simulated.

6 Obtaining Period Volume Distributions

Let V_t denote the reservoir volume at the end of decision period t, x_t the water volume released during decision period t, E_t the volume of water lost due to evaporation during decision period t, e_t the evaporation rate per unit of average exposed water surface area during decision period t, and A_t the exposed water surface area at the end of decision period t, where $t \in \mathcal{T}$, for some set $\mathcal{T} = \{0, 1, \ldots, T-1\}$ of decision periods. According to

Assumption VIII of §3 it follows that

$$V_t = V_{t-1(\mathrm{mod}\,T)} + I_t - x_t - E_t, \quad t \in \mathcal{T},$$

while according to *Assumption III* it follows that

$$E_t = e_t \left(\frac{A_{t-1(\mathrm{mod}\,T)} + A_t}{2} \right), \quad t \in \mathcal{T}.$$

Furthermore, the exposed surface area of the reservoir is related to the stored water volume according to some reservoir shape characteristic f in the sense that

$$A_t = f(V_t), \quad t \in \mathcal{T}.$$

A preliminary release strategy may be determined according to the demand profile and the sluice release parameters. Let D_t denote the water demand during decision period t and let $x_{\min}$ and $x_{\max}$ denote respectively the minimum and maximum possible release volumes during any decision period. Then

$$x_t = \begin{cases} x_{\max} & \text{if } D_t \geq x_{\max}, \\ D_t & \text{if } D_t \in (x_{\min}, x_{\max}), \\ x_{\min} & \text{if } D_t \leq x_{\min}, \end{cases}$$

for all $t \in \mathcal{T}$.

Using the current reservoir volume, the stochastically simulated inflows for the remaining decision periods of the hydrological year and the preliminary release strategy described above, a distribution of the reservoir volume at the end of the hydrological year, resulting from this strategy, may be obtained. This distribution may be analysed, using standard statistical methods for inference, to obtain an estimate of the expected reservoir end volume

$$\mu_{V_T} = \frac{\sum_{j=1}^{K} V_j^*}{K} \approx \sum_{i=1}^{S} y_i \overline{V_i^*}, \tag{1}$$

where K denotes the number of simulation replications and V_j^* denotes the end volume obtained by simulation replication j. This estimate may be compared to the expected end volume, as obtained from the distribution, to asses the effect of the distribution's bin sizes. In (1), S denotes the number of bins used to obtain the empirical distribution, y_i denotes the proportion of end volumes residing within bin number i, and $\overline{V_i^*}$ represents the midpoint between the upper and lower limits of bin i.

The estimate μ_{V_T} will be compared to a target end volume specified by the decision maker. If the estimator falls outside a certain tolerance band centred around the target value (also specified by the decision maker), the release strategy should be adjusted with the aim of centring the end volume distribution on the target value.

In our model, seven factors are taken into account for this adjustment: the number of remaining decision periods, denoted by n, the end volume estimate μ_{V_T} in (1), the target end volume, denoted by V_A^*, a tolerance $\alpha \in (0, 1]$ within which the target end volume should be met, the minimum and maximum sluice release parameters, and user-specified

weight factors which represent each demand period's sensitivity to not meeting water demand for that period, denoted by $w_t \in [0,\ 1]$, where a lower value represents a more sensitive period. Let μ_w denote the mean of the user-defined weight factors. Our proposed adjustment process of the preliminary release strategy is iterative in nature. Each iteration of this process is accomplished in two stages. First

$$x_t' = x_t + \left(\frac{\mu_{V_T} - V_A^*}{n}\right) \times \frac{w_t}{\mu_w}$$

is computed, after which

$$x_t'' = \begin{cases} x_{\max} & \text{if } x_t' \geq x_{\max}, \\ x_t^{'} & \text{if } x_t' \in (x_{\min},\ x_{\max}), \\ x_{\min} & \text{if } x_t' \leq x_{\min} \end{cases}$$

is determined for each remaining decision period t. Periods with sensitive demand are also expected to be periods of higher demand. By definition, the first strategy satisfies demand as well as possible, and thus larger releases are made during sensitive periods, possibly equalling the maximum release capacity. If the end volume estimate exceeds $(1+\alpha)V_A^*$, more water will be released during less sensitive periods than sensitive ones, so as to expedite the centring of the distribution. If the end volume estimate is less than $(1-\alpha)V_A^*$, the water volume released during non-sensitive decision periods will be decreased in greater proportions than the water volume released during sensitive periods, so as to ensure that demand is met as best possible.

During each iteration of this adjustment procedure, the end volume distribution is re-calculated and a new end volume estimate obtained, closer to the target value, until the estimate falls within the interval $[(1-\alpha)V_A^*, (1+\alpha)V_A^*]$. Once the distribution is thus centred on the target value, the current release strategy may be used as starting point for developing tradeoff strategy suggestions.

7 Quantifying Risk

According to *Assumption VII* of §3, the repeatability of a strategy is taken to depend on the reservoir volume during the transition between hydrological years. The probability of water shortage associated with reservoir volumes during this transition may be determined by equating the starting and target volumes in the model, for a simulated hydrological year, and solving the model for a range of volumes. The number and sizes of shortage occurrences may then be obtained by comparing the release strategy for a given reservoir transition volume with the demand profile. An expected percentage of unmet demand may then be associated with a given transition volume.

In any period during an actual hydrological year, the probability of ending at a certain volume may be obtained from the end volume probability distribution resulting from the current release strategy. The percentage of expected unmet demand during subsequent years, related to this volume, will have been obtained by the approach described in the previous two sections. The risk of not meeting future hydrological years' demand as a result of a current strategy, can therefore be expressed numerically.

In the case of a dry year, when the notion of risk requires special attention, tradeoff decisions between the fulfilment of the current year's demand and future repeatability may be required. Future repeatability here refers to a level of confidence in the ability to repeatedly fulfil the irrigation demands of subsequent years. If, due to particularly low reservoir storage levels during a dry year, the proposed strategy which centres the end volume distribution on the specified target value for repeatability to within the acceptable tolerance interval fails to meet the current year's demand adequately, the decision maker may prefer to aim for a lower target ending volume. The fulfilment of the current year's demand will thereby be improved, but at the cost of a decrease in the security of future years' water supply. The improvement in demand fulfilment, as well as the decrease in security, may finally be quantified and compared.

8 Conclusion

A water release model, which may form the basis of a DSS for open-air irrigation reservoir operation, was proposed in this paper. The model may be used to quantify the risks associated with annual repeatability and expected water shortages resulting from a specific reservoir release strategy. In addition, the model may be used to generate release strategies with the aim of ending the hydrological year at a certain reservoir storage level, thereby allowing a decision maker to review and compare tradeoff strategy choices.

The new model proposed in this paper forms part of a larger, ongoing research project on reservoir release strategy management at Stellenbosch University. Further work will include incorporating the model proposed here into a computerised decision support system and validating the decision support system by applying it to a special case study involving Keerom Dam.

References

[1] Butcher WS, 1971, *Stochastic dynamic programming for optimum reservoir operation*, Journal of the American Water Resources Association, **7(1)**, pp. 115–121.

[2] Huang G & Loucks D, 2000, *An inexact two-stage stochastic programming model for water resources management under uncertainty*, Civil Engineering Systems, **17(2)**, pp. 95–118.

[3] Jain S, Das A & Srivastava D, 1999, *Application of ANN for reservoir inflow prediction and operation*, Journal of Water Resources Planning and Management, **125(5)**, pp. 263–271.

[4] Reddy MJ & Kumar DN, 2006, *Optimal reservoir operation using multi-objective evolutionary algorithm*, Water Resources Management, **20(6)**, pp. 861–878.

[5] Reznicek K & Cheng T, 1991, *Stochastic modelling of reservoir operations*, European Journal of Operational Research, **50(3)**, pp. 235–248.

[6] Rizzo ML, 2008, *Statistical computing with R*, Chapman & Hall/CRC, London.

[7] Skiena SS, 2015, *Global Risks 2015 Report*, [Online], [Cited March 9th, 2015], Available from `http://www3.weforum.org/docs/WEF_Global_Risks_2015_Report15.pdf`.

[8] Strauss JC, 2014, *A decision support system for the release strategy of an open-air irrigation reservoir*, Final Year Industrial Engineering Undergraduate Project, Stellenbosch University, Stellenbosch.

[9] Swennenhuis J, 2006, *CROPWAT version 8.0*, [Online], [Cited April 30th, 2015], Available from `http://www.fao.org/nr/water/infores_databases_cropwat.html`.

[10] Van Vuuren JH & Gründlingh WR, 2001, *An active decision support system for optimality in open air reservoir release strategies*, International Transactions in Operational Research, **8(4)**, pp. 439–464.

[11] Yeh WW, 1985, *Reservoir management and operations models: A state-of-the-art review*, Water Resources Research, **21(12)**, pp. 1797–1818.

Proceedings of the 2015 ORSSA Annual Conference
pp. 44–53
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

An evaluation of self-organisation in traffic control with respect to varying distances between adjacent intersections in a road corridor

SJ Movius* JH van Vuuren†

Abstract

Traffic congestion is a major concern in most cities all over the world. The economy, health of the population and the environment are all affected negatively by heavily congested roads. A recently proposed solution to ease traffic congestion in busy road networks involves the implementation of self-organising signal control algorithms at signalised intersections. In particular, an algorithm inspired by the theory of inventory control, an algorithm inspired by the chemical process of osmosis and finally an algorithm that is a hybrid of the first two have recently been proposed for traffic control at signalised intersections. These algorithms seem to be promising in terms of reducing vehicle delays as well as the propagation of uninterrupted traffic flow, known as *green waves*, through adjacent intersections. These three self-organising algorithms have previously been compared to an existing fixed-time algorithm as well as to two other existing self-organising algorithms. This comparison took place in a simulated environment facilitating the calculation of a number of performance measure indicators in order to determine which algorithms were most effective. The results revealed that in a corridor road network, the hybrid algorithm outperformed the others overall, while in a grid road network, the osmosis-inspired algorithm was the most effective. These results were, however, obtained under the assumption that neighbouring intersections were equally spaced from one another, which is a highly unlikely occurrence in practice. In this paper, the algorithms are compared in a more realistic, simulated environment where neighbouring intersections lie at varying distances from one another. It is verified to what extent the self-organising algorithms still outperform the existing algorithms by facilitating the formation of so-called green waves of uninterrupted traffic flow through adjacent intersections in the context of this added complexity.

Key words: Traffic signal control, Self-organisation.

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 16446305@sun.ac.za

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

1 Introduction

The main cause of traffic congestion is the over-utilisation of roads which leads to dense, stop-and-go traffic [6]. A viable method of reducing traffic congestion involves optimisation of traffic signal control algorithms employed at signalised intersections. Improved traffic signal control may serve to dilute concentrated traffic in road networks by increasing the efficiency of signal duration times, leading to reduced vehicle delay and a more dispersed use of road networks. There are two main types of traffic signal control: *fixed-time control* and *vehicle-actuated control*. Fixed-time control consists of the specification of predetermined signal cycle times based on expected traffic flow densities for various times of the day. Since traffic volumes often fluctuate dramatically over the course of a single day, actual demand is typically not met by such predetermined cycle times and, as a result, green times are often either too long or too short. Vehicle-actuated control, on the other hand, is capable, at least to some extent, of adapting according to real-time traffic conditions, by selecting cycle phases and signal timings to best suit the current conditions of the road network. Unlike fixed-time control, vehicle-actuated control is responsive to changes in traffic flow, but requires the implementation of vehicle detection equipment in order to register the prevailing traffic conditions.

The optimisation problem associated with centralised traffic signal control is NP-hard and realistic instances of this problem typically cannot be solved in real time [7, 10]. The use of a decentralised traffic control system is advantageous as the problem of traffic control at each intersection may be viewed as an isolated problem, requiring no information on how signals are controlled at neighbouring intersections. The decentralised paradigm of *self-organisation* has been suggested as an appropriate approach toward developing effective traffic signal control algorithms. Self-organisation occurs when there is an increase in the order present in a system without any form of external control [3]. It may be an effective approach to traffic control, not only because it is decentralised, but also because it can lead to the natural *emergence* of coordination between intersections. Such emergence is observed when the system exhibits novel behaviour on a macro level as a result of interactions between system elements at the micro level [3]. Three self-organising traffic signal control algorithms recently proposed by Einhorn *et al.* [4] appear to be effective in terms of being capable of reducing vehicle delay time in road networks. None of these algorithms requires predetermined parameter values, but they require the use of radar detection equipment mounted at each intersection.

The first algorithm is inspired by the theory of inventory control and attempts to minimise the virtual costs associated with vehicle delay. The second algorithm is inspired by the chemical process of osmosis, which takes into account the push force of vehicles approaching an intersection as well as the pull force of the empty space on the other side of the intersection which may be occupied by vehicles. The third algorithm is a hybrid of the first two, attempting to exploit the best characteristics of both, while simultaneously maximising the intersection utilisation.

These three algorithms were compared by Einhorn *et al.* [4] in a specially designed simulated road network. The simulation model was, however, only able to accommodate road networks with equally spaced intersections, which is not a realistic assumption. The objective in this paper is to evaluate to what extent the algorithms are effective in reducing

vehicle delay time when intersections are at varying distances from one another. The reason for this evaluation is the anticipation that in order to generate so-called green waves of uninterrupted traffic flow through adjacent intersections along a road corridor, it is expected to be advantageous if the distances between successive intersections are uniform. The simulation test bed developed by Einhorn *et al.* [4] is generalised in this paper so as to allow for non-uniform spacings between successive intersections along a road corridor. The validity of the findings of Einhorn *et al.* [4] is then tested in this generalised simulation setting.

The paper is organised as follows. A brief literature review is conducted in §2 on the use of self-organisation in traffic signal timings, affording special attention to the three self-organising algorithms proposed in [4]. The design of our simulation experiment is presented in §3, after which the simulation results are reported and interpreted in §4. Some concluding remarks follow in §5 after which the the paper closes in §6 with a number of ideas with respect to possible future work related to the work reported here.

2 Literature review

A brief literature review is conducted in this section on the previous use of self-organisation in traffic control and its associated advantages and shortcomings.

2.1 Early use of self-organisation in traffic signal timings

The implementation of self-organisation in traffic signal control is a relatively new idea which has yet to be implemented in practice on a large scale. The self-organising approach towards traffic signal control brings with it multiple benefits, including short-term flexibility, long-term adaptability and reduced cost associated with computing time [5, 9]. Gershenson [5] states that traffic signal control is more of an adaption problem, rather than an optimisation problem, due to typically unpredictable changes in traffic volume. He proposed a self-organising traffic signal control algorithm called *SOTL-request*. This algorithm gives right-of-way preference to platoons of vehicles rather than to individual vehicles, in order to promote the uninterrupted migration of large groupings of vehicles through an intersection during a single signal phase. This is achieved by changing a red signal once the number of vehicles queued at an intersection reaches a certain threshold, while if this number falls below the threshold, the signal remains red for longer, allowing time for more vehicles to join the queue [2, 5]. A variation on this algorithm, called *SOTL-phase*, incorporates an additional minimum time constraint in order to prevent rapid switching of signals. A third algorithm, known as *SOTL-platoon* is similar to *SOTL-phase*, but with additional restrictions designed to regulate the size of the platoons that travel through the intersection. The latter algorithm was compared to the widely implemented SCATS algorithm [8] in a simulated environment and was shown to significantly outperform it [12]. The three self-organising algorithms mentioned above have been described as robust, but a common disadvantage of these algorithms is that a number of parameter values must be determined in order for them to work effectively. It is, however, not stated in the literature how to go about selecting appropriate parameter values for a given traffic scenario [4].

Another example of self-organisation in traffic signal control has been proposed by Lämmer and Helbing [7]. Their algorithm was inspired by the oscillatory changes in the counter-flow of pedestrians through a narrow bottleneck. The algorithm makes use of a strategy combining optimisation and stabilisation in the road network. The optimisation phase aims to minimise total vehicle delay experienced by vehicles that are within a certain distance from the intersection. In order to prevent queues from growing too long, the stabilisation phase attempts to keep the maximum vehicle queue length below a certain threshold. This algorithm was shown in [4] to be effective in reducing vehicle delay and outperformed existing state-of-the-art adaptive control schemes. It is also capable of outperforming any solution based on fixed-time control cycles provided that the parameter values of the algorithm are chosen judiciously [10].

2.2 The self-organising algorithms of Einhorn et al.

As mentioned in the introduction, three traffic signal control algorithms were proposed in [4], each inspired by a self-organising process.

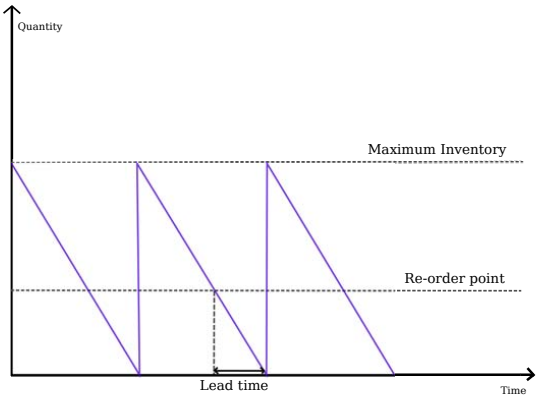


Figure 1: *Demand in the basic EOQ model.*

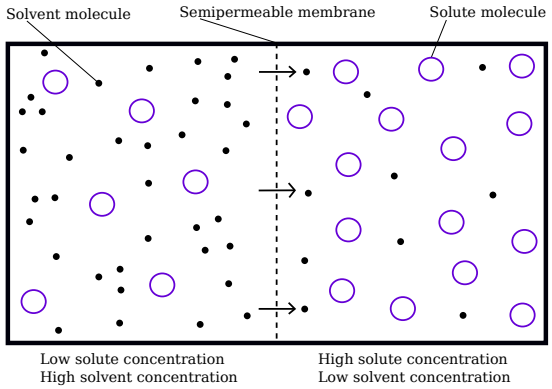


Figure 2: *The process of osmosis.*

2.2.1 An algorithm inspired by inventory control

The first of the algorithms by Einhorn *et al.* [4] is a self-organising algorithm known as the *inventory traffic signal control algorithm* (I-TSCA), which is based on the theory of inventory control and attempts to minimise a virtual cost associated with the total delay time of vehicles in a network. The *economic order quantity* (EOQ) model in inventory control requires the determination of the reorder point in time as well as the associated reorder quantity of a particular product held in inventory, as illustrated in Figure 1 [11]. In traffic control, these two variables correspond to the point in time at which signal switching must take place and the amount of green time allocated, respectively [4]. The algorithm functions by calculating the total cost associated with allocating green time to each signal phase, selecting the phase which results in the lowest total cost and awarding it green time. The use of radar detection technology at intersections is assumed as a prerequisite for implementing the algorithm so that the number of approaching vehicles can be observed, as well as their associated speeds and distances from the intersection,

in order to facilitate the calculation of the "demand" for each signal phase. In [4] it was found that the I-TSCA performed most effectively in uniformly spaced road corridors and city grid topologies under lighter traffic conditions, due to the fast switching propensity of the algorithm.

2.2.2 An algorithm inspired by osmosis

If two liquids of different solute concentrations are separated by a semipermeable membrane such that the solute molecules cannot pass through it, the solvent molecules of the liquid with a lower solute concentration pass through the membrane into the liquid with the higher solute concentration [1]. This is known as the *process of osmosis*, depicted in Figure 2, and results in a balance of the solute concentration on both sides of the membrane through pressure exertion. The presence of these pressures is the reason this process is ideal for implementation in traffic control.

The second self-organising algorithm by Einhorn *et al.* [4] is called the *osmosis traffic signal control algorithm* (O-TSCA). The algorithm likens the solvent molecules in the process of osmosis to vehicles approaching the intersection, which is analogous to the semipermeable membrane through which the solvent molecules (vehicles) pass. The solute molecules correspond to the empty space along the roadway on the other side of the intersection not occupied by vehicles. Vehicles approaching the intersection will exert a push pressure on the system in a manner similar to how solvent molecules are "pushed" through the semipermeable membrane in osmosis, while the empty space just beyond the intersection exerts a pull force on vehicles through the intersection to the empty road space on the other side of the intersection. Einhorn *et al.* [4] found that the O-TSCA performs well in uniformly spaced road corridors and city grid topologies under heavier traffic conditions as it tends to give preference to large platoons of vehicles, switching signals less often.

2.2.3 A hybrid algorithm

Under light traffic conditions it was found that the I-TSCA outperformed the O-TSCA as the latter algorithm tended to award green times that were too long under light traffic conditions, while the faster switching signals of the I-TSCA better suited these traffic conditions. Under heavy traffic conditions, on the other hand, the O-TSCA outperformed the I-TSCA, as mentioned. This is due to the long green times typically awarded by the O-TSCA under such conditions which are indeed necessary to alleviate heavy traffic density, while the I-TSCA switched signals too frequently. A third self-organising algorithm, known as HYBRID, was therefore proposed by Einhorn *et al.* [4], which incorporates both the I-TSCA and the O-TSCA in order to capitalise on the advantages of both algorithms. It employs these algorithms together with an *intersection utilisation maximisation supervisory mechanism* (IUMSM) in order to ensure that the intersection is not under-utilised as a result of green times that are too long or too short.

3 Simulation experiment

We generalised the simulation framework proposed in [4] to account for road networks that do not exhibit equally spaced intersections. Three different scenarios are considered in our simulation evaluation experiment within the paradigm of this generalisation. A corridor roadway comprising four equally spaced intersections is considered in the first scenario (as illustrated in Figure 3), while a corridor of four intersections is also considered in our second scenario, but with the difference that the distance between the middle two adjacent intersections is a multiple of those between the flanking pairs of intersections (as illustrated in Figure 4). Lastly, a corridor of four intersections is yet again considered in our third scenario in which intersections occur at uneven distances from one another (as illustrated in Figure 5).

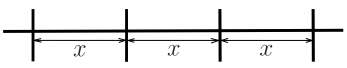


Figure 3: *Scenario 1.*

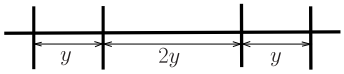


Figure 4: *Scenario 2.*

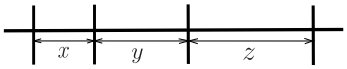


Figure 5: *Scenario 3.*

In each of these three experiments, ten simulation replications are performed under the exact same arrival conditions in order to ensure a fair comparison. In contrast, Einhorn *et al.* [4] performed 30 simulation replications in their experiments, but we found that the difference between the averages of the five *performance measure indicators* (PMIs) over ten replications and over 30 replications was not significant. We adopt the same five PMIs used by Einhorn *et al.* [4], namely, the average mean and maximum vehicle delay time, the average normalised vehicle delay time, the average mean number of vehicle stops made and the average mean normalised number of vehicle stops made.

4 Simulation results

In this section, the performance results of each of the three self-organising algorithms proposed by Einhorn *et al.* [4] are presented for various traffic flow densities for the three scenarios described in §3.

4.1 Scenario 1

The scenario in Figure 3 was already implemented and evaluated by Einhorn *et al.* [4], but the implementation was repeated in this paper for validation purposes. The five PMI values achieved by the three algorithms of Einhorn *et al.* [4] over the ten simulation runs for Scenario 1 are shown in Tables 1 and 2 for light and heavy traffic flow conditions, respectively.

Under light traffic conditions the hybrid algorithm was found to be the best performing algorithm, with a normalised delay time of 1.18 — significantly outperforming I-TSCA and O-TSCA. Although O-TSCA causes longer delay times than Hybrid, O-TSCA is the best performing algorithm with respect to the mean number of stops made in the system. The mean maximum delay experienced by vehicles under O-TSCA was 106.35 — a significantly larger value than those achieved by both Hybrid and I-TSCA.

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	13.94	17.70	11.80
Normalised delay time	1.22	1.29	1.18
Mean number of stops	1.23	0.81	1.00
Normalised number of stops	0.81	0.55	0.62
Average maximum delay time	95.31	106.35	89.52

Table 1: *PMI results for each algorithm under light traffic conditions in Scenario 1.*

In roadways experiencing a heavier flow of traffic, Hybrid was again the most effective in terms of achieving the lowest average mean delay time of 21.87 seconds, while O-TSCA and I-TSCA obtained corresponding values of 23.67 and 24.63, respectively. Once again, O-TSCA was the most effective with respect to the mean number of stops made, significantly outperforming both I-TSCA and hybrid. O-TSCA also obtained the lowest average maximum delay time of 108.83. Since it was the worst performing algorithm in terms of maximum delay time under lighter traffic conditions, the simulation results suggest that O-TSCA is more efficient for Scenario 1 under heavier traffic.

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	24.63	23.67	21.87
Normalised delay time	1.39	1.41	1.34
Mean number of stops	1.36	0.92	1.07
Normalised number of stops	0.89	0.65	0.67
Average maximum delay time	131.73	108.83	122.69

Table 2: *PMI results for each algorithm under heavier traffic conditions in Scenario 1.*

4.2 Scenario 2

The five PMI values achieved by the three algorithms of Einhorn *et al.* [4] over the ten simulation runs for Scenario 2 are shown in Tables 3 and 4 for light and heavy traffic flow conditions, respectively.

Under light traffic conditions, the performances of Hybrid and I-TSCA are similar in Scenario 2 to those in Scenario 1, but the performance of O-TSCA is significantly worse. Its average delay time is longer than in the previous scenario although the difference is not statistically significant at a 95% confidence interval. The average maximum delay time of O-TSCA, however, is 199.15, which is significantly worse than the corresponding value of 106.35 achieved by the algorithm in the first scenario.

In heavier traffic conditions, once again, there is very little difference between the performance of hybrid and I-TSCA in this scenario, and with Scenario 1. The performance of O-TSCA worsens (as it did under lighter traffic conditions) as the distances between intersections become non-uniform. The average maximum delay time, for example, increased from 108.83 in the equidistant network of Scenario 1 to 151.60 in the road corridor of Scenario 2.

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	13.84	23.27	11.90
Normalised delay time	1.22	1.33	1.18
Mean number of stops	1.22	0.86	1.00
Normalised number of stops	0.79	0.51	0.62
Average maximum delay time	90.97	199.15	91.34

Table 3: *PMI results for each algorithm under light traffic conditions in Scenario 2.*

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	23.66	29.14	21.81
Normalised delay time	1.38	1.45	1.34
Mean number of stops	1.31	1.07	1.07
Normalised number of stops	0.86	0.68	0.67
Average maximum delay time	129.23	151.60	123.90

Table 4: *PMI results for each algorithm under heavier traffic conditions in Scenario 2.*

4.3 Scenario 3

The five PMI values achieved by the three algorithms of Einhorn *et al.* [4] over the ten simulation runs for Scenario 3 are shown in Tables 5 and 6 for light and heavy traffic flow conditions, respectively.

Under low traffic flow densities, the O-TSCA is the only algorithm performing significantly worse in this scenario than it did in Scenario 1. While the mean vehicle delay time is not dramatically different from that in the second scenario, it is 25.94, while originally it was 17.70 in Scenario 1. The average maximum delay time significantly worsened from 106.35 in Scenario 1 (as well as from 199.15 in Scenario 2) to 230.88 in Scenario 3.

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	13.86	25.94	11.80
Normalised delay time	1.22	1.34	1.18
Mean number of stops	1.23	0.84	1.00
Normalised number of stops	0.80	0.48	0.62
Average maximum delay time	94.28	230.88	89.52

Table 5: *PMI results for each algorithm under light traffic conditions in Scenario 3.*

Under heavier traffic conditions, Hybrid and I-TSCA once again did not yield results that are significantly different from those in Scenario 1. O-TSCA's PMIs are all similar to those of O-TSCA in Scenario 2, obtaining a delay time of 27.41 and average maximum delay time of 148.51. It is also noted that its average maximum delay time is significantly less than the corresponding value under lighter conditions.

Algorithm	I-TSCA	O-TSCA	Hybrid
Mean delay time	24.07	27.41	21.99
Normalised delay time	1.38	0.80	1.34
Mean number of stops	1.34	1.02	1.08
Normalised number of stops	0.87	0.67	0.68
Average maximum delay time	132.02	148.51	130.31

Table 6: *PMI results for each algorithm under heavier traffic conditions in Scenario 3.*

5 Conclusion

From the results in §4 we deduce that varying distances between intersections only seems to have a significant impact on the performance of O-TSCA. This indicates that the effectiveness of this algorithm relies on the equal spacing of intersections in order to perform effectively. Under light traffic conditions we observed that both the mean delay time and the average maximum delay time worsen with a loss of uniformity of intersection spacing along the corridor. However, under heavier traffic conditions the results yielded by O-TSCA in Scenarios 2 and 3, while worse than those in Scenario 1, do not differ significantly from one another. It was already known prior to this study that O-TSCA performs better under heavier traffic conditions. This may well be the reason why the PMI-values of the algorithm do not vary much from Scenario 2 to Scenario 3 under heavier traffic conditions. O-TSCA is not suited for use under light traffic conditions, which may be the cause of the worsening results that come with the loss of uniformity in the network spacing.

6 Future work

There are a number of ways in which this work can be taken further. The effectiveness of the three self-organising algorithms considered in this paper may be measured when taking into account a broad range of realistic scenarios. First, a similar experiment may be carried out in the context of a road network consisting of a grid of intersections, rather than merely a corridor roadway. Secondly, pedestrian phases may be incorporated into the signals timings at intersections, which is another way in which to add realism to the simulation model. Finally, the size of the network may be increased in order to observe how the algorithms perform in differently sized road networks.

References

[1] Cath TY, Childress AE & Elimelech M, 2006, *Forward osmosis: Principles, applications, and recent developments*, Journal of Membrane Science, **281(1)**, pp. 70–87.

[2] Cools SB, Gershenson C & D'Hooghe B, 2008, *Self-organizing traffic lights: A realistic simulation*, Springer, London.

[3] De Wolf T & Holvoet T, 2005, *Emergence versus self-organisation: Different concepts but promising when combined*, Engineering Self-Organising Systems, **3464**, pp. 1–15.

[4] Einhorn MD, Burger AP & Van Vuuren JH, *Self-organising traffic control inspired by inventory theory and the process of osmosis*, European Journal of Operational Research, Submitted.

[5] Gershenson C, 2008, *Self-organizing traffic lights*, arXiv preprint nlin/0411066, pp. 1–12.

[6] Goodwin P, 2004, *The economic costs of road traffic congestion*, The Rail Freight Group, University College London, London.

[7] Lämmer S & Helbing D, 2008, *Self-control of traffic lights and vehicle flows in urban road networks*, Journal of Statistical Mechanics: Theory and Experiment, **2008(04)**, pp. 1–30.

[8] Lowrie P, 1982, *The Sydney coordinated adaptive traffic system — Principles, methodology, algorithms*, Proceedings of the 207$^{\text{th}}$ International Conference on Road Traffic Signalling, London.

[9] Placzek B, 2014, *A self-organizing system for urban traffic control based on predictive interval microscopic model*, Engineering Applications of Artificial Intelligence, **34**, pp. 75–84.

[10] Szklarski J, 2010, *Cellular automata model of self-organizing traffic control in urban networks*, Bulletin of the Polish Academy of Sciences: Technical Sciences, **58(3)**, pp. 435–441.

[11] Winston WL, 2004, *Operations research: Applications and algorithms*, 4th Edition, Cengage Learning, Belmont (CA).

[12] Zhang L, Garoni TM & De Gier J, 2013, *A comparative study of macroscopic fundamental diagrams of arterial road networks governed by adaptive traffic signal systems*, Transportation Research Part B: Methodological, **49**, pp. 1–23.

Proceedings of the 2015 ORSSA Annual Conference
pp. 54–61
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

An integer linear programming formulation for collateral optimisation

PG Reynolds* SE Terblanche†

Abstract

Collateral management, driven by regulatory pressure, is becoming an ever more complex area requiring sophisticated systems and technology. Collateral optimisation aims to optimise funding costs and balance sheet utilisation when allocating assets to meet a range of liabilities. This process can be modelled as an optimisation problem with the objective to minimise the cost of the posted collateral whilst meeting all collateral calls and satisfying a range of constraints such as collateral eligibility criteria. In practice, constraints such as lot sizes, concentration limits and cardinality, result in a problem that becomes difficult to solve in a reasonable amount of time. This paper investigates some of the practical restrictions in collateral optimisation and presents an integer linear programming formulation of the problem.

Key words: Finance, Collateral management, Collateral optimisation, Integer linear programming

1 Introduction

In the investment industry, *collateral* refers to assets that are used to secure a lending transaction that are forfeited in the event of default. Collateral serves as a form of guarantee that if the borrower is unable to pay the lender, the lender has the right to sell the collateral to recover the outstanding amounts owed by the borrower. In finance, trading in *over-the-counter* (OTC) derivatives markets creates counterparty credit exposure. The *mark-to-market* (MTM) value of a transaction varies over the lifetime of the deal. If the transaction has a positive MTM value and the counterparty to the deal defaults, expected profits on the deal may be lost. Collateral is taken as security against non-payment or default by a counterparty and collateralisation serves as a form of bilateral insurance that is used to mitigate counterparty credit risk.

In the wake of the financial crisis new regulations were put in place to reduce systemic counterparty risk and are significantly altering the way OTC derivatives are cleared, settled, collateralised and reported. The *European Markets Infrastructure Regulation* (EMIR)

*School of Computer, Statistical and Mathematical Sciences, North-West University, Private Bag X6001, Potchefstroom, 2520, South Africa

†Centre for Business Mathematics and Informatics, North-West University, Private Bag X6001, Potchefstroom, 2520, South Africa, email: fanie.terblanche@nwu.ac.za

states that financial counterparties within the G20 economies will be obliged to clear certain classes of OTC derivatives through a *Central Clearing Counterparty* (CCP) [6]. Similarly, in the United States the Wall Street Reform and Consumer Protection Act, known as the Dodd-Frank Act, imposes mandatory initial and variation margins that will increase the value of collateral held against these trades [7]. Collateral eligibility standards are getting stricter under Basel III and Solvency II, increasing the demand for highly liquid, high quality collateral such as major currencies and government bonds [9]. Non-cleared bilateral trades will also be subject to higher capital adequacy requirements [2].

Within financial institutions the collateral management function is the area of the organisation responsible for reducing credit risk in unsecured financial transactions. On a daily basis, the collateral management team track the MTM exposures for the various counterparties, make and receive margin calls, process settlements for the transfer of cash and securities, manage reconciliations and resolve disputes. Collateral management has evolved rapidly over the last couple of years, driven by regulatory pressures, and is becoming an ever more complicated area requiring sophisticated systems and technology across interrelated divisions within the bank. It is no longer seen as an ancillary back office function but rather there is a movement towards integrated collateral management within the front office. As the cost of collateral increases and given the amount of collateral flows within an investment bank on a daily basis, optimally managing one's collateral and posting collateral more efficiently than one's counterparties can have a significant impact on overall profit.

Many industry "white papers" have been published online, describing optimisation as part of the collateral management process, but there appears to be limited academic literature dealing specifically with the collateral optimisation problem. The driving forces behind collateral optimisation and techniques that can be used to optimise collateral are outlined by Seagroatt [10]. A good introduction to the collateral allocation problem is given by Allen and Hellaby [1] and numerical methods are mentioned. Optimally allocating collateral to loans is discussed by Gürtler *et al.* [8] who formulate an allocation problem for minimising regulatory capital requirements under Basel II. Collateral optimisation, re-use and the transformation of collateral in the Dutch financial sector is discussed by Capel and Levels[4] who also highlight the increase in operational risks due to the inherent complexity in collateral optimisation processes and systems.

Applying real-world restrictions with integer-based constraints has been extensively studied in relation to Markowitz' mean-variance portfolio optimisation model [5]. Bonami and Miguel [3] use integer variables to model the need to diversify investments, the non-profitability of small investments and lot size constraints. This paper aims to apply similar techniques to the collateral optimisation problem. Section 2 gives an overview of the problem and Section 3 presents a formulation of the model. Results are discussed in Section 4 with Section 5 presenting a summary and conclusion.

2 Collateral Optimisation

OTC derivatives transactions, where trading is done directly between two counterparties, as opposed to trading on an exchange, are usually done under an ISDA[1] Master Agreement. This agreement serves as a contract between the parties where they agree general terms governing future transactions in order to facilitate the prompt negotiation of future transactions and a focus on only deal-specific differences. The agreement will specify netting rules so that the calculation of exposures can take place on a net basis and the netting of payments allows only a single payment to be exchanged between counterparties. The master agreement includes a *Credit Support Annex* (CSA) that regulates the credit support (collateral) for these derivative transactions. The CSA documents the negotiated margin and collateral terms, including thresholds, minimum transfer amounts, eligible collateral, *etc.*

A simplified example is considered in Appendix A where a party engages in OTC derivatives trading with two counterparties as outlined in Table 1. For this example these trades create net positive exposures for the counterparties (Table 2) and thus they have the right to call for collateral. These transactions are covered by the CSAs listed in Table 3. The CSA will specify the type of collateral that is allowed to be posted, for example government bonds with a remaining maturity greater than a year. Given these eligibility criteria and the attributes of the securities (Table 4) we create an eligibility matrix (Table 5). This matrix form allows for the use binary digits to specify whether collateral is eligible or not when formulating the constraints.

Thus the party will need to draw from the available assets in their collateral inventory (Table 6) to meet the margin call requirements. In essence the collateral optimisation problem is that of optimally allocating assets to the various calls in such a way as to minimise the overall cost of the collateral being posted out and maximising the value of the collateral being retained in the collateral inventory (Figure 1).

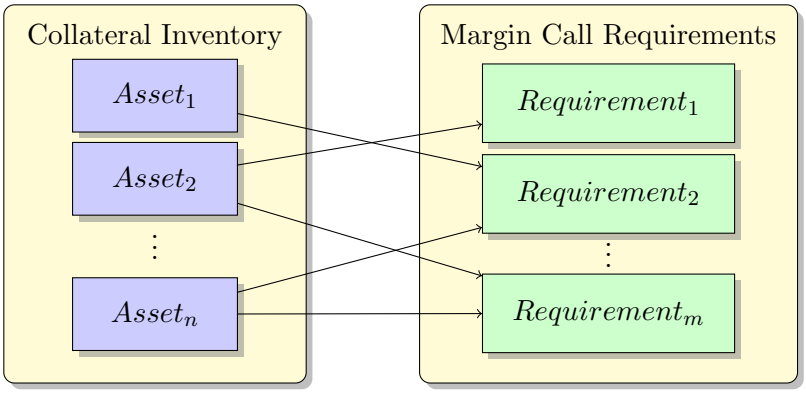


Figure 1: *Asset allocation.*

[1]International Swaps and Derivatives Association (ISDA)

3 Formulation

The objective is to minimise the value of posted collateral when allocating assets from the collateral inventory to meet the margin call requirements. Let $\mathcal{A}$ be the index set of assets in the collateral inventory, let $\mathcal{R}$ be the index set of requirements that we are obliged to meet and let $\mathcal{I}$ be the index set of issuers. In order to filter assets from a particular issuer $i \in \mathcal{I}$, we make use of the index set $\mathcal{A}(i)$.

Let $x_{ar} \in \mathbb{Z}_+$ be a decision variable that represents the number of units of an asset $a \in \mathcal{A}$ allocated to a margin call requirement $r \in \mathcal{R}$. Let p_a be the price per unit for each asset and let e_{ar} be the eligibility of an asset (Table 5) to be posted as collateral for a specific requirement. We introduce a preference weighting w_a that acts as an additive penalty on assets in order to keep these in the collateral inventory. For example, we may prefer to retain high quality liquid assets. Let l_a represent the fact that assets are usually only allocated in discrete lot sizes. The resulting collateral optimisation problem is given by

$$\text{minimise} \sum_{a\in\mathcal{A}} \sum_{r\in\mathcal{R}} x_{ar} p_a w_a l_a$$

$$\text{subject to} \quad \sum_{r\in\mathcal{R}} x_{ar} l_a \leq u_a, \qquad \forall a \in \mathcal{A}, \tag{1}$$

$$\sum_{a\in\mathcal{A}} x_{ar} e_{ar} p_a l_a \geq v_r, \qquad \forall r \in \mathcal{R}. \tag{2}$$

Availability constraints (1) prevent the allocation of more units of an asset than are available in the inventory, where u_a is the number of available units for each asset (Table 6).

Requirement constraints (2) ensure that the value of the posted collateral at least meets the value of the required amount v_r of the margin call (Table 2).

In addition to the lot size defined previously, we introduce a minimum lot size constraint where m_a denotes the minimum number of units of an asset that can be allocated to meet a requirement. Let y_{ar} be a binary indicator variable such that

$$y_{ar} = \begin{cases} 1, & \text{if asset } a \text{ is allocated to requirement } r, \\ 0 & \text{otherwise.} \end{cases}$$

The minimum lot size constraint is

$$y_{ar} m_a \leq x_{ar} l_a \leq y_{ar} u_a. \tag{3}$$

The logic employed in (3) ensures that if y_{ar} is 1, then $x_{ar} \geq m_a$, while if y_{ar} is 0 then x_{ar} is forced to 0.

In order to reduce fragmentation, by preventing the allocation of a large number of different assets to a specific call, the number of assets assigned can be constrained not to exceed a

level C_r by incorporating the constraint

$$\sum_{a \in \mathcal{A}} y_{ar} \leq C_r, \forall r \in \mathcal{R}. \tag{4}$$

Concentration limits could be added to diversify the composition of assets that are retained in the collateral inventory by industry sector or issuer. Consider the diversification constraints that limit the allocation of assets from a specific issuer when meeting a particular call while distinguishing between two levels of diversification.

Asset level diversification places a limit d_{ar} on the proportion of an asset a that is allowed to be assigned to requirement r

$$\frac{x_{ar} e_{ar} p_a l_a}{\sum\limits_{k \in \mathcal{A}} x_{kr} e_{kr} p_k l_k} \leq d_{ar}, \forall a \in \mathcal{A}, \forall r \in \mathcal{R}. \tag{5}$$

Issuer level diversification places a limit d_{ir} on the proportion of assets from issuer i that are allowed to be assigned to a requirement r

$$\frac{\sum\limits_{a \in \mathcal{A}(i)} x_{ar} e_{ar} p_a l_a}{\sum\limits_{k \in \mathcal{A}} x_{kr} e_{kr} p_k l_k} \leq d_{ir}, \forall i \in \mathcal{I}, \forall r \in \mathcal{R}. \tag{6}$$

4 Results

The formulation was tested in an implementation using the IBM ILOG CPLEX package on the synthetic data listed in Appendix A, with security prices manipulated in order to create an illustrative example.

Ignoring the additional allocation constraints and only including the availability (1) and requirement (2) constraints we see that the two requirements are matched exactly by posting collateral to the value of $R\,100$ to the one call and $R\,200$ to the other, with an overall allocation cost of $R\,300$. No cash is allocated to either call because of the higher preference for cash over other assets. No equities are allocated to the call where they are ineligible and as many government bonds as possible are retained in the inventory as they have a higher preference than both corporate bonds and equities, leaving an overall inventory worth $R\,140$ (Figure 2).

Incremental lot size constraints were tested by increasing the available units of equities from 30 to 31 and setting the lot size equal to 31. Here we see that for the agreement which allows equities as eligible collateral, 31 units are allocated to the call, resulting in a total allocation cost for the posting of $R\,103$ and a total allocation cost across calls equal to $R\,303$. The value of the collateral in the inventory remains at $R\,140$.

Similarly, the minimum lot size constraints (3) were evaluated by setting the minimum number of equity units to 31 and lot size to 1. The results show that the 31 units are allocated to the relevant call.

To test the cardinality constraints (4), the inequality was changed from $\leq$ to $\geq$ in order to force the allocation of 4 asset classes, as opposed to the optimal 2. We can see that in order to satisfy the constraint, assets across all four classes are now allocated to calls, including drawing from the cash pool which was previously avoided because of its higher weighting. This results in one requirement being met with $R\,100$ worth of collateral and the other with $R\,204$, both now including a single unit of cash with $R\,136$ remaining in the inventory.

The asset diversification constraints (5) were tested by specifying a maximum asset allocation proportion of 80%. Previously 30 units of equities at $R\,3$ each were allocated to a call, resulting in $R\,90$ of the $R\,100$ being met by the allocation (*i.e.* 90%). The impact of introducing this additional constraint is that cash and government bonds are allocated to satisfy the requirement thus reducing the percentage equity allocation.

Finally, the issuer diversification constraints (6) were tested by including available units of corporate bonds and equities from Telkom and reducing the available units of similar asset types from other issuers to zero. This results in 100% allocation of assets from the same issuer. However, once we introduce a limit on the percentage allocation from a single issuer we see that government bonds are included in the allocation so as to satisfy the constraint.

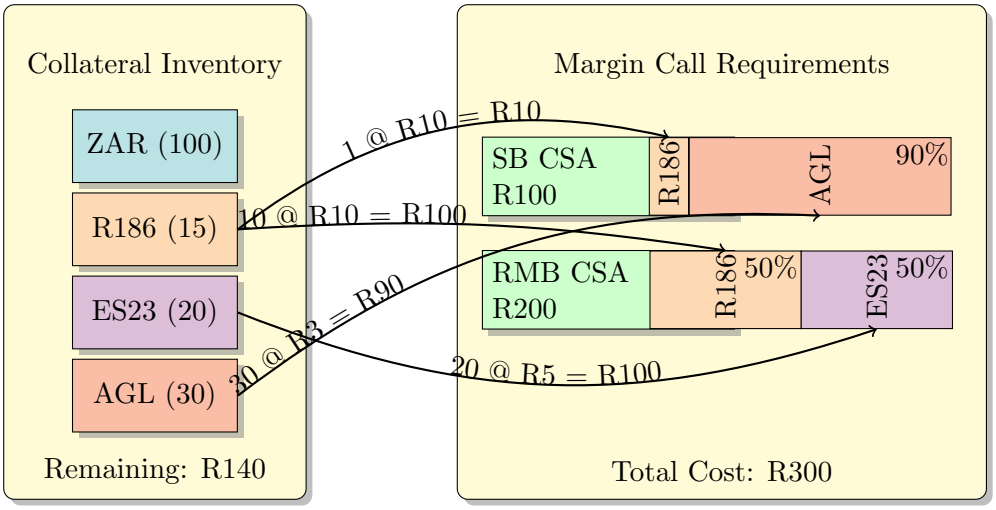


Figure 2: *Allocation costs.*

5 Conclusion

This paper proposed an integer linear programming formulation of the collateral optimisation problem. The model was tested on a simple data set. Future work will validate the implementation on a larger, more realistic set of data and investigate some of the computational complexity issues related to the problem. Possible alternate formulations could consider the ability to substitute collateral and a more detailed specification of funding costs in the objective function.

6 Appendix A: Simple Example

Id	Product	Counterparty	MTM
1	Interest Rate Swap	Standard Bank	-300
2	Forward Rate Agreement	Standard Bank	50
3	Forward Rate Agreement	Standard Bank	150
4	Interest Rate Swap	Rand Merchant Bank	-300
5	Interest Rate Swap	Rand Merchant Bank	100

Table 1: *Trade exposures.*

Agreement	Required Amount	Maximum Number Assets
SB CSA	R100	3
RMB CSA	R200	3

Table 2: *Margin calls.*

Id	Type	Counterparty	Eligible Collateral
SB CSA	CSA	Standard Bank	ZAR Cash South African Government Bonds Corporate Bonds Equities (JSE Top 40)
RMB CSA	CSA	Rand Merchant Bank	ZAR Cash South African Government Bonds Corporate Bonds

Table 3: *Agreements.*

Id	Type	Issuer	Issuer Type	Coupon	Maturity	ISIN[2]
ZAR	Cash	Republic of South Africa	Government			
R186	Bond	Republic of South Africa	Government	10.5	2026/12/21	ZAG000016320
ES23	Bond	Eskom Holdings	Corporate	10	2023/01/25	ZAG000074212
TL20	Bond	Telkom SA Limited	Corporate	6	2020/02/24	ZAG000021528
TKG	Equity	Telkom SA Limited	Corporate			ZAE000044897
AGL	Equity	Anglo American plc	Corporate			GB00B1XZS820

Table 4: *Securities.*

	ZAR	R186	ES23	TL20	TKG	AGL
CSA SB	1	1	1	1	1	1
CSA RMB	1	1	1	1	0	0

Table 5: *Eligibility matrix.*

[2]International Securities Identification Number (ISIN)

Id	Available Units	Price	Value	Weighting	Lot Size	Minimum Units
ZAR	100	1	$R\,100$	10	1	1
R186	15	10	$R\,150$	5	1	1
ES23	20	5	$R\,100$	3	1	1
TL20	0	5	$R\,0$	2	1	1
TKG	0	2	$R\,0$	1	1	1
AGL	30	3	$R\,90$	1	1	1

Table 6: *Collateral inventory.*

7 References

[1] Allen T & Hellaby E, 2013, *SunGard: Collateral Optimization - How It Really Works*, [Online], [Cited June 26th, 2014], Available from `http://finance.flemingeurope.com/webdata/4201/WP_CollateralOptimization_HowItReallyWorks_2013.pdf`

[2] Bank for International Settlements, 2013, *Margin requirements for non-centrally cleared derivatives*, [Online], [Cited September 2nd, 2013], Available from `http://www.bis.org/publ/bcbs261.pdf`

[3] Bonami P & Lejeune MA, 2009, *An Exact Solution Approach for Portfolio Optimization Problems Under Stochastic and Integer Constraints*, Operations Research, **57**, pp. 650–670

[4] Capel J & Levels A, 2014, *Collateral optimisation, re-use and transformation Developments in the Dutch financial sector*, De Nederlandsche Bank Occasional Studies, **12(5)**

[5] Cesarone F & Scozzari A & Tardella F, 2009, *Efficient algorithms for mean-variance portfolio optimization with hard real-world constraints*, Giornale dellIstituto Italiano degli Attuari, **12**, pp. 37–56

[6] Council of the European Union, 2012, *REGULATION (EU) No 648/2012 on OTC derivatives, central counterparties and trade repositories*, [Online], [Cited July 4th, 2012], Available from `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R0648`

[7] Gregory, 2012, *Counterparty Credit Risk and Credit Value Adjustment: A Continuing Challenge for Global Financial Markets*, 2nd Edition, John Wiley & Sons, United Kingdom

[8] Gürtler M & Heitchecker D & Hibbeln M, 2007, *Optimizing Credit Risk Mitigation Effects of Collaterals Under Basel II*, In *Operations Research Proceedings 2006*, Karlsruhe, Germany

[9] JPMorgan Chase & Co., 2012, *Regulatory Reform and Collateral Management: The Impact on Major Participants in the OTC Derivatives Markets*, [Online], [Cited September 13th, 2012], Available from `https://www.jpmorgan.com/tss/DocumentForEmail/Regulatory_Reform_and_Collateral_Manag-ement/1320476294035`

[10] Seagroatt M, 2012, *4sight Financial Software: Collateral Optimisation in a Centrally Cleared World*, [Online], [Cited October 20th, 2012], Available from `http://www.4sight.com/media/2310/4sight%20Whitepaper%20-%20Collateral%20Optimisation%20in%20a%20Centrally%20Cleared%20World.pdf`

Proceedings of the 2015 ORSSA Annual Conference
pp. 62–71
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

A modelling framework for shelf space allocation of fresh produce at a local retailer

J Lötter* JH van Vuuren†

Abstract

Retailers may gain a competitive advantage through the efficient management of shelf space, which is considered to be one of the most valuable resources in a supermarket. Several mathematical models have in the past been proposed to assist with the allocation of shelf space to the various products on offer. The most basic of these models appeared during the 1960s, but since then several variations, enhancements and improvements to these models have appeared. Products may, however, be classified into different categories according to a variety of characteristics. Considering a product's shelf life, for example, it can be categorised as perishable or non-perishable. Shelf space allocation decisions become more complicated in the area of fresh produce, due to the seasonality of these products, their short shelf lives and freshness-dependent demand. A modelling framework is proposed in this paper which may be used as the basis for practical decision support in respect of shelf space allocation for fresh produce at a local retail outlet.

Key words: Shelf space allocation, Retailing.

1 Introduction

South Africa has a competitive retailing industry. The expansion of this industry during the past two decades may be attributed to factors such as trade liberation after 1994 and extensive recent urbanisation [14]. Five top-performing retailers in South Africa were among the 250 largest retailers in the 2015 Global Powers of Retailing report [13].

Store managers in the retailing industry are faced with complex periodic decisions related to shelf space allocation and display periods for fresh produce. The complexity of these decisions may be attributed to the short shelf lives and seasonality of fresh produce, as well

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 16444949@sun.ac.za

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

as their freshness-dependent demand. A competitive advantage may be gained through the efficient management of shelf space in the fresh produce department of a retail outlet.

The research in this paper forms part of an ongoing project at Stellenbosch University in which the aim is to propose a fresh produce shelf allocation modelling framework and to incorporate this framework into a flexible, user-friendly decision support system which may be used as a practical support tool in respect of shelf space allocation decisions related to fresh produce within the South African retailing industry. The study is performed in collaboration with one of the top five South African retailers, which prefers to remain anonymous. A specific outlet of this retailer has been selected as the focus of a case study demonstrating workability of the decision support tool and underlying modelling framework.

The remainder of the paper is organised as follows. Literature on topics related to perishable products and shelf space allocation decisions in respect of these products is reviewed in Section 2. In Section 3, we propose a modelling framework for fresh produce shelf space allocation at the local retailing partner. Section 4 contains a brief conclusion, and some ideas with respect to possible future work is outlined in Section 5.

2 Literature review

Fresh produce exhibits certain characteristics that make it a unique type of product, but which also complicate decisions related to its inventory control and shelf space allocation. Due to its short shelf life, fresh produce is generally considered a perishable product, a class of products which is the topic of discussion in Section 2.1. Certain notions related to the mathematical modelling of perishable products are discussed in Section 2.2. Two shelf space allocation models which have inspired our proposed modelling framework are finally presented in Section 2.3.

2.1 Perishable products

The thousands of products in a supermarket may be classified into different categories based on a variety of characteristics. Contributing to the earliest shelf space allocation research, Brown and Tucker [5] reportedly partitioned products into three classes based on their responsiveness to changes in the amount of shelf space allocated to them. Products that are sold at a rate that is independent of how much shelf space is allocated to them (such as spices) are classified as *unresponsive products*, whereas products whose sales rates are slightly dependent on shelf space allocation (breakfast cereals, canned food, *etc.*) form part of the class of *general use products*. Finally, *occasional purchase products* are only sold when a large amount of shelf space is allocated to them so as to increase their visibility (sardines and nuts, for example). Classification with respect to product shelf life is often also observed when managing products in a store. Products are then categorised as either perishable or nonperishable.

Typical perishable products are meat, seafood, dairy products, eggs and fresh produce (fruit and vegetables). Perishable products exhibit certain characteristics which may influence shelf space allocation decisions. Compared to nonperishable products, the most

prevalent characteristic of a perishable product is its short shelf life [17]. In order to slow down the process of their deterioration, perishable products also require special storage conditions. Deterioration occurs quicker in higher temperatures [16], which is generally avoided by displaying products on refrigerated shelves. Frozen products are not categorised as perishable products, because freezing reduces the deterioration process rate significantly [17].

The quality and variety of perishable products displayed contribute to a customer's perception of the store quality and may be the primary reason why a customer prefers a certain store above another [11]. Fresh perishable products also portray an efficient supply chain and a high demand for the specific products. Often, new perishable products are procured to replenish shelves before all the products on display have been sold. This poses a decision as to whether new and old stock should be displayed together, which is a decision unique to perishable products [11]. Some retailers separate old and new stock, preferring to sell the older products at a discounted price. Products can be sold separately in the same store in order to prevent products of poor quality from affecting fresh ones. Alternatively, older products can be sold from a different retail outlet in a lower income area [10]. Stores which choose not to separate old and new stock do so to minimise administration related to product prices and to minimise the amount of shelf space allocated to a specific product [11].

Other product characteristics that have been identified as contributing toward distinguishing between perishable and nonperishable products include average weekly sales, the coefficient of variation in weekly sales, delivery frequency, case pack size and minimum inventory [17]. Higher weekly sales and less variation in average weekly sales are typically associated with perishable products. The minimum inventory level of perishable products is higher, deliveries of perishable products are made more often and perishable products' case packs typically contain fewer units.

The supply chain of perishable products is referred to as a cold chain, and consists of several processes followed to maintain special conditions from the time of harvest until products reach the end customer [9]. A well-managed, temperature-controlled supply chain is normally associated with a competitive advantage [2]. Although several advanced technologies exist for improving temperature control, factors such as delays in deliveries are difficult to control and negatively influence product quality [1]. An inefficient cold chain leads to waste [9], which is an unnecessary expense that should be avoided as far as possible.

2.2 Mathematical modelling of perishable products

For mathematical modelling purposes, perishable products are generally classified into three categories, namely products with fixed shelf lives, products whose quality decays proportionally over time, or products with random shelf lives [18]. Products with fixed shelf lives have predetermined expiry dates and an entire batch of products (of the same age) perishes at the same time [12]. Medicines and most food items form part of this category [7]. The second category refers to perishable products which decay at a rate that is directly proportional to the amount of the product present (also known as exponential decay). Examples of this category are radioactive materials and chemicals [8]. Perishable

products have random shelf lives when the time to spoilage is not known beforehand and differs from product to product. Fresh produce generally forms part of this last category [7].

2.3 The shelf space allocation problem

Several mathematical models have been proposed since the 1960s to assist with the allocation of shelf space at retail outlets. One of the most cited models, proposed by Corstjens and Doyle [6] in 1981, takes into account a product's contribution towards profit, its responsiveness to changes in the amount of shelf space allocated to it (space elasticity), cross-elasticities among products and costs associated with displaying the product. Constraints in this model include the total amount of shelf space available, and upper and lower display size limits per product. The aim of the model is to maximise overall profit as a result of product exposure to consumers. A literature review on the shelf space allocation problem revealed twelve models that are based specifically on the model developed by Corstjens and Doyle. Other models that follow different approaches are, however, also abundant in the literature.

Bai and Kendall [4] proposed a particularly useful model for the shelf space allocation of fresh produce by combining a deteriorating inventory model with a shelf space allocation model. Their assumptions differ from those underlying previous attempts at modelling shelf space allocation in the literature. In previous attempts, fresh produce had traditionally been considered a special class of perishable products, which implies a fixed deterioration rate and that all products have the same value unless they have expired. Bai and Kendall claimed that it is possible to predict the expiry dates of fresh produce, by using advanced cooling technology in the fresh produce supply chain. Their model is built on the assumption that the freshness condition of fresh produce decreases continuously although the value is not entirely lost by the time the products expire. They also assume that the demand for fresh produce depends on the volume and the freshness of the products displayed. Freshness is a consumer measurement of quality. In other models it has typically been assumed that demand depends on inventory as a whole, but due to the scarcity of shelf space, only a fraction of inventory can usually be displayed. In the model of Bai and Kendall, an ordering policy is determined for fresh produce as well as the amount of shelf space to be allocated to each product. They used the generalised reduced gradient method to solve instances of the model. Bai, Burke and Kendall [3] also used a metaheuristic and a hyperheuristic to solve instances of the same model approximately.

Although numerous models have been proposed for both shelf space allocation and inventory control, none of them was developed specifically for fresh produce prior to the model of Bai and Kendall [4]. Only one other similar model appears in the literature. Piramuthu and Zhou [15] proposed an extension to the model of Bai and Kendall in 2013. They drew into question Bai and Kendall's assumption that all the displayed items of the same product are of the same quality. According to Piramuthu and Zhou, items are exposed to varying environmental conditions and therefore advanced technologies should be employed to track individual item deterioration. In contrast to the approach of Bai and Kendall, they modelled demand as a function of both item freshness and allocated shelf space. Furthermore, they computed the effective amount of shelf space allocated to

items of a specific product, which is less than the actual allocated shelf space because of the influence of deteriorated items on the product demand. The rest of their model is similar to that of Bai and Kendall.

3 Model development

The local retail partner wishes to gain a competitive advantage by improving operations in its fresh produce department as a result of streamlining decisions related to order quantities, display periods and shelf space allocation. The general nature of operation at an outlet of the local retailer in question is described in Section 3.1. The development of our modelling framework tailored for the specific retailer is presented in Section 3.2, and the model is finally proposed in Section 3.3.

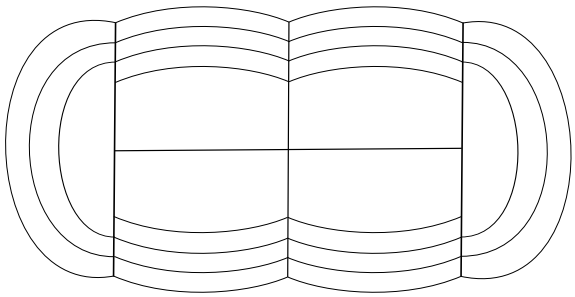


Figure 1: *Top-view layout of a fruit and vegetable boat at the local retailer.*

3.1 The local retailer

As in most retail outlets, the layout of the fresh produce department at the local retail partner varies significantly from those employed in other aisles and shelves. Fresh produce is displayed in a single loose-standing produce display, also referred to as a fruit and vegetable boat, as well as in refrigerated shelves. Figure 1 contains a simplified top-view representation of the fruit and vegetable boat of the outlet considered. The only calculations incorporated in the current shelf space allocation method is the products' rates of sale. Products which sell the fastest are given the most shelf space, and the most popular products are displayed on the semi-circular sides. The products are displayed in compartments, which may be shared by more than one product or may be reserved for one type of product.

With respect to lead time, products are either on a 2-day or a 3-day delivery roster. Some products are delivered from the distribution centre two days after having been ordered and some three days. Product orders are, however, placed on a daily basis. Once the delivery process reaches a steady state, products are delivered whenever necessary. Products are ordered in multiples of a so-called pack size, which is the number of units that are packaged together at the distribution centre.

The local retail outlet identified for case study purposes operates with minimal backroom

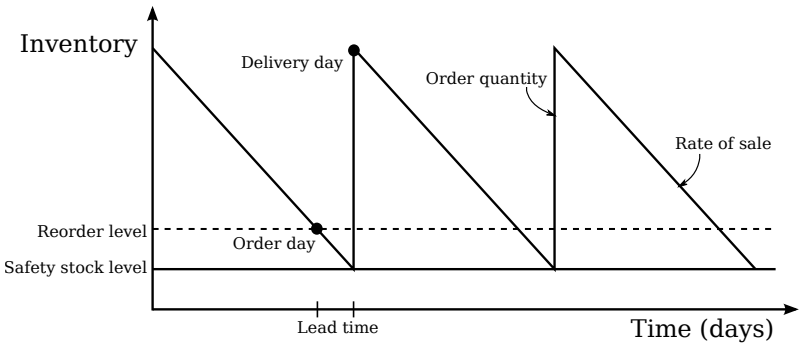


Figure 2: *The inventory control policy adopted at the local retailer.*

inventory. Only 20 percent of the outlet floor space is dedicated to inventory storage, with no space to store additional fresh produce. The entire fresh produce inventory is therefore displayed on the shelves. As in most inventory control systems, the local retail partner incorporates the notion of a safety stock in inventory replenishment decisions. The method is illustrated in Figure 2. In order to prevent stock-outs and increase customer satisfaction, the retailer focuses on placing orders in time so that the safety stock is never compromised.

In Figure 2, the relationship between order quantity, lead time and rate of sale can be seen. Rate of sale describes how the products leave the store, while the lead time and order quantity together describe the replenishment of the products. Shelf space allocation decisions are influenced by this relationship as well as the expected shelf life of a specific product.

The local retailer seeks to gain a competitive advantage by improving stock turnover, eliminating (or at least minimising) product waste and not having too much capital tied up in stock. In respect of inventory control, the main focus is on maintaining the safety stock level, the second most important aspect is the replenishment of products and almost as important as this are decisions related to the actual display of products in the store.

A shelf space allocation model will typically be solved twice per year in order to determine the allocation of the available shelf space for an entire season (summer or winter). The model of Bai and Kendall [4], described in Section 2.3, may be applied to form the basis of decision support for the retail partner. The model is therefore used as a foundation from which we develop a modelling framework tailored for specific use by the retail partner. Some model adaptations are, however, required in order to better represent requirements at the local retailer's outlets. These adaptations are described in the following section.

3.2 Suggested model adaptations

In the model of Bai and Kendall [4], two factors are accounted for that are not present at the local retail outlet, namely backroom inventory and a second, discounted selling period. The associated variables are therefore excluded in the adapted model. Some of the modelling assumptions adopted by Bai and Kendall are also not valid for the local problem. For example, lead time is considered to be zero in the model of Bai and Kendall.

Some expressions in the model of Bai and Kendall will, on the other hand, have to be simplified due to data unavailability constraints and to decrease the complexity of their modelling framework. This section contains a systematic discussion of how we adapted the model of Bai and Kendall to suit the situation at the local retailer.

Instead of calculating the demand for fresh produce from scale parameters, elasticity values, decaying rates and inventory levels, the rates of sale will be used instead. These rates are a good representation of demands for the different fresh produce products. This is the preferred approach because trends exist in the rate of sales data and because all the parameters present in the Bai and Kendall demand formula are not available.

The model of Bai and Kendall does not account for safety stock, but it does allow for the possibility of surplus inventory at the end of a decision cycle. Such surplus inventory should be sold at a discounted price. The local retailer currently does not have this option. Therefore, the incorporation of a surplus inventory value into the model should be replaced by a constant representing the safety stock, which can be determined from the lead time and daily demand of a product.

The other two types of decision variables calculated in the model of Bai and Kendall, procurement quantity and number of facings, are relevant to the local retailer. The number of facings will, however, be taken as the number of product units allocated to a compartment of the fruit and vegetable boat — not only the number entirely visible to the customer, as the name might suggest. The number of facings will, however, be equal to the sum of the procurement quantity and the safety stock, because there is no backroom inventory. The procurement quantity will be the quantity ordered from the retailer's distribution centre. The quantity must be in multiples of the pack size, as mentioned. To account for this requirement, an additional constraint should be added to the model.

Due to the shelf layout of the fresh produce department at the local retailer's outlets, the total shelf space available has to be measured in number of compartments. The amount of shelf space required for a facing of a product unit will then be taken as the portion of a compartment taken up by one item (one divided by the number of items that fit into a compartment).

We have confirmed that the selling prices (neglecting the use of discounted selling prices), unit costs and fixed costs are available from the local retailer. Data on product shelf lives, and lower and upper bounds on the number of product facings are also available. A simplified formula will be used for determining the holding cost during a decision cycle at the local retailer. Only one type of period occurs during each decision cycle at the local retailer, as opposed to the normal period and discounted period in the model of Bai and Kendall [4]. The holding cost per decision period may be determined from the average inventory cost and the acquisition cost per product.

After affecting all these model adaptations, the total profit can still be calculated as the sum of all the individual product profits. The objective of the model for the local retailer will remain the same as that of Bai and Kendall. The local retailer does, however, not merely want to make as much profit as possible — it also values customer satisfaction. Because customer satisfaction is not easily measured, the way in which it will be accounted for in the model will vary from outlet to outlet, based on customer segmentation. For

example, a retail outlet in a neighbourhood with many school children should be able to provide for weekly purchases of lunchbox snacks put together by mothers on Sunday evenings. Customer segments of other retail outlets may, on the other hand, exhibit a need for specific products to be sold together.

3.3 Model formulation

As mentioned, the objective in our model is to maximise the overall profit as contributed by the individual product profit functions. Let $\mathcal{P} = \{1, \ldots, |\mathcal{P}|\}$ be the set of fresh produce products to be displayed at the outlet, and define the decision variable q_i as the order quantity of product $i \in \mathcal{P}$ for a decision cycle.

Let d_i denote the average daily demand for product $i \in \mathcal{P}$. From Figure 2, the expression for the length of the decision cycle is deduced as a function of the order quantity and is represented by $T_i = q_i/d_i$ for all $i \in \mathcal{P}$. Let L_i denote the lead time of product $i \in \mathcal{P}$. Then $s_i = L_i d_i/2$ is the safety stock level of product $i \in \mathcal{P}$. Furthermore, let a_i denote the acquisition cost of product $i \in \mathcal{P}$ and let $h_i = a_i(\frac{q_i}{2} + s_i)$ be the holding cost of product $i \in \mathcal{P}$ during a decision cycle. The number of items of product $i \in \mathcal{P}$ displayed on the shelves may therefore be denoted by $f_i = q_i + s_i$.

In addition, let the characteristics of product $i \in \mathcal{P}$ be denoted by p_i (the selling price), o_i (the fixed ordering cost), e_i (the amount of shelf space taken up by one unit of the product), ℓ_i and u_i (the lower and upper limits on the number of product facings), t_i (the expected shelf life), and k_i (the pack size). Furthermore, let S denote the total amount of shelf space available and let c_s be the cost of shelf space per unit of space. Then the average profit of product $i \in \mathcal{P}$ per unit time may be expressed as $P_i = \frac{1}{T_i}[(p_i - a_i)q_i - o_i - h_i] - c_s f_i e_i$ and the objective of the model is to

$$\text{maximise} \quad \sum_{i \in \mathcal{P}} P_i(q_i)$$

subject to the constraints

$$\begin{aligned}
\sum_{i \in \mathcal{P}} f_i e_i &\leq S, && && (1)\\
\ell_i &\leq f_i \leq u_i, && i \in \mathcal{P}, && (2)\\
q_i &> s_i, && i \in \mathcal{P}, && (3)\\
0 &< T_i \leq t_i, && i \in \mathcal{P}, \text{ and} && (4)\\
\frac{q_i}{k_i} &\in \{1, 2, 3, \ldots\}, && i \in \mathcal{P}. && (5)
\end{aligned}$$

Constraint (1) ensures that the allocated shelf space does not exceed the available amount of space, while constraint set (2) limits the number of facings allocated to product $i \in \mathcal{P}$ between its minimum and maximum values. Constraint set (3) furthermore ensures that the order quantity is more than the safety stock level, while constraint set (4) restricts the length of the decision period to be shorter than the expected product shelf life. Constraint set (5) finally ensures that the order quantity is a multiple of the pack size.

4 Conclusion

Upon reviewing the literature related primarily to perishable products, a modelling framework for the shelf space allocation of fresh produce at a local retail partner was proposed in this paper. This framework was based on a model by Bai and Kendall [4], but included several adaptations to better represent the situation experienced at the local retail partner. The objective of the model is to maximise overall profit subject to a number of constraints developed according to the situation at the local retail partner.

5 Further work

Developing the modelling framework here in order to portray the operation of the retail partner more realistically is an ongoing process. The model proposed is currently still in a basic form, and more detail will be added as our understanding of the requirements of and processes followed by the local retail partner deepens over time. After a sufficient level of model detail has been achieved, the modelling framework will be verified and validated in the context of the case study outlet mentioned in Section 1, using real data.

It is envisaged that the modelling framework may eventually be incorporated into a computerised decision support system. Although the modelling framework and decision support system will be developed specifically for the local retailer, it is anticipated that many of the requirements and specifications identified at the case study retail outlet are commonly experienced at local supermarkets. The proposed framework and system are therefore expected to be generic to some extent.

References

[1] Aiello G, La Scalia G & Micale R, 2012, *Simulation analysis of cold chain performance based on time-temperature data*, Production Planning & Control: The Management of Operations, **23(6)**, pp. 468–476.

[2] Aung MM & Chang YS, 2014, *Temperature management for the quality assurance of a perishable food supply chain*, Food Control, **40**, pp. 198–207.

[3] Bai R, Burke EK & Kendall G, 2008, *Heuristic, meta-heuristic and hyper-heuristic approaches for fresh produce inventory control and shelf space allocation*, Journal of the Operational Research Society, **59**, pp. 1387–1397.

[4] Bai R & Kendall G, 2008, *A model for fresh produce shelf-space allocation and inventory management with freshness-condition-dependent demand*, INFORMS Journal on Computing, **20(1)**, pp. 78–85.

[5] Brown WM & Tucker WT, 1961, *Vanishing shelf space*, Atlanta Economic Review, **9**, pp. 9–13.

[6] Corstjens M & Doyle P, 1981, *A model for optimizing retail space allocations*, Management Science, **27(7)**, pp. 822–833.

[7] Goyal S & Giri B, 2001, *Recent trends in modeling of deteriorating inventory*, European Journal of Operational Research, **134(1)**, pp. 1–16.

[8] Gürler Ü & Özkaya BY, 2008, *Analysis of the (s,S) policy for perishables with a random shelf life*, IIE Transactions, **40**, pp. 759–781.

[9] Joshi R, Banwet D, Shankar R & Gandhi J, 2012, *Performance improvement of cold chain in an emerging economy*, Production Planning & Control: The Management of Operations, **23**, pp. 817–836.

[10] Kar S, Bhunia AK & Maiti M, 2001, *Inventory of multi-deteriorating items sold from two shops under single management with constraints on space and investment*, Computers and Operations Research, **28**, pp. 1203–1221.

[11] Li Y, Cheang B & Lim A, 2012, *Grocery perishables management*, Production and Operations Management, **21(3)**, pp. 504–517.

[12] Liu L & Shi DH, 1999, *(s, S) model for inventory with exponential lifetimes and renewal demands*, Naval Research Logistics, **46**, pp. 39–56.

[13] National Retail Federation, 2015, *2015 Top 250 global powers of retailing*, [Online], [Cited February 26th, 2015], Available from `https://nrf.com/news/2015-top-250-global-powers-of-retailing`.

[14] Peyton S, Moseley W & Battersby J, 2015, *Implications of supermarket expansion on urban food security in Cape Town, South Africa*, African Geographical Review, **34(1)**, pp. 36–54.

[15] Piramuthu S & Zhou W, 2013, *RFID and perishable inventory management with shelf-space and freshness dependent demand*, International Journal of Production Economics, **144(2)**, pp. 635–640.

[16] Qin Y, Wang J & Wei C, 2014, *Joint pricing and inventory control for fresh produce and foods with quality and physical quantity deteriorating simultaneously*, International Journal of Production Economics, **152**, pp. 42–48.

[17] Van Donselaar K, Van Woensel T, Broekmeulen R & Fransoo J, 2006, *Inventory control of perishables in supermarkets*, International Journal of Production Economics, **104(2)**, pp. 462–472.

[18] Wang X & Li D, 2012, *A dynamic product quality evaluation based pricing model for perishable food supply chains*, Omega, **40(6)**, pp. 906–917.

Proceedings of the 2015 ORSSA Annual Conference
pp. 72–80
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

A new vehicle routing problem with application to pathology laboratory service delivery

A Smith* A Colmant* L Oosthuizen† JH van Vuuren‡

Abstract

Accurate and reliable clinical laboratory testing is an important component of a public health approach to disease management in resource-limited settings, with a safe and reliable transportation network playing an integral role in the delivery of the required services. The collection of specimens from a multitude of specimen collection stations and the subsequent transportation of these specimens to respective laboratories for processing, poses a serious logistical challenge to any pathological testing organisation. The specimen collection process is modelled in this paper as a large tetra-objective *vehicle routing problem* (VRP) which may be used as the basis of a decision support system capable of aiding pathological laboratory services in respect of cost-effective planning, routing and scheduling of a large dedicated fleet of vehicles responsible for the delivery of specimens to laboratories. The model builds on a combination of various well-known variants of the celebrated VRP in the literature, but also exhibits novel features, such as an incompatibility between certain types of specimens and laboratories in terms of testing facilities and capabilities available at the laboratories, the equalisation of specimen testing workload across laboratories, limitations in the rates at which the various laboratories can analyse specimens, and time window constraints within which specimens have to be delivered to laboratories.

Key words: Vehicle routing problem, Pathology laboratory service.

1 Introduction

A consultation held in January 2008 in Maputo, Mozambique served to draw up, in collaboration with the World Health Organisation, the Centre for Disease Control and Prevention, the United States Agency for International Development, the American Society for Clinical Pathology, the Clinton Foundation, the Bill and Melinda Gates Foundation and the Supply Chain Management System [5], documentation containing recommendations

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 16678591@sun.ac.za

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

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

for health care services. The documentation provides a framework for a tiered, integrated network of pathology laboratories with the aim of strengthening laboratory capacity in resource-limited settings. Amongst several other African states, South Africa took part in the development of this framework (and is also a signatory of the Maputo declaration). The South African *National Health Laboratory Service* (NHLS), in fact, consists of such a tiered laboratory network.

These laboratories are partitioned into four tiers: primary laboratories (tier 1), district laboratories (tier 2), regional laboratories (tier 3) and national laboratories (tier 4). The various laboratories have increasing levels of resources and capabilities available with respect to the processing of test specimens as their tier level increases. The tiered level of a laboratory system and the types of testing performed at each level may vary depending on the population served, physical infrastructure available, the level of service available, water and electricity available, road conditions and the availability of trained technical personnel in-country [5].

The purpose of this paper is to put forward a novel variation on the well-known *vehicle routing problem* (VRP) which may be used as a basis for providing decision support for pathology laboratories in respect of the efficient and effective routing and scheduling of its dedicated fleet of specimen collection vehicles.

An acceptable trade-off between four objectives are pursued in this model formulation, namely to minimise the cost associated with the specimen collection routing schedule, to balance the analysis workload at each laboratory, to minimise the longest time spent by a vehicle on the road and finally to minimise the number of vehicles required to implement the specimen collection routing schedule.

This paper is organised as follows. A brief review is conducted in §2 of the large body of literature on the VRP and its variations, after which the objectives and constraints of our VRP formulation are introduced and motivated in §3. The paper closes in §4 with a number of ideas for possible future work related to the proposed formulation.

2 Literature study

The VRP was first introduced into the operations research literature in a paper by Dantzig and Ramser [7] in 1959 who were concerned with the real-world application of delivering gasoline to gas stations. The first mathematical formulation of the VRP was proposed in the paper and an algorithmic solution approach was suggested for the VRP, formulated simply as the celebrated *Travelling Salesman Problem* (TSP) with the addition of a capacity constraint. The VRP should, in fact, rather be viewed as a combination of the TSP and the well-known *bin packing problem*. The algorithm originally proposed by Dantzig and Ramser was limited to small instances of the problem, but in 1964 Clarke and Wright [4] proposed an efficient greedy heuristic for obtaining good solutions to larger instances of the VRP. While the VRP is a generalisation of the TSP, it is much more difficult to solve than the TSP. Exact algorithms exist for the TSP which routinely solve instances with hundreds or thousands of vertices [1] while the best exact algorithms for the VRP can currently only solve instances with roughly a hundred vertices [3, 9].

Significant research interest has been generated by the VRP over the past fifty years. Toth and Vigo [13] have suggested a classification system for variations on the VRP in terms of:

- the underlying transportation network structure,
- the type of transportation requests,
- the constraints that affect each route individually (intra-route constraints),
- the vehicle fleet composition and their home locations,
- various inter-route constraints, and
- the optimisation objectives.

There are numerous approaches toward solving variations on the VRP, depending on the size of the instance and user requirements. The two main approaches to solving these problems involve the use of exact algorithms or metaheuristics. The favoured exact approach has been column generation, introduced by Desrosiers [8] and usually applied to *VRPs with Time Windows*. In instances of the standard *Capacitated VRPs* it often under-performs, however, and so a *branch-and-cut* approach is usually favoured instead, but this superior approach is still limited to instances with fifty customers or less. In 2006, Fukasawa *et al.* [9] introduced a branch-and-cut-and-price algorithm which proved to be more effective in handling larger instances. Continual improvement in exact algorithmic approaches has been made, with instances of more than 150 customers served by 12 vehicles being solved exactly by Contardo [6].

Heuristic solution approaches are almost as old the problem itself, with Dantzig and Ramser [7] introducing a basic heuristic based on the successive matching of customers by the solution of linear programming relaxations and the removal of fractional solutions by trial and error. Since then, numerous *constructive* and *improvement* heuristics have been developed. More recently, effective *metaheuristics* have also been designed which are powerful enough to solve large instances approximately within seconds, with solutions often achieving to within one percent of the optimal objective function value [13]. The most common metaheurisitcs applied to variations on the VRP are tabu search, ant colony optimisation, particle swarm optimisation and simulated annealing [13].

3 Mathematical modelling

In this study we demonstrate how the specimen collection problem of a tiered pathological laboratory service can be translated into the mathematical formulation of a new type of VRP.

3.1 Set notation and parametric configuration

Let $\mathcal{V}^r = \{1, \ldots, |\mathcal{V}^r|\}$ denote the set of homogeneous vehicles that make up the pathological testing service's specimen collection fleet. It is assumed that this set is large enough

to facilitate a feasible specimen collection routing and scheduling solution at a 100% service level, as is required by most health care organisations. The homogeneity of the fleet implies that all the vehicles have the same finite freight capacity C_{max} and autonomy level μ (the maximum assignable route length, measured in units of expected time duration). Let b_k represent the home depot within the set of all depots $\mathcal{V}^b = \{1, \ldots, |\mathcal{V}^b|\}$ of vehicle k, from which it sets out on its collection route and to which it must return upon completion of its delivery tour. Let $G = (\mathcal{V}, \mathcal{E})$ represent an undirected graph with vertex set $\mathcal{V} = \mathcal{V}^e \cup \mathcal{V}^d$, representing the set of all specimen collection stations from which pathological specimens have to be collected for analysis $\mathcal{V}^e$ together with the set $\mathcal{V}^d$ of all laboratories to which specimens may be delivered, and edge set $\mathcal{E}$ representing all road connections between destinations in $\mathcal{V}^e$ and $\mathcal{V}^d$. Every point in $\mathcal{V}$ is assumed to be reachable from every other point. Suppose the edge $(i, j) \in \mathcal{E}$ is expected to be traversed in t_{ij} time units at an associated cost c_{ij}. Every specimen collection point $i \in \mathcal{V}$ (laboratory, respectively) has a time window $[a_i, b_i]$ associated with it during which specimens can be collected (delivered, respectively) as well as a service time S_i associated with handling a batch of pathological specimens there. Denote the set of all pathological specimen types by $\mathcal{V}^c = \{1, \ldots, |\mathcal{V}^c|\}$. Each specimen of type $o \in \mathcal{V}^c$ also has an expiration time τ_o before which it must be processed at a laboratory. Every specimen collection station may potentially require different types of specimen collection based on demographic variability and fluctuating demand. Therefore, let $\mathcal{Q}^i = \{q_1^i, \ldots, q_{|\mathcal{V}_c|}^i\}$ represent the specimen collection requirements at specimen collection station i, where q_o^i denotes the volume of specimens of type $o \in \mathcal{V}^c$ requiring collection. Furthermore, define

$$\alpha_{io} = \left\lceil \frac{q_o^i}{C_{max}} \right\rceil,$$

where C_{max} denotes the cargo capacity of a delivery vehicle. Then α_{io} denotes the number of vehicles that have to be scheduled to visit specimen collection station $i \in \mathcal{V}^e$ for the collection of specimens of type $o \in \mathcal{V}^c$. Similar to each specimen collection point, every laboratory $d \in \mathcal{V}^d$ has varying processing capabilities. We utilise the parameter

$$\delta_{do} = \begin{cases} 1, & \text{if specimen type } o \in \mathcal{V}^c \text{ may be transported to laboratory } d \in \mathcal{V}^d \\ 0, & \text{otherwise} \end{cases}$$

to control the delivery of specimens to capable laboratories which can actually analyse these specimens. The different tiered laboratories have different processing rates associated with them, and the decision variable β_{do} denotes the rate at which laboratory $d \in \mathcal{V}^d$ is able to process specimens of type $o \in \mathcal{V}^c$. A maximum processing capacity is also associated with each specimen type at each laboratory. The parameter γ_{do} represents this maximum processing ability at laboratory $d \in \mathcal{V}^d$ in respect of specimens of type $o \in \mathcal{V}^c$.

3.2 Model formulation

In the model formulation, decision variables are required to keep track of the movement and allocation of vehicles to specimen collection stations and laboratories. The decision variable

$$x_{ijk} = \begin{cases} 1, & \text{if vehicle } k \text{ is scheduled to traverse arc } (i, j) \text{ in } \mathcal{E} \\ 0, & \text{otherwise} \end{cases}$$

monitors the movement of vehicle k, while the decision variable

$$y_{iko} = \begin{cases} 1, & \text{if vehicle } k \text{ is scheduled to collect type } o \text{ specimens from customer } i \\ 0, & \text{otherwise} \end{cases}$$

is required to monitor which vehicle services each specimen collection station and what types of specimens each vehicle should transport, since the specimen collection stations exhibit varied service demand. Finally, the decision variable

$$z_{diko} = \begin{cases} 1, & \text{if vehicle } k \text{ is scheduled to transport specimens of type } o \text{ from} \\ & \text{specimen collection station } i \text{ to laboratory } d \\ 0, & \text{otherwise} \end{cases}$$

is required.

Following the discussion in §1, the aim of our model is to pursue an acceptable trade-off between optimising four objective functions. The first of these objectives is to minimise the costs associated with transportation of all the specimens from the specimen collection stations at which they originate to the appropriate laboratories where they are to be analysed, as is standard in most VRPs. This objective may be formulated as

$$\text{minimise} \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{V}} \sum_{k\in\mathcal{V}^r} c_{ij} x_{ijk}.$$

Our second objective is to balance as much as possible the workload of laboratories in terms of the time required to analyse specimens, which may be formalised as

$$\text{minimise} \max_{d} \sum_{i\in\mathcal{V}^e} \sum_{k\in\mathcal{V}^r} \sum_{o\in\mathcal{V}^c} \frac{q_i^o}{\beta_{do}} z_{diko}.$$

The third objective is to balance the workload of the delivery vehicles in terms of their total travel time, which may be expressed as

$$\text{minimise} \max_{k} \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{V}} x_{ijk} t_{ij}.$$

Our final objective is to

$$\text{minimise} \sum_{k\in\mathcal{V}^r} \sum_{j\in\mathcal{V}} x_{b_k jk},$$

i.e. to minimise the number of vehicles required for specimen collection at a service level of 100% by reducing the number of trips departing from home depots.

The model includes numerous constraints reflecting the requirements of the tiered pathological testing service in respect of the transportation of pathological specimens. The first such constraint set is

$$M \sum_{i\in\mathcal{V}} x_{ijk} \;\geq\; \sum_{o\in\mathcal{V}^c} y_{jko}, \quad j \in \mathcal{V}^e,\;\; k \in \mathcal{V}^r,$$

which ensures, if vehicle k is scheduled to collect at least one specimen from specimen collection station j, that vehicle k must traverse an arc of the form (i, j) for some $i \in \mathcal{V}^e$ at some point along its route. Here M is a large number (any number larger than $|\mathcal{V}|$ will do). The flow conservation constraint set

$$\sum_{i \in \mathcal{V}} x_{ijk} - \sum_{\ell \in \mathcal{V}} x_{j\ell k} = 0, \quad j \in \mathcal{V}, \ k \in \mathcal{V}^r$$

states that if vehicle k arrives at location j, then the same vehicle must traverse an arc departing from vertex j, for all $j \in \mathcal{V}$ and all $k \in \mathcal{V}^r$.

Another flow constraint set is required to control the delivery of specimens. In this respect, the constraint set

$$y_{iko} - \sum_{d \in \mathcal{V}^d} z_{diko} \delta_{do} = 0, \quad i \in \mathcal{V}^e, \ k \in \mathcal{V}^r, \ o \in \mathcal{V}^c$$

states that if vehicle $k \in \mathcal{V}^r$ is scheduled to collect specimens of type $o \in \mathcal{V}^c$ from specimen collection station $i \in \mathcal{V}^e$, then the same vehicle must deliver these specimens at exactly one feasible laboratory $d \in \mathcal{V}^d$.

It is safe to assume that pathological testing services utilise numerous depots where vehicles may be stored overnight for security reasons. As mentioned in §3.1, all vehicles are assumed to have home depots associated with them. The constraint set

$$\sum_{i \in \mathcal{V}^e} x_{b_k ik} - \sum_{j \in \mathcal{V}^d} x_{jb_k k} = 0, \quad k \in \mathcal{V}^r$$

ensures that vehicle k begins and ends its route at its home depot $b_k \in \mathcal{V}^b$, for all $k \in \mathcal{V}^r$.

As previously stated, each specimen collection station may require collection of different types of specimens by the vehicles and so the constraint set

$$\sum_{k \in \mathcal{V}^r} y_{iko} = \alpha_{io}, \quad i \in \mathcal{V}^e, \ o \in \mathcal{V}^c$$

is incorporated to ensure that vehicles service specimen collection stations appropriately. Let C_{ik} denote the volume of freight in vehicle $k \in \mathcal{V}^r$ when it leaves facility $i \in \mathcal{V}$. The constraint set

$$C_{ik} + \sum_{o \in \mathcal{V}^c} q_o^j y_{jko} - \sum_{o \in \mathcal{V}^c} q_o^j z_{jiko} - C_{max} \leq (1 - x_{ijk}) M', \quad i \in \mathcal{V}, \ j \in \mathcal{V}, \ k \in \mathcal{V}^r,$$

is included to ensure that no vehicle capacity constraint is exceeded. More specifically, this constraint set ensures, if vehicle $k \in \mathcal{V}^r$ travels from facility $i \in \mathcal{V}$ directly to facility $j \in \mathcal{V}$, that its freight contents after leaving facility j (*i.e.* its contents after leaving facility i plus the freight collected at facility j less the content delivered at facility j if it is a laboratory) does not exceed the vehicle capacity C_{max}. Here M' is again a large number.

Let T_{ik} be the expected arrival time of vehicle k at destination $i \in \mathcal{V}^e \cup \mathcal{V}^d$. The constraint set

$$T_{ik} + S_i + t_{ij} - T_{jk} \leq (1 - x_{ijk}) M'', \quad i \in \mathcal{V}, \ j \in \mathcal{V}, \ k \in \mathcal{V}^r$$

is included to monitor the arrival time of vehicle k at customer j. This constraint set ensures, if vehicle $k \in \mathcal{V}^r$ travels from location $i \in \mathcal{V}$ to location $j \in \mathcal{V}$, that the time instant at which it starts to service the facility at j is bounded from below by the time instant at which it started servicing the facility at i together with the combined service time duration at i and the travel time from i to j. Here M'' is yet again a large number.

Apart from the multiple problem objectives, the novelty of the problem is illustrated in the next constraint set. Each specimen has a certain time window associated with it during which it remains viable. The time window depends on the specimen type, available storage techniques within the vehicles and specimen collection stations, and the potential use of the specimen in question. The constraint set

$$\tau_o - z_{diko} T_{dk} \geq 0, \quad d \in \mathcal{V}^d, \;\; i \in \mathcal{V}^e, \;\; k \in \mathcal{V}^r, \;\; o \in \mathcal{V}^c$$

is incorporated to ensure, if vehicle k is scheduled to transport specimens of type $o \in \mathcal{V}^c$ from specimen collection station $i \in \mathcal{V}^e$ to laboratory $d \in \mathcal{V}^d$ at a later stage along its route, that the difference between collection time and delivery time of these specimens does not exceed its expiration time τ_o.

It is required that the services provided by pathological testing organisations and the respective laboratories are not twenty-four hour operations, but should take place within acceptable time windows associated with each collection point and laboratory. The constraint set

$$a_i \leq T_{ik} \leq b_i, \quad i \in \mathcal{V}^e \cup \mathcal{V}^d, \;\; k \in \mathcal{V}^r$$

states that vehicle k may not arrive at a specimen collection station or laboratory outside of its associated time window.

If the planning schedule is determined on a daily basis, it is highly unlikely that vehicles will require refueling before returning to their home depots or that the route times will exceed the maximum assignable times stated in standard labour regulations. The constraint set

$$T_{b_k k} \leq \mu, \quad k \in \mathcal{V}^r$$

may nevertheless be incorporated to ensure that vehicle k does not undertake a route which is expected to take longer to complete than its time autonomy level in cases where the overall scheduling is not one day.

Finally, laboratory $d \in \mathcal{V}^d$ is limited in that there is a maximum capacity γ_{do} associated with the processing of specimens of type $o \in \mathcal{V}^c$. The constraint set

$$\sum_{i \in \mathcal{V}^e} \sum_{k \in \mathcal{V}^r} z_{diko} q_o^i \leq \gamma_{do}, \quad d \in \mathcal{V}^d, \;\; o \in \mathcal{V}^c$$

ensures that the processing capabilities of laboratory $d \in \mathcal{V}^d$ with respect to specimen type $o \in \mathcal{V}^c$ is not exceeded.

4 Future work

The model proposed in the previous section may serve as a starting point in delivering a real-life applicable solution to pathological testing organisations. There are, however,

characteristics which have either been simplified or neglected in our model and which will have to be addressed in future in order to create a more realistic model representation.

One such limitation involves the evolution of data. In our formulation the input data were assumed to be static and known a *priori*, but in a real-life application the model would be more useful if it were able to accommodate a dynamic evolution of data and determine when disturbances in the data evolution are considerable enough to warrant triggering the computation of a new routing solution. *Dynamic* VRP formulations of this nature exist in the literature which are able to deal with information revealed over time concerning either customer demand or location.

The notion that it may be more profitable to outsource certain routes due to the isolated nature of certain specimen collection stations may also be investigated and is referred to as *Routing with Profits and Service Selection* in the literature. This variant was first introduced in conjunction with a TSP and then applied later to a VRP. Applying it specifically to pathological collection stations, it would be beneficial to combine the routing costs and profits into a single objective. This specific problem is referred to as *Profitable Tour Problem* and examples of such formulations appear in [2, 10, 12]. Also related to the possibly isolated locations of certain facilities is the option of collecting specimens from positions that are close enough to specimen collection stations. Formulations accommodating this option are referred to as the *Multi-Vehicle Covering Tour Problem* in the literature. Hachicha [11] developed three heuristics specifically to solve this problem.

Due to the nature of the problem formulation presented in §3, metaheurisitcs will be required to obtain good trade-off solutions to instances of our VRP. The VRP formulation proposed in this paper forms part of a larger, ongoing project on local health care decision support at Stellenbosch University. A comparative study of applying different metaheurisitcs to a real case study instance of our VRP will be performed as part of this ongoing project.

References

[1] Applegate DL, 2011, *The traveling salesman problem: A computational study*, Princeton University Press, Princeton (NJ).

[2] Archer A, 2011, *Improved approximation algorithms for prize-collecting Steiner tree and TSP*, SIAM Journal on Computing, **40(2)**, pp. 309–332.

[3] Baldacci R, Christofides N & Mingozzi A, 2008, *An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts*, Mathematical Programming, **115(2)**, pp. 351–385.

[4] Clarke GU & Wright JW, 1964, *Scheduling of vehicles from a central depot to a number of delivery points*, Operations Research, **12(4)**, pp. 568–581.

[5] *Consultation on technical and operational recommendation for clinical laboratory testing harmonization and standardization*, [Online; accessed 11th May 2015], Available at: `http://www.who.int/healthsystems/round11_9.pdf`.

[6] Contardo C, 2011, *A new exact algorithm for the multi-depot vehicle routing problem under capacity and route length constraints*, Technical Report, CIRRELT 2011-44, Université de Québec á Montreal, Québec.

[7] Dantzig GB & Ramser JH, 1959, *The truck dispatching problem*, Management Science, **6(1)**, pp. 80–91.

[8] Desrosiers J, Soumis F & Desrochers M, 1984, *Routing with time windows by column generation*, Networks, **14(4)**, pp. 545–565.

[9] Fukasawa R, 2006, *Robust branch-and-cut and price for the capacitated vehicle routing problem*, Mathematical Programming, **106(3)**, pp. 491–511.

[10] Goemans MX & Williamson DP, 1995, *A general approximation technique for constrained forest problems*, SIAM Journal on Computing, **24(2)**, pp. 296–317.

[11] Hachicha M, 2000, *Heuristics for the multi-vehicle covering tour problem*, Computers & Operations Research, **27(1)**, pp. 29–42.

[12] Nguyen VH, 2010, *A primal-dual approximation algorithm for the asymmetric prize-collecting TSP*, Combinatorial Optimization and Applications, **6508**, pp. 260–269.

[13] Toth P & Vigo D, 2014, *Vehicle routing: problems, methods, and applications*, MOS-SIAM Series on Optimization, SIAM, Philadelphia (PA).

Proceedings of the 2015 ORSSA Annual Conference
pp. 81–90
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Off-gas power generation optimisation using a mixed integer linear programming model

PvZ Venter* S.E. Terblanche† M. van Eldik‡

Abstract

In industry engineering plants typically have a variety of interlinked production chains, where process flows are dependent on upstream events and operated by default or manual settings. Burnable off-gasses, generated via operation processes, are in most cases utilized as energy sources. Raw material feeds may fluctuate over time, leading to varying off-gas production and potentially resulting in inefficient energy resource usage. It is common practice to generate steam from off-gas, in boiler houses, where excess steam is allocated for power generation. One of the problems is that unused off-gasses are burned and released atmosphere, where the energy potential is nullified. This paper proposes a mixed integer linear programming model that, under specified conditions, optimises power generation through the efficient use of energy from off-gasses. Empirical results are based on real world data and show the practical use of the model within a manufacturing context. Results generated by the model was compared to an engineering plant's operational philosophy and yielded an improved power generation potential.

Key words: Off-gas, Power generation optimisation, Mixed linear integer programming.

1 Introduction

The engineering manufacturing industry has numerous different types of plant layouts, where a plant refers to a process or combination of processes, delivering an end result but not necessarily limited to a single product. These processes have integrated production chains, where continuous process flows are dependent on upstream events. A plant may operate as an isolated entity or be integrated with other plants, forming part of a section or sub-section. For any production chain a number of by-products may be generated. A

*School for Mechanical and Nuclear Engineering, North-West University, Private Bag X6001, Potchefstroom, 2520, South Africa email: 12330825@nwu.ac.za

†Centre for Business Mathematics and Informatics, North-West University, Private Bag X6001, Potchefstroom, 2520, South Africa

‡School for Mechanical and Nuclear Engineering, North-West University, Private Bag X6001, Potchefstroom, 2520, South Africa

gaseous by-product, also known as an off-gas, may potentially combust when ignited in an oxygen enriched environment. These off-gasses may be utilized as energy resources.

It is common practice to generate steam, in boiler houses, from residual burnable off-gasses that are not otherwise consumed by the production processes. If sufficient excess steam is produced, power generating steam turbines may be installed to exploit the excess available energy. When raw material feeds to the process plants, including the chemical qualities thereof, experience variations over time, it could consequently lead to fluctuating off-gas production trends. Since steam flow production and turbine feeds are dependent on off-gas availability, for the engineering plant under consideration, inconsistent off-gas flows will effectively cause alternating power generation. Alternating power generation itself is not a problem for a plant where the core business is not power generation, but it is problematic if generation capacities go to waste, due to inefficient resource utilization.

To the best of the authors' knowledge no *Mixed Integer Linear Programming* (MILP) model currently exists, where the objective of power generation optimisation is driven by the optimal control of turbines, for the case of fluctuating steam.

Fazlollahi and Maréchal [2] investigated the integration of biomass and natural gas with energy systems, using multi-objective evolutionary algorithms and a MILP. The aim was to minimize cost and *Carbon Dioxide* (CO_2) emissions, while matching energy supply and demand, as part of the design and operation of the energy system. To link energy supply and demand, Fazlollahi *et al.* [3] investigated the design and operation of an energy system. A MILP was used in the developing of optimisation methods.

A MILP and Lagrangian relaxation was used by Thorin *et al.* [6] to develop a deterministic optimisation model for the optimisation of *Combined Heat-and-Power* (CHP) systems. A CHP system is used for power and heat generation and is also known as co-generation systems. For the sizing of a co-generation system, Beihong and Weiding [1] proposed a MILP, with constraints that includes energy demands and equipment effectiveness. The objective was the minimization of annual operating cost.

Fumero *et al.* [4] proposed a MILP for multi-product batch plants, when addressing the design and scheduling thereof. The objective was to minimize investment cost and still satisfy demand. For site utility system optimisation, Hui and Natori [5] discussed mixed integer programming models, both linear and non-linear. A multi-period mixed integer model was further introduced to optimise decisions related to the stop or start of equipment and boiler maintenance scheduling.

In this paper, a MILP model is proposed that, under specified conditions, optimises power generation through the efficient energy use from off-gasses. In the following section, technical aspects of the problem that needs to be considered in the model formulation are discussed. The complete model formulation is given in Section 3, followed by computational results based on real data.

2 Engineering considerations

An industry wide practice is to operate plants with manually controlled or default settings, without taking into account any instantaneous flow variations. These fluctuations typically result in inefficient off-gas energy resource utilization. In addition to this, unused off-gasses cannot be stored and are typically burned and released into atmosphere, nullifying the energy potential. When this off-gas production behaviour results in fluctuating steam productions, turbine steam flows will be affected. Every turbine has a minimum and maximum allowable mass flow limit. Turbines should furthermore only be tripped for planned interval maintenance. As for any equipment type, turbines also have an estimated life expectancy and every start-up depletes the projected life span, which implies that trip occurrences should be minimized. A protection measure for turbines in trip is to let a sufficient available steam period elapse before the machine may return back online.

Consider a plant that experiences fluctuating residual steam used for power generation. Figure 1 depicts hypothetical steam mass flows over time for such a plant. Furthermore, assume that the plant has only one turbine installed that operates between a minimum and maximum mass flow of 15 to 75 units. Given that the turbine is operational at time zero, it is clear that all steam flow above 75 units cannot be utilized for power generation. At time period 20, a dip below the minimum operational requirement is experienced. The turbine trips and all power generation potential are lost during the off time period from 20 till 80, whereafter start-up is initiated. During the time lapse from 20 till 80, there are some instances where the turbine could have been operational, but none of these periods provided an adequate time span to satisfy the minimum sufficient available steam criteria.

Figure 1 displays only a single occurrence where deficient steam causes a turbine to trip. When these occurrences are regularly observed, significant power generation capacity is lost and turbine life expectancy drastically shortened. For multiple turbines, residual steam needs to be distributed in such a manner that the system delivers maximum power generation over time. With sufficient available steam and all turbines in operation, it is only a matter of loading the turbines above the minimum flow requirements, starting at the most efficient and ending with the least efficient one.

A problematic situation is when all the machines cannot stay operational, due to limited steam supply. Not only the efficiency, but also the turbine capacity contributes to the decision to take a unit off line. It is not necessarily the best option to keep the most efficient turbine operational, or the machine with the smallest or largest generation capacity. When mass flow distributions to selected turbines are manually controlled or default settings are used, other turbines may trip, due to the controlled turbines that use more than the minimum required threshold. Not only is this problematic for the tripped machine, but as a consequence some residual steam and therefore power generation potential may go to waste.

As mentioned above, before every machine start-up, a predetermined time period needs to elapse. During this period, sufficient steam must be present at any given moment to support every operational turbine and still result in adequate available steam, to keep the turbines due for start-up online. When operation occurs manually, is it expected from the operator to decide if there was sufficient steam available during the preceding time

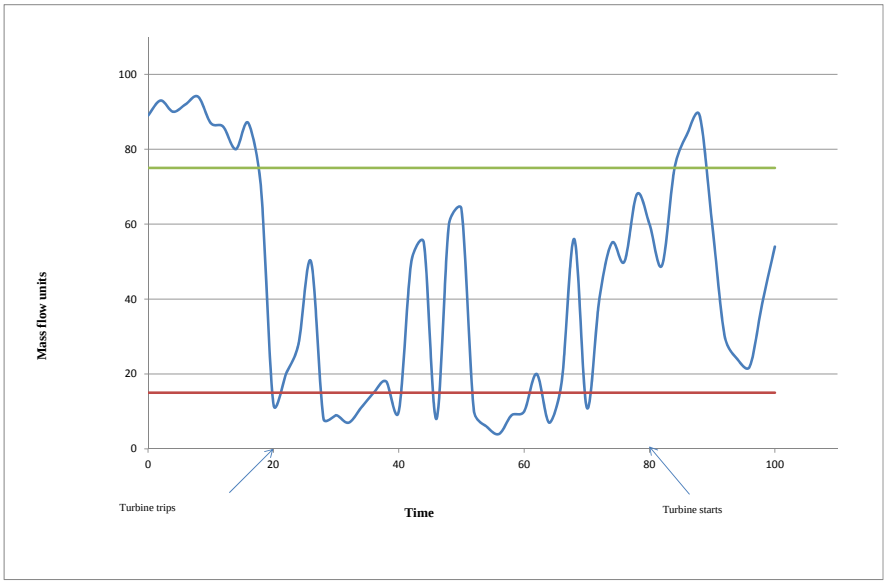


Figure 1: *Hypothetical steam flow scenario over time.*

period. It is fairly common that this time period carries on far longer than the requirement, especially during night or weekend shifts when there are fewer plant personnel.

3 Model

This proposed MILP model uses historic data and determines how steam distribution should have been controlled amongst any configuration of turbines for optimal power generation. The results can then be used for strategic operational scheduling or future planning of capital expenditures. Input to the model includes off-gas mass flows and plant steam usage over time. Existing technology can be implemented that will allow for accurate in-time off-gas mass flow predictions. These predictions will allow the MILP model to be used as an in-time control algorithm for power generation optimisation. This study works on a simplified model, not taking onto account the dynamic behaviour typically found in a plant set-up.

The primary decision variables are concerned with the operational status of each turbine over time. Although the objective function will aim to keep all turbines running at maximum capacity, the available steam at any point in time may trigger some turbines to receive less steam or even trip. An important factor to consider is that no turbine may ever be tripped if there is sufficient steam available to keep it operational. This constraint will always hold, even if tripping a turbine will yield more power generation.

The following parameter definitions are required to formulate the problem as a MILP

$\mathcal{I}^T$ index set of all turbines, where $\mathcal{I}^T = \{1, 2, ..., |\mathcal{I}^T|\}$,

$\mathcal{I}^G$ index set of all burnable off-gasses, where $\mathcal{I}^G = \{1, 2, ..., |\mathcal{I}^G|\}$,

δ^R index of start-up time range $\delta^R = \{1, 2, ..., |\delta^R|\}$,

$\mathcal{T}$ index set of time, where $\mathcal{T} = \{1, 2, ..., |\mathcal{T}|\}$,

M^B combined boiler house steam production capacity [ton/h],

m^g_{it} off-gas flow from gas $i \in \mathcal{I}^G$ at time $t \in \mathcal{T}$ [m^3/h],

m^s_t total residual steam at time $t \in \mathcal{T}$ [ton/h],

m^W_t total steam consumed by the works at time $t \in \mathcal{T}$ [ton/h],

CV_i calorific value of gas $i \in \mathcal{I}^G$ [MW/ton],

δ minimum time with sufficient steam availability that needs to elapse before a tripped turbine may be started up [h],

η^T_i efficiency of turbine [ton/kW],

f^s_i conversion factor for off-gas $i \in \mathcal{I}^G$ into steam [ton/MWm^3],

L_i lower bound on steam capacity for turbine $i \in \mathcal{I}^T$ [ton/h], and

U_i upper bound on steam capacity for turbine $i \in \mathcal{I}^T$ [ton/h].

In order to model the operation, the following decision variables are required:

m^s_{it} mass flow of steam to turbine $i \in \mathcal{I}^T$ at time $t \in \mathcal{T}$ [ton/h],

$$y^T_{it} = \begin{cases} 1 & \text{if turbine } i \in \mathcal{I}^T \text{ is operational at time } t \in \mathcal{T}, \\ 0 & \text{if turbine } i \in \mathcal{I}^T \text{ is in trip at time } t \in \mathcal{T}, \end{cases}$$

$$y^L_{it} = \begin{cases} 1 & \text{if turbine } i \in \mathcal{I}^T \text{ is in trip for at least } \delta \text{ hours, up till time } t \in \mathcal{T}, \\ 0 & \text{if not,} \end{cases}$$

$$y^o_{it} = \begin{cases} 1 & \text{if turbine } i \in \mathcal{I}^T \text{ is operational from time } t-1 \in \mathcal{T} \text{ to } t \in \mathcal{T}. \\ 0 & \text{if not.} \end{cases}$$

$$y^b_{it} = \begin{cases} 1 & \text{if turbine } i \in \mathcal{I}^T \text{ that is in trip at time } t-1 \in \mathcal{T}, \text{ may be brought back online at time } t \in \mathcal{T}, \\ 0 & \text{if not,} \end{cases}$$

$$y^h_{it} = \begin{cases} 1 & \text{if turbine } i \in \mathcal{I}^T \text{ is tripped from time } t-1 \in \mathcal{T} \text{ to } t \in \mathcal{T}, \text{ and} \\ 0 & \text{if not.} \end{cases}$$

The model starts with off-gas mass flows, either predicted or measured. Under the assumption that an optimal strategy is used by the operational plant to utilize all possible off-gasses, within boiler restrictions, the residual steam calculation is

$$m^s_t = \min\left(M^B, \sum_{i \in \mathcal{I}^G} (m^g_{it} CV_i f^s_i)\right) - m^W_t, \quad \forall t \in \mathcal{T}. \tag{1}$$

The objective is to optimise power generation over the time horizon $\mathcal{T}$, *i.e.* the summation of steam flow to each turbine over the efficiency of that turbine

$$\max\left(\sum_{i \in \mathcal{I}^T} \sum_{t \in \mathcal{T}} m^s_{it} / \eta^T_i\right). \tag{2}$$

The first set of constraints ensures that total steam distribution to all turbines at time

$t \in \mathcal{T}$, may not exceed the available steam for that time period

$$\sum_{i \in \mathcal{I}^T} m_{it}^s \leq m_t^s, \quad \forall t \in \mathcal{T}. \tag{3}$$

Steam distribution to each turbine, at any time $t \in \mathcal{T}$, may not exceed the maximum allowable flow to that machine

$$m_{it}^s \leq U_i y_{it}^T, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T}. \tag{4}$$

For every operational turbine at time $t \in \mathcal{T}$, the steam distribution may not be less than the minimum rated flow

$$m_{it}^s \geq L_i y_{it}^T, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T}. \tag{5}$$

Constraint set (6) ensures that turbine $i \in \mathcal{I}^T$ is operational at time $t-1 \in \mathcal{T}$, may not be tripped at time $t \in \mathcal{T}$, if there exists sufficient steam to keep it operational

$$m_{it}^s y_{it}^h - \sum_{j \in \mathcal{I}^T} (y_{jt}^T L_j) \leq y_{it}^h L_i, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > 1. \tag{6}$$

Constraint sets (7) and (8) check whether turbine $i \in \mathcal{I}^T$ that was operational at time $t-1 \in \mathcal{T}$, is still operational at time $t \in \mathcal{T}$

$$y_{it-1}^T + y_{it}^T \leq 1 + y_{it}^o, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta \text{ and} \tag{7}$$

$$y_{it-1}^T + y_{it}^T \geq 2y_{it}^o, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta. \tag{8}$$

Constraint sets (9) to (11) determine whether turbine $i \in \mathcal{I}^T$ that was operational at time $t-1 \in \mathcal{T}$, is tripped at time $t \in \mathcal{T}$

$$y_{it-1}^T - y_{it}^T \leq y_{it}^h, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > 1, \tag{9}$$

$$y_{it-1}^T \geq y_{it}^h, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > 1, \text{ and} \tag{10}$$

$$y_{it}^T \leq 1 - y_{it}^h, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > 1. \tag{11}$$

Constraint sets (12) to (14) determine if the minimum time period has elapsed for a turbine in trip. Steam availability is not yet accounted for in these constraints

$$\sum_{d \in \delta^R} (1 - y_{it-d}^T) \geq \delta y_{it}^L, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta \text{ and} \tag{12}$$

$$\sum_{d \in \delta^R} (1 - y_{it-d}^T) \leq \delta - 1/2 + y_{it}^L, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta. \tag{13}$$

For the time period $t \in \delta^R$, a turbine may be started up if there is sufficient steam during that period

$$y_{it}^L = 1, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t < \delta + 1. \tag{14}$$

If turbine $i \in \mathcal{I}^T$ is off-line and the minimum δ has not yet passed, the machine may not be brought back online

$$y_{it}^T \leq y_{it}^L + y_{it-1}^T, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > 1. \tag{15}$$

The following set of constraints is to verify that sufficient steam existed for turbine $i \in \mathcal{I}^T$ in trip, to have been kept operational during the previous δ time period

$$\sum_{i \in \mathcal{I}^T} ((y^T_{it-d} + y^b_{it}) L_i) \leq m^s_{t-d}, \quad \forall t \in \mathcal{T} : t > \delta, d \in \delta^R. \tag{16}$$

To verify that turbine $i \in \mathcal{I}^T$ in trip may be brought back online, constraint sets (17) to (19) are used

$$y^b_{it} \geq y^T_{it} - y^T_{it-1}, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta, \tag{17}$$

$$y^b_{it} \leq |y^T_{it} - y^T_{it-1}|, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta, \text{ and} \tag{18}$$

$$y^T_{it} - y^T_{it-1} + 1.5 \geq y^b_{it}, \quad \forall i \in \mathcal{I}^T, t \in \mathcal{T} : t > \delta. \tag{19}$$

In the following section this MILP model is implemented on real world data to illustrate its practical use within a manufacturing plant context. Scenario results from the MILP are compared to that of the engineering plant's operation philosophy. The MILP solved within 5 minutes. Future expansion of the MILP will lead to a larger time calculation intensive model.

4 Optimisation results

From real measured steam flow productions and off-gases burned into atmosphere, potential residual steam was determined for this engineering plant during the time period. The results in Figure 2 display the potential residual steam in ton/h over the time period. This potential steam flow availability was used for two scenarios; one where the manual operation philosophy of the plant was used and another where the MILP model was incorporated. Steam production fluctuations are evident in Figure 2 and will reflect in the power generation.

Typical power generation capacity for this set-up is the ability to generate up to 35 MW. Assume that this plant's power generation potential consists of two turbines, namely a 10 MW and a 25 MW machine and that both turbines have the same conversion efficiency from steam to electricity, *i.e.* 5 ton/MW. For the turbines to stay operational the 10 MW machine requires a minimum steam flow of 17.5 ton/h, whereas the 25 MW machine requires a minimum flow of 25 ton/h. This plant has a power generation philosophy that is manually operated. The manual operation philosophy is to set the smaller turbine at an 80% power generation capacity and thereafter load the 25 MW turbine. The 25 MW turbine will thus only be operational if the 10 MW machine is generating power at a rate of 8 MW and sufficient steam is available to sustain the 25 MW turbine. Under manual operations it is realistic to assume that no start-up will occur during the hours of 0:00 till 6:00 A.M. and on Sundays. A typical, plant specific, time period of 15 hours is chosen for the minimum sufficient steam criteria and will hold for both the manual operations and MILP model.

Figure 3 depicts the two turbine's scenario results for the manual operation procedure and the MILP control optimisation model. The two left side graphs show the results for the manually operated system and the two right handed graphs that of the MILP model. The

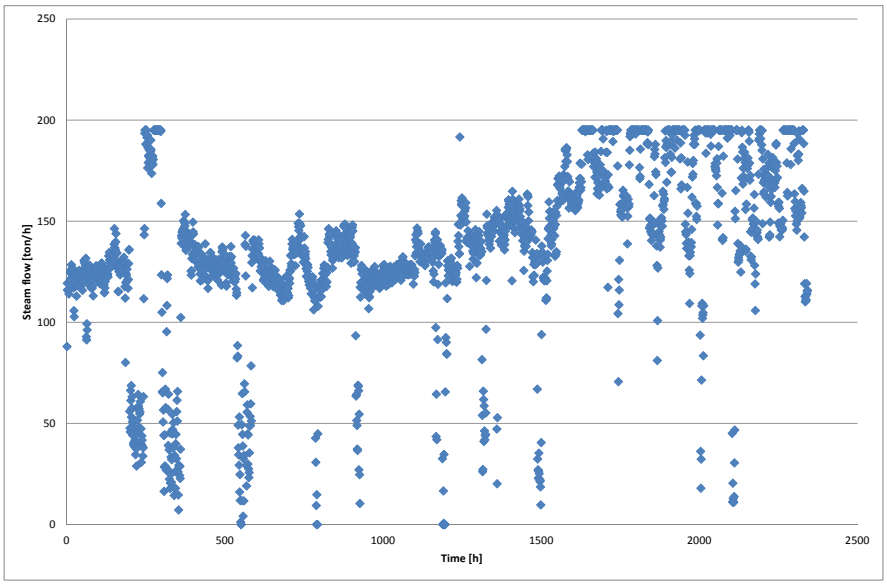


Figure 2: *Hypothetical steam flow scenario over time.*

manual operation strategy to keep the 10 MW turbine at a fixed operating point of 8 MW is clearly evident from the top left graph. The 10 MW turbine does not exceed a power generation rate of 8 MW and operates at this set point for over 88% of the time.

As mentioned above, the two right hand sided graphs of Figure 3 are generated results from the MILP control algorithm. Under the MILP model, power generation starts with only the 25 MW turbine that is operational. Once the 10 MW turbine is brought online, there is a visible difference between the operation strategy from the manual to MILP approach. Whereas the manually operated 10 MW turbine is operating at the 8 MW set point for over 88% of the time, the optimisation model permits the 10 MW machine to receive any possible flow, within the operational steam flow limits and simultaneously allows for the 25 MW turbine to receive steam. For the MILP model, the 10 MW turbine operates at a rate of 8 MW and higher for 25% of the time. When comparing the 25 MW turbine for the two different operational approaches, the MILP optimisation model clearly allows for the 25 MW turbine to receive more steam flow over the time period. For the MILP model power generation from the 25 MW turbine is at a rate of 20 MW and higher over 82% of the time, compared to the manual operations that allow for less than 30% of the time to generate power in this interval.

The manual operations scenario yields an average power generation of 7.30 MW for the 10 MW with 8 trips while 16.86 MW is generated by the 25 MW turbine with 12 trips. Power generation potential lost due to the inability to utilize the steam amounts to 3.39 MW for the time period. Under the MILP model, the 10 MW turbine produces an average power generation of 4.20 MW with 8 trips while the 25 MW machine delivers 20.89 MW with 9 trips. The combined power generation amounts to an average of 25.09 MW, 0.93 MW or 3.82% higher than for the manual operations. Even though the objective of the MILP model is to optimise power generation and not to reduce turbine trips, operation under the control algorithm yields a 15% reduction in combined turbine trips. If the cost

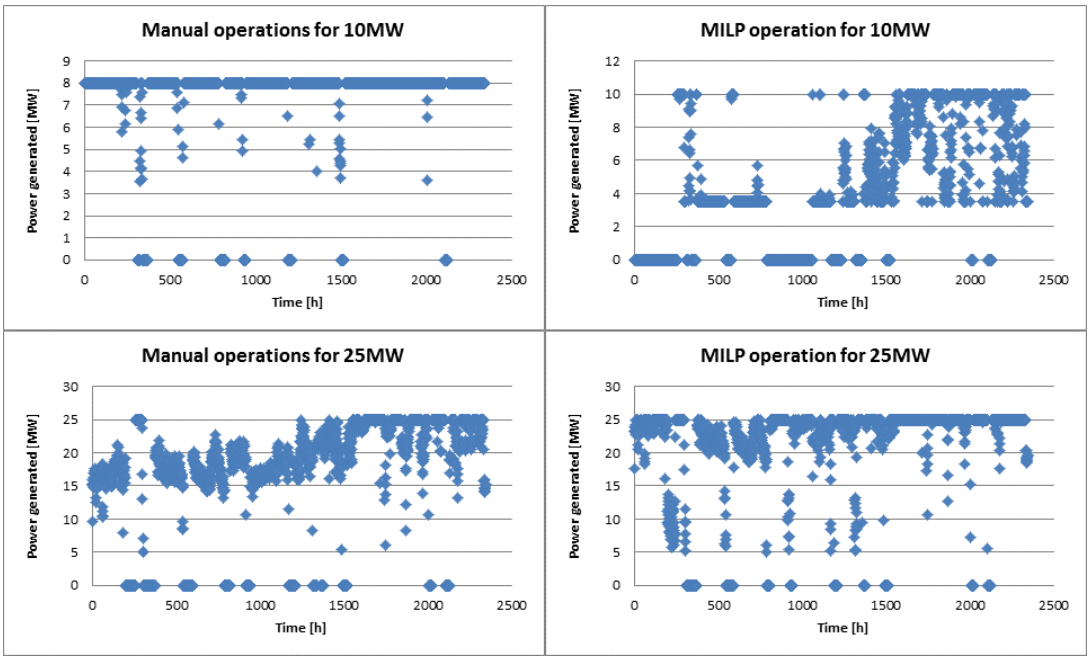


Figure 3: *Generated results for power generation per turbine over time interval.*

of electricity for the company is R0.50 /kWh, then a saving of 0.93 MW will result in an electricity cost saving of approximately R4 million over a one year period. Note that this potential saving would result from optimising available energy resources and not changing plant or production processes.

5 Summary and conclusion

For the engineering production industry a scenario was depicted where process off-gasses generate steam in boiler houses with residual steam being utilized for power generation. For this scenario a MILP optimisation model was developed. Under specified constraints, the model maximizes power generation in an environment where turbines do not always receive full capacity steam flow and even with times need to be tripped, due to steam shortages. The model was implemented, using data of an existing engineering plant and yielded a potential 3.82% increase in power generation and a 15% reduction in combined turbine trips, when compared to the specified manual operating philosophy, over the time period investigated.

References

[1] Beihong Z & Weiding L, 2006, *An optimal sizing method for cogeneration plants*, Energy and Buildings, **38**, pp. 189–195.

[2] Fazlollahi S & Maréchal F, 2013, *Multi-objective, multi-period optimization of biomass conversion technologies using evolutionary algorithms and mixed integer linear programming (MILP)*, Applied Thermal Engineering, **50**, pp. 1504–1513.

[3] Fazlollahi S, Mandel P, Becker G & Maréchal F, 2012, *Methods for multi-obejtive investment and operating of complex energy systems*, *Energy*, **45**, pp. 12–22.

[4] Fumero Y, Corsano G, & Montagna JM, 2013, *A Mixed Integer Linear Programming model for simultaneous design and scheduling of flowshop plants*, *Applied Mathematical Modelling*, **37**, pp. 1652–1664.

[5] Hui CW & Natori Y, 1996, *An industrial application using mixed-integer programming technique: a multi-period utility system model*, European Symposium on Computer Aided Process Engineering, **6(B)**, pp. S1577–S1582.

[6] Thorin E, Brand H & Weber C, 2005, *Long-term optimization of cogeneration systems in a competitive market environment*, *Applied Energy*, **81**, pp. 152–169.

Proceedings of the 2015 ORSSA Annual Conference
pp. 91–101
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Radio transmission tower placement in cellular telephone communication networks

T Schmidt-Dumont* JH van Vuuren†

Abstract

Mobile telecommunication has become an essential communication channel in the modern world. Network providers are faced with the challenge of providing as many people in as many different areas as possible with network service. Multiple factors have to be taken into account when radio transmitter placement decisions are made. Generally, maximum area terrain surface coverage, as well as fail-safe mutual area coverage by at least two transmitters are of prime importance. This results in a bi-objective facility location problem with the goal of achieving an acceptable trade-off between maximising total area coverage by all transmitters in the network, and maximising areas covered by at least two radio transmitters. The network planning problem for second generation networks can be decomposed into the above-mentioned coverage problem and a subsequent real-time frequency assignment problem. For technical reasons, the frequency assignment and coverage objectives cannot be separated in third and fourth generation networks. The focus of this paper is on the planning of second generation networks. In particular, a suitable framework is proposed for evaluating the effectiveness of a given set of placement locations for a network of radio transmitters with respect to both maximum total and mutual area coverage, taking into account obstruction of the line of sight and the first Fresnel ellipsoid between the transmitter and the receiver, which are required to be unobstructed for effective transmission. This is followed by the formulation of a bi-objective facility location model suitable for use as the basis of a decision support system for identifying high-quality trade-offs between maximising total area coverage and maximising mutual area coverage.

Key words: Facility location, Transmitter location, Wireless network planning.

1 Introduction

Mobile telecommunication has revolutionised the modern world. Smartphones and similar devices are used on a daily basis to communicate through many different types of electronic media. This has sparked a trend, especially among the younger generation, of always

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: 17000807@sun.ac.za

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

having to be connected and up-to-date on what is going on, not only in their own lives, but also in the lives of others. Similar trends may be seen among business people, who can now use their smartphones or tablets to complete almost any business transaction. It has, therefore, become an absolute priority for cellular telephone network providers to cover as much area as possible in their service provision. Currently, network providers have the choice of using a combination of second, third or fourth generation networks for their service provision.

Second Generation (2G) networks use the *Time Division Multiple Access* (TDMA) protocol to partition the bandwidth bought by the mobile provider into frequency channels of a specific bandwidth, generally in the order of 200 kHz. These channels are then assigned to receivers, each call having one channel allocated to it. Should there be no available channels when a new call is made, the call is blocked until a channel opens to which the call can be assigned. The allowable time during which a call can be blocked is limited and depends on the policy adopted by the network provider. If a channel does not open up during the allowable blocking time, the call is terminated. Channel assignment to the receiver can be performed in one of two ways: *Fixed Channel Assignment* (FCA) or *Dynamic Channel Assignment* (DCA) [6]. In FCA, each base station only has a limited number of channels allocated to it over the available frequency band. These are then assigned to the receivers as calls are made. In DCA, however, the entire bandwidth is available for use by all transmitters and different assignment policies are in place according to which new calls are treated. The general objective in any channel assignment policy is to minimize the number of blocked calls. The channel assignment is usually the final step of the network planning process, but is an operational task, as opposed to a strategic task, and hence repeated in an online fashion, whereas other planning aspects, such as transmitter placement decisions, are strategic and are performed once-off in an off-line fashion [6].

Apart from 2G networks, network providers may also choose from either *Third Generation* (3G) or *Fourth Generation* (4G) networks, also known as *Long Term Evolution* (LTE) networks. These networks use different bandwidth assignment protocols, are more focused on achieving high data download speeds, and are thus usually established in urban locations, where the demand for high download speeds is ever growing. 2G networks, however, are more focused on voice transmission and as a result more common in semi-urban or rural areas.

The choice of the type of network and the resulting placement of radio transmitters forming the network is of primary importance to network providers, especially when taking into account the prospective growth of smartphone users in Africa. Reed *et al.* [8] state that "the number of smartphone connections will rise from about 79 million at the end of 2012 to 412 million by 2018, according to forecasts by Informa." It is, however, not only the number of new smartphone connections that is expected to achieve such impressive growth. 2G networks and feature phones are expected to remain a key aspect of mobile networks in Sub Saharan Africa where, due to the relatively low *Gross Domestic Product* (GDP), smartphones remain beyond reasonable levels of affordability for a large portion of the population. This will especially be the case in semi-urban and rural areas, where new mobile networks are established [2].

Decision support frameworks for transmitter location problems in cellular telephone net-

works in the literature are mostly based on single-objective optimisation models. In rare instances where multiple placement objectives are incorporated into the underlying optimisation models, these objectives are usually combined into a single weighted-sum model objective. In non-convex models this practice of combining objectives is known to mask Pareto-optimal solutions. Instead, a bi-objective modelling framework is proposed in this paper for uncovering high-quality trade-off solutions to the radio transmitter location problem.

The paper is organised as follows. A concise review is given in §2 of the literature on facility location models which have been used previously to solve similar network planning problems. In §3, a framework is established for evaluating the effectiveness of a given set of transmitter locations. This framework takes into account the obstruction of the line of sight as well as the first Fresnel ellipsoid between transmitter and receiver, both of which are required for effective transmission. A bi-objective facility location model suitable for use as the basis of a decision support system able to identify high-quality trade-off solutions between maximising total area, and maximising mutual area coverage is put forward in §4. Some preliminary computational results are presented in §5 in order to demonstrate the working of the modelling approach. The paper finally closes with a brief conclusion and ideas for future work in §6.

2 Literature Review

For second generation networks, the network planning problem may be decomposed into two distinct phases: *coverage planning*, which involves antennae placement in order to achieve maximum service coverage, and *capacity planning* which involves frequency assignment planning [1]. The coverage planning problem has generally been modelled using variations on the celebrated set covering problem described in the operations research literature. Amaldi *et al.* [1] describe this problem, known in the context of radio transmitter network planning as the *coverage problem*, as follows: Given an area where service provision has to be guaranteed, determine those locations where the radio transmitters should be placed and specify their configurations such that each point (or user) in the service area receives an adequate signal level.

Two main modelling approaches have been adopted in the literature to solve instances of the coverage problem [1]. The first approach follows a continuous optimisation strategy. A specified number, k (say), of base stations are to be located at any site within the given space which is to be covered, where the antennae co-ordinates are the continuous variables of the problem. This space may exclude certain forbidden areas in which no transmitter placements are allowed. In certain cases, other parameters, such as the antennae orientations and/or the transmission power may also be considered as variables. Amaldi *et al.* [1] claim that the most important element of this type of optimisation model is the propagation prediction model used to estimate the signal intensity at each point in the coverage area. Various functions have been developed over the years for signal estimation, ranging from simple empirical models, such as those developed by Hata [3], to more sophisticated ray tracing methods, such as that discussed by Iskander and Yun [4]. The objective function of the coverage problem is usually determined by some measure of the

quality of service, such as the largest minimum signal intensity at any location [1]. Due to the high complexity of typical propagation loss functions, global optimisation techniques are usually employed to tackle these problems, as illustrated by Sherali *et al.* [9].

The second coverage problem modelling approach involves the use of discrete mathematical models. In this approach, a number of test sites or demand nodes representing users of the network have to be identified in the service area. Instead of allowing base stations to be placed at any location in the coverage area, discrete mathematical models restrict the positioning of these base stations to a set of so-called *candidate sites*. In these models, the area covered by each base station is determined *a priori*, generally using a radio wave propagation predictor and taking the surrounding topology and morphology of the terrain into account [6]. The area covered by each candidate site is therefore assumed to be known in such an optimisation model.

Krzanowski and Raper [5] explain that in both the continuous and discrete modelling paradigms, *total cover problems* require the determination of the minimum number of facilities required to meet all the demand. *Partial cover problems* arise in contrast when the number of facilities to be placed is fixed and the locations have to be chosen so as to maximise the demand that can be covered using the limited number of facilities. A further extension of the partial cover problem is the so-called *general cover problem*, in which the objective is to minimise the maximum distance between a facility and the demand points it covers. Mathar and Niessen [6] demonstrate how the coverage problem is an extension of the classical minimum cost set covering problem in the operations research literature.

Due to the large dimensions of the optimisation problems typically involved in radio transmitter facility location planning problems, metaheuristics are often employed as approximate optimisation techniques. Simulated annealing has, for example, been used by Mathar and Niessen [6] in an instance where the complexity of the optimisation problem places an optimal solution out of reach. Krzanowski and Raper [5] instead used a hybrid genetic algorithm designed to take the surrounding geography into account during the site selection process.

3 Measuring Coverage

For an area to be considered covered, an unobstructed line of visibility between the transmitter and receiver should at the very least be achieved. If this direct line of visibility is obstructed, then a situation of *Non Line Of Sight* (NLOS) is said to prevail. Radio wave transmission does, however, not only depend on a clear line of visibility between transmitter and receiver. Radio transmission generates an infinite family of nested ellipsoids called *Fresnel ellipsoids*. These ellipsoids all have both the transmitter and receiver at their foci. For effective transmission, the innermost of this family of ellipsoids, called the *first Fresnel ellipsoid*, should also be unobstructed. If an unobstructed line of visibility exists between a transmitter and receiver, but the first Fresnel ellipsoid is partially obstructed, *Near Line Of Sight* (nLOS) is said to have been achieved, whereas if both the direct line of visibility and the first Fresnel ellipsoid between the transmitter and receiver are unobstructed, then (full) *Line Of Sight* (LOS) is said to have been achieved. These notions are graphically illustrated in Figure 1 (a)–(c).

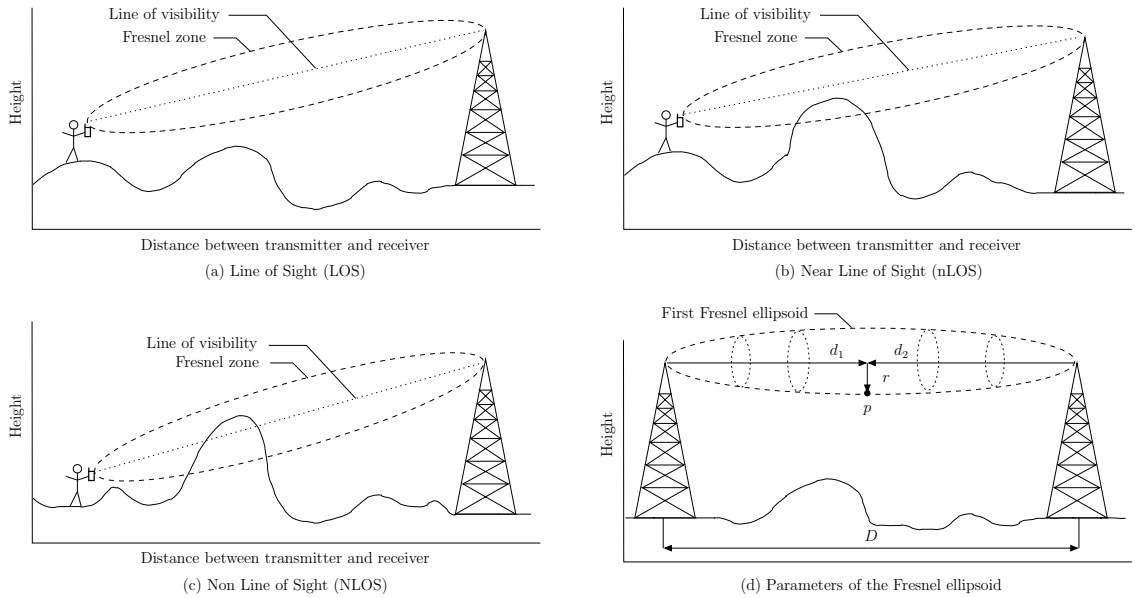


Figure 1: *Various notions related to line of sight and the first Fresnel ellipsoid between a radio transmitter and receiver.*

The radius of the first Fresnel ellipsoid at any point p between the transmitter and receiver is given by

$$r = \sqrt{\frac{\lambda d_1 d_2}{D}}, \tag{1}$$

where d_1 represents the horizontally projected distance between p and the transmitter, d_2 represents the horizontally projected distance between p and the receiver, $D = d_1 + d_2$ is the total distance between the transmitter and receiver, and λ represents the wavelength of the transmitted signal. These parameters are graphically illustrated in Figure 1 (d).

Our decision support framework for radio transmission tower placement is based on a discrete facility location modelling approach, as discussed in §2. The input to the process of determining coverage of an area by a given set of transmitters is a matrix of entries corresponding to a rectangular grid of placement candidate sites (which are also the coverage demand points) containing terrain elevations above sea level for some specified area of interest. For a demand point in this area to be considered covered by a potential transmitter, an unobstructed LOS (*i.e.* an unobstructed line of visibility as well as an unobstructed first Fresnel ellipsoid) has to exist between the transmitter and the demand point. Bresenham's well-known line drawing algorithm, which is widely used in computer graphics to determine which pixels need to be coloured in when drawing straight lines on screen displays, may be used to determine those entries in the matrix which form the line between the transmitter and receiver locations under investigation. Detailed information on Bresenham's line drawing algorithm may be found in [7].

At each of the demand points along the line determined by Bresenham's line drawing algorithm, the difference in elevation between the lower boundary of the first Fresnel ellipsoid and the demand point's elevation above sea level are compared. This is done

using the equation of the straight line of visibility between the transmitter and the demand point, and subtracting the radius of the first Fresnel ellipsoid from the height of the line of visibility. The distances d_1, d_2 and D in (1) may be approximated using the theorem of Pythagoras. Only if the elevation of the lower boundary of the first Fresnel ellipsoid between transmitter candidate site i and demand point j is higher than the elevation above sea level for all points along the line determined by the Bresenham line drawing algorithm between i and j, the demand point is considered to be covered by the transmitter candidate site. In this case we populate the entry in row i and column j of a *coverage matrix* with the value $a_{ij} = 1$. Otherwise the value $a_{ij} = 0$ is entered into the coverage matrix.

4 Mathematical Model

Suppose that k transmission towers are to be located at some subset of the transmitter candidate sites, as described in §3. A coverage importance value c_j and a mutual importance value C_j is associated with candidate site $j \in \{1, \ldots, n\}$. The aim of the model is to achieve an acceptable trade-off between maximising the accumulated coverage importance value z of candidate sites actually covered by the k transmission towers and maximising the accumulated mutual importance value Z of those candidate sites that are covered by at least two of the k transmission towers.

We employ the decision variables

$$x_i = \begin{cases} 1 & \text{if a radio transmission tower is placed at site } i \\ 0 & \text{otherwise} \end{cases}$$

for all $i = 1, \ldots, n$ as well as the auxiliary variables

$$y_j = \begin{cases} 1 & \text{if site } j \text{ is covered by at least one transmission tower} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Y_j = \begin{cases} 1 & \text{if site } j \text{ is covered by at least two transmission towers} \\ 0 & \text{otherwise} \end{cases}$$

for all $j = 1, \ldots, n$. The objectives are to

$$\text{maximise } z = \sum_{j=1}^{n} c_j y_j \tag{2}$$

and to

$$\text{maximise } Z = \sum_{j=1}^{n} C_j Y_j \tag{3}$$

subject to the constraints

$$\sum_{i=1}^{n} x_i \quad \leq \quad k \tag{4}$$

$$\sum_{i=1}^{n} a_{ij}x_i \geq y_j, \qquad j = 1, \ldots, n \tag{5}$$

$$\sum_{i=1}^{n} a_{ij}x_i \geq 2Y_j, \qquad j = 1, \ldots, n \tag{6}$$

$$x_i, y_i, Y_i \in \{0, 1\}, \qquad i = 1, \ldots, n. \tag{7}$$

In the above formulation, the objectives in (2) and (3) are conflicting in the sense that increasing z (usually achieved by spacing the transmission towers far apart) typically decreases Z which is, in turn, maximised by placing transmission towers not too far apart. Constraint (4) restricts the number of transmission towers placed to at most k. Constraint sets (5) and (6) are respectively the coverage and mutual coverage requirements. In these linking constraints, the parameter a_{ij} takes the value 1 if the first Fresnel ellipsoid between sites i and j is sufficiently unobstructed, as described in §3. Finally, constraint set (7) enforces the binary nature of the model variables.

5 Model Solution

The framework for determining coverage of a given area as described in §3, as well as the mathematical model of §4, is applied in this section to a real data set containing the elevation data for an area in the Western Cape. The set contains the elevation data for $n = 400$ points forming a 20×20 matrix. The latitude distance between two successive points in the matrix is 308.1 metres, while the longitude distance between two successive points is 256.6 metres. A surface plot of these elevation data is shown in Figure 2.

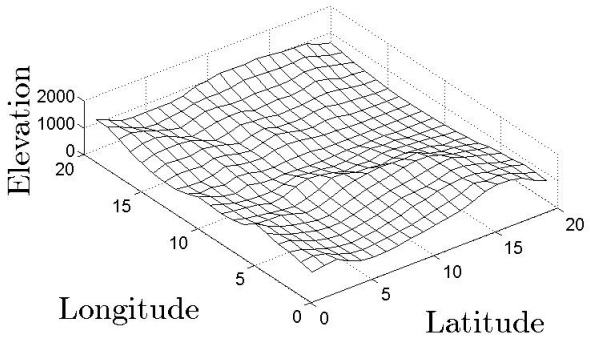


Figure 2: *Surface plot of the elevation data used in the model.*

The frequency used to determine the wavelength required for the calculation of the radii of the first Fresnel ellipsoids is 1 800 MHz, which is commonly used in 2G networks. At this frequency, the wavelength λ is 0.167 metres. The results of the framework for determining coverage are shown in Figure 3 (a)–(b). Figure 3 (a) corresponds to the placement of a transmission tower at latitude position 1 and longitude position 5, while Figure 3 (b) corresponds to the placement of a transmission tower at latitude position 8 and longitude position 1. In these plots, the white area represents the area covered by the transmitter,

whereas the black area depicts the parts not covered by the transmitter. For the purpose of demonstration, a road running through the area under consideration is given a single coverage importance value of 1. This importance matrix contains the values of $c_1, \ldots, c_{400}$ in (2) and is shown graphically in Figure 4 (a) where white denotes ones and black denotes zeros. Parts of two towns in the bottom left-hand and top right-hand corners are given a mutual coverage importance rating of 1, as can be seen in Figure 4 (b). The mutual coverage matrix depicted in Figure 4 (b) using the same colour coding scheme contains the values of $C_1, \ldots, C_{400}$ in (3).

The Pareto fronts for this instance of the model (2)–(7) are shown in Figure 5 for the situations in which $k = 2$ and $k = 3$ transmission towers are to be placed. The corresponding locations of the transmitters, together with the percentages of both the single and mutual area coverage achieved by these locations, (measured as the fraction of the areas associated with an importance weighting of 1), are shown in Tables 1 and 2.

Solution	Location 1	Location 2	Single Coverage	Mutual Coverage
1	(10,7)	(11,7)	78.79%	88.06%
2	(1,6)	(9,8)	100%	71.64%
3	(10,7)	(9,8)	93.94%	86.57%
4	(11,7)	(8,9)	96.97%	82.09%

Table 1: *The percentages of the single area as well as mutual area demand coverage achieved by the Pareto-optimal transmission tower placement pairs, located at Location 1 and Location 2.*

It is interesting to note that most of the Pareto-optimal transmitter placements, returned by the model occur in the area ranging from (13,6) to (8,9). This indicates that the best coverage will be achieved by placing a transmitter in that region. Another interesting fact is that in both the cases where $k = 2$ and where $k = 3$, there is a solution covering all the single area coverage demand points which have been given an importance weighting of $c_j = 1$ in (2). In neither case, however, could all the demand for mutual coverage, corresponding to demand points which had been assigned a value $C_j = 1$ in (3) be met. This indicates that in order to provide mutual coverage in those areas, at least four transmitters would have to be placed.

Solution	Location 1	Location 2	Location 3	Single Coverage	Mutual Coverage
1	(13,6)	(12,7)	(8,9)	96.97%	94.03%
2	(13,6)	(8,9)	(17,19)	100%	91.04%
3	(13,6)	(11,7)	(18,20)	69.70%	95.52%

Table 2: *The percentages of the single area as well as mutual area demand coverage achieved by the Pareto-optimal transmission tower placement triples, located at Location 1, Location 2 and Location 3.*

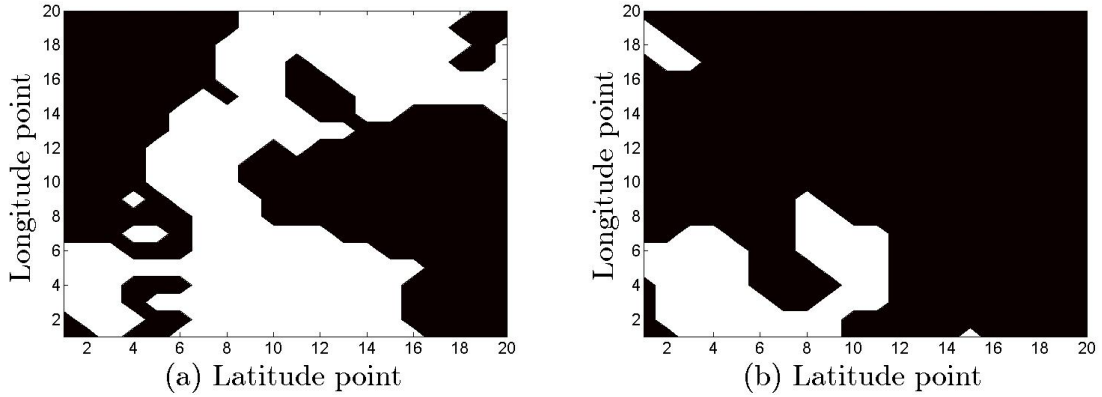


Figure 3: *Viewshed of the area covered by a transmitter placed at (a) point (1,5) and (b) point (8,1).*

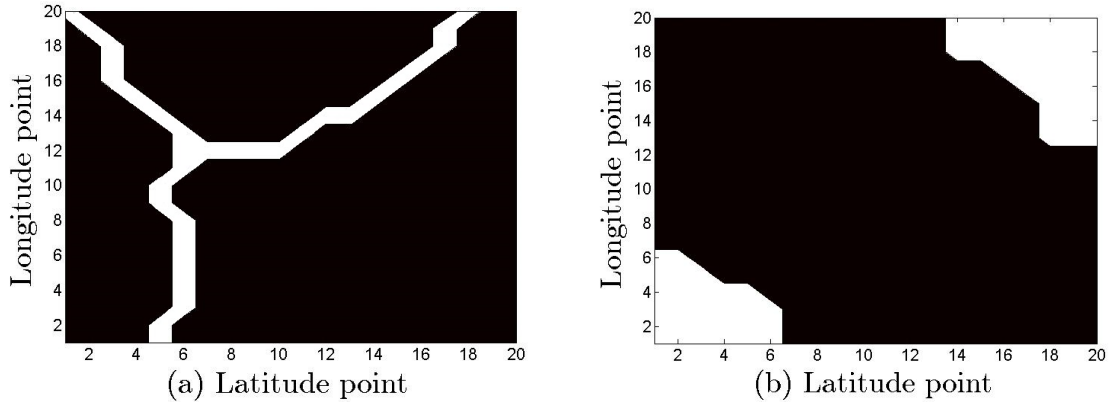


Figure 4: *Contour plot illustrating the importance rating of the area which is to be covered (a) by at least one transmitter and (b) by at least two transmitters.*

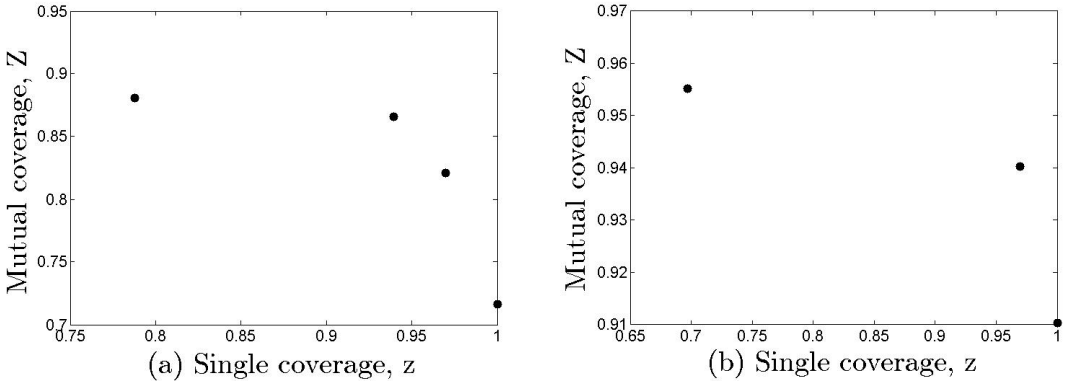


Figure 5: *The Pareto fronts for (a) $k = 2$ and (b) $k = 3$ transmitter placements.*

6 Future Work

The work reported in this paper forms part of a larger, ongoing project on facility location decision support at Stellenbosch University involving various researchers. The next step will be to incorporate radio signal propagation loss into the framework for the evaluation of coverage of a given set of transmitter locations. The propagation loss at a point r incurred from a transmitter located at a point r_0 is defined as the ratio of transmitted power at r_0, $P_t(r_0)$, to the received power at r, $P_t(r)$ [4]. The propagation loss of electromagnetic waves in free space is given by

$$L(dB) = 10 \log \left[\frac{P_t(r_0)}{P_r(r)} \right] = -10 \log \left[\frac{G_t G_r \lambda^2}{(4\pi)^2 D^2} \right], \tag{8}$$

where G_t represents the transmitter gain, G_r represents the receiver gain, λ represents the wavelength of the transmitted signal and D is the distance between the transmitter and the receiver. The propagation loss may not exceed a specified threshold at which the power reaching the receiver will be too small to ensure effective transmission. The expression in (8), however, only allows for the computation of losses along a path loss in free space, ignoring any obstructions by trees, buildings or similar structures which may also have a significant influence on the reduction of the transmitted power. To be able to take this into account, the model developed by Hata and Okumura [3] for path losses in areas where such obstructions may occur is applicable.

The Pareto-optimal solutions reported in §5 were found by brute force (*i.e.* considering all $\binom{400}{2} = 79\,800$ and all $\binom{400}{3} = 10\,586\,800$ location combinations, respectively). As a result, of the high computational complexity, only small instances of the transmitter placement facility location problem can be thus solved. A suitable metaheuristic, such as simulated annealing, will, however, be implemented so that larger instances of the problem can be solved.

References

[1] Amaldi E, Capone A, Malucelli F & Mannino C, 2006, *Optimization problems and models for planning cellular networks*, pp. 917–939 in Resende M & Pardalos PM (Eds), *Handbook of optimization in telecommunications*, Springer, New York (NY).

[2] Anonymous, 2014, *The Mobile Economy Sub-Saharan Africa 2014*, (Unpublished) Technical Report, Groupe Speciale Mobile Association.

[3] Hata M, 1980, *Empirical formula for propagation loss in land mobile radio services*, IEEE Transactions on Vehicular Technology, **29(3)**, pp. 317–325.

[4] Iskander MF & Yun Z, 2002, *Propagation prediction models for wireless communication systems*, IEEE Transactions on Microwave Theory and Techniques, **50(3)**, pp. 662–673.

[5] Krzanowski R & Raper J, 1999, *Hybrid genetic algorithm for transmitter location in wireless networks*, Computers, Environment and Urban Systems, **23(5)**, pp. 359–382.

[6] Mathar R & Niessen T, 2000, *Optimum positioning of base stations for cellular radio networks*, Wireless Networks, **6(6)**, pp. 421–428.

[7] McKinney AL & Agarwal KK, 1992, *Development of the Bresenham line algorithm for a first course in computer science*, The Journal of Computing Sciences in Colleges, **8**, pp. 70–81.

[8] Reed M, Jotishky N, Newman M, Mbongue T & Escofet G, 2013, *Africa Telecoms Outlook 2014 Maximising digital service opportunities*, (Unpublished) Technical Report, Informa.

[9] Sherali HD, Pendyala CM & Rappaport TS, 1996, *Optimal location of transmitters for micro-cellular radio communication system design*, IEEE Journal on Selected Areas in Communications, **14(4)**, pp. 662–673.

Proceedings of the 2015 ORSSA Annual Conference
pp. 102–111
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Toward decision support for firebase locations in Table Mountain National Park

T Meyer* R Reed† JH van Vuuren‡

Abstract

The Table Mountain National Park is situated in the Cape Peninsula within the South African Western Cape. The park comprises three disconnected natural areas, surrounded by the city of Cape Town and the Atlantic Ocean. Effective fire management is of the utmost importance within the park due to the facts that rare species of fynbos are found there and the City of Cape Town borders the park. There are currently three permanent firebases in the park, one in each section of the park and a fourth semi-mobile firebase which serves all three sections. Each permanent firebase is responsible for the section in which it is located. Since it is critical that fire-fighters respond to fires as quickly as possible, it is important that firebases are located in such a way that access to all areas of the park can be achieved in the shortest time possible. Three basic facility location models are used in this paper to determine the effectiveness of the current permanent firebase locations as well as to uncover acceptable trade-offs between minimising the number of firebases in the park and minimising the response times of fire fighting teams to fires breaking out in the park.

Key words: Firebase location, Facility location model, Table Mountain National Park.

1 Introduction

Table Mountain National Park (TMNP, also referred to in this paper as *the park*) is situated in the Cape Peninsula, and is surrounded on all sides be either the City of Cape Town or by the Atlantic Ocean [4, 10]. The park comprises three disconnected natural areas, as shown in Figure 1. The City of Cape Town and the TMNP are popular tourist destinations. The safety of visitors to the park and that of people living and visiting areas in close proximity to the park is therefore of the utmost importance.

*Department of Logistics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa
†Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: ryan4reed@gmail.com
‡**(Fellow of the Operations Research Society of South Africa)**, Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: vuuren@sun.ac.za

Fynbos, the dominant indigenous vegetation of the park, constitutes approximately 90% of the park's vegetation. It is an evergreen scrub-land vegetation characterised by reed-like Restios, fine-leafed Ericas and leather-leafed Proteas which is both fire-prone and fire-adapted [4], meaning that it requires a specific fire-regime[1] in order to maintain its diversity and ecosystem processes [10]. The TMNP is, however, greatly invaded by alien shrubs and trees which are not only spread by fire, but also threaten the biodiversity of the area, increasing biomass and hence the intensities of fires in the park. This seriously influences the recovery and survival of Fynbos and other natural vegetation after fires in the area [4].

It is therefore important that effective fire management takes place, and that the limited resources available for fire-fighting in the park are managed in the best manner possible. The park management is responsible for the difficult task of balancing the need to combine fire with other measures of alien plant control, maintaining features that provide safe boundaries, maintaining post-fire mosaic patterns and providing available manpower to control and contain fires so as to reduce the risk of damage to the urban fringe of the park.

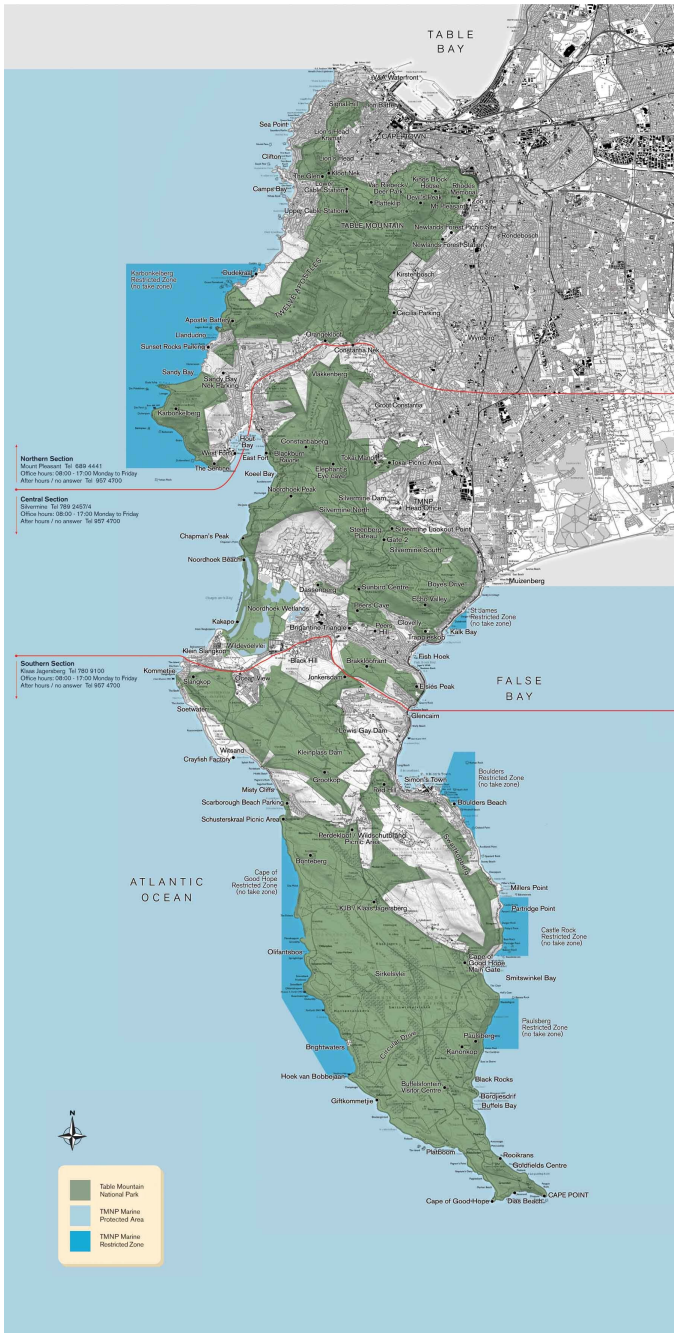


Figure 1: *Map of Table Mountain National Park.*

The TMNP *Fire Management Plan* [10] has two main objectives: to ensure the conservation and continued survival of viable populations of all indigenous biota in the area, and to minimise the potential and actual damage caused to property by fires [10]. Major factors exacerbating the problem of providing fire protection include poor planning and

[1] A fire regime describes a suitable frequency, size, season and intensity of fires for optimal re-population of biota.

unplanned wildfires as this sets back alien plant control operations and is a severe risk to the urban park fringe.

There are currently three permanent and independent firebases in the park; one firebase in each of the three (northern, central and southern) sections of the park, while a fourth semi-mobile firebase, situated in the northern section, is on standby for all three of the areas and plays an integrating role between these firebases [10]. These firebases are under extreme pressure to provide rapid response using the least amount of resources. Fire-management teams have to keep costs to a minimum, and the most important factor influencing the cost of fighting a fire is the time it takes for fire-fighters to respond to a fire; *response* being measured as the length of time elapsed between when a fire is reported and when the first line of fire-fighters arrive at the fire. The longer it takes for a team to reach a fire, the more out-of-control a fire becomes which, in turn, increases the cost of maintaining and extinguishing that fire. This is a serious concern in the northern and central sections of the park which border on private property within the City of Cape Town and thus have to be protected at all costs.

The locations of firebases in the park significantly affect the time it takes for fire-fighters to respond to a fire, and an optimal placement of these firebases can therefore greatly reduce the costs of fire management. For example, a series of devastating wildfires in the TMNP early in March 2015, resulting from a combination of particularly strong winds and high temperatures, led to the closure of various public roads between the three sections of the park. Traffic congestion on the roads that remained open severely delayed the response times of fire-fighting teams having to access the northern and southern sections of the park from firebases in the central section and elsewhere in the peninsula, giving rise to a realisation that each park section should be equipped with an adequate number of firebases which are strategically located. Following these devastating wildfires, the effectiveness of the current placement of firebases is being re-examined and alternative, potentially better firebase locations are sought in order to limit the damage caused in and around the park by such fires in the future and also to decrease the expected cost of fire management in the TMNP, if possible.

Our purpose in this paper is to take first steps in this respect, by attempting an answer to the following basic question: What is the smallest number of firebases required to cover each section of the TMNP, given a certain coverage response radius for each firebase. Although not full and sufficiently practical firebase location decision support for the park, an answer to this basic question may go some way towards helping the park management to form a reasonable idea of what size of area would have to be covered by each firebase, given a (budget) restriction on the number of firebases that can be placed in each section of the park.

The paper is organised as follows. Sample work from the literature on decision support within the context of firebase location is provided in §2, after which three popular facility location models are reviewed from the operations research literature in §3. These models are then applied to the TMNP in §4, after which we discuss the results obtained from these models. The paper closes in §5 with a number of ideas in respect of possible future work in terms of improving the realism of the three simple models employed here, so that the models can be used as the basis for practical firebase location decision support in future.

2 Selected literature on firebase location decision support

The location of fire-fighting facilities in various geographical areas has been extensively researched in the literature. We provide three examples of this kind of work in this section. Lui *et al.* [7] used a *Geographical Information System* (GIS) in conjunction with an ant colony algorithm to determine optimal locations of fire stations. The GIS was used to determine possible locations for fire stations and possible routes that may be followed from fire stations to accident sites. An ant colony algorithm was implemented to solve the problem of locating fire stations with the aim of serving as much as possible of a given jurisdiction area discretised into a grid of location cells, and to minimise the response times to these cells. This was applied to Singapore, and the primary objective was to cover the transportation routes of hazardous materials.

Schreuder [9] determined the smallest number of fire stations, the locations of these fire stations and the number of first attendance pumpers required by the city of Rotterdam so that each point in the city could be reached within a prescribed time. The city of Rotterdam was divided into a number of districts whose borders coincided with natural fire-breaks, such as rivers and wide roads. These districts were then classified according to different levels of risk. A network approach was adopted to determine a set of possible locations for fire stations and a set-covering model was implemented, using the locations obtained from the network approach, to select the smallest number of fire stations which could fulfil the first attendance requirements.

Chow and Regan [2] compared a static *k-server p-median problem* to a *chance-constrained dynamic k-server relocation problem* for fighting regional wild-land fires. Their objective was to solve the problem of locating and relocating air tanker initial attacks. They showed that a dynamic resource location model using a rolling horizon forecast of future conditions is able to obtain more cost-effective results than a static model.

3 Three popular facility location models

All of the work cited in §2 was based on generalisations of three basic facility location models in the operations research literature, namely the *Set Covering Location Model* (SCLM), the *Maximum Coverage Location Model* (MCLM) and the *Uncapacitated Facility Location Model* (UFLM) [3]. These three models are reviewed briefly in this section.

3.1 The Set Covering Location Model

The objective in the well-known SCLM is to minimise the number of facilities required to cover demand generated at a set of demand points. The demand generated at a demand point is generally considered to be met by a facility if the demand point is close enough to (or in the region of service or influence) of the facility. Denote the set of demand points by $\mathcal{I} = \{1, 2, \ldots, |\mathcal{I}|\}$ and the set of facility candidate locations by $\mathcal{J} = \{1, 2, \ldots, |\mathcal{J}|\}$. Furthermore, denote the distance between demand point $i \in \mathcal{I}$ and candidate site $j \in \mathcal{J}$ by d_{ij} and let D denote the coverage distance radius of a facility. Let $\mathcal{N}_i = \{j \mid d_{ij} \leq D\}$ be the set of all candidate facility locations that cover demand point $i \in \mathcal{I}$ and define the

decision variables

$$x_j = \begin{cases} 1 & \text{if a facility is located at candidate site } j \\ 0 & \text{otherwise.} \end{cases}$$

Then the objective in the SCLM is to

$$\text{minimise} \quad Z = \sum_{j \in \mathcal{J}} x_j \tag{1}$$

subject to the constraints

$$\sum_{j \in \mathcal{N}_i} x_j \geq 1, \qquad i \in \mathcal{I}, \tag{2}$$

$$x_j \in \{0, 1\}, \quad j \in \mathcal{J}. \tag{3}$$

Constraint set (2) ensures that each demand point is covered by at least one facility, while constraint set (3) specifies the binary nature of the decision variables.

3.2 The Maximum Coverage Location Model

The SCLM may, however, require an unpractically large number of facilities to be placed in order to ensure that *every* demand point is covered by a facility. In the MCLM, the objective is instead to locate at most a predetermined number of facilities, p (say), in order to maximise their combined coverage. Upon defining the additional auxiliary variables

$$y_i = \begin{cases} 1 & \text{if demand point } i \text{ is covered by at least one facility} \\ 0 & \text{otherwise,} \end{cases}$$

the objective in the MCLM is to

$$\text{maximise} \quad Z' = \sum_{i \in \mathcal{I}} h_i y_i \tag{4}$$

subject to the constraints

$$\sum_{j \in \mathcal{N}_i} x_j \geq y_i, \qquad i \in \mathcal{I}, \tag{5}$$

$$\sum_{j \in \mathcal{J}} x_j \leq p, \tag{6}$$

$$y_i, x_j \in \{0, 1\}, \quad i \in \mathcal{I}, \ \ j \in \mathcal{J}. \tag{7}$$

The parameter h_i in (4) represents an importance rating associated with covering demand point $i \in \mathcal{I}$. Constraint set (5) ensures that demand point i is only considered covered if there is at least one facility in its neighbourhood $\mathcal{N}_i$, while constraint (6) ensures that at most p facilities are placed.

3.3 The Uncapacitated Facility Location Model

In the UFLM, each demand point $i \in \mathcal{I}$ must be assigned to exactly one, $j \in \mathcal{J}$ (say), of p facilities, incurring a cost c_{ij} in the process which is typically proportional to the distance d_{ij} between the points i and j. Upon defining the alternative auxiliary variables

$$z_{ij} = \begin{cases} 1 & \text{if demand point } i \text{ is assigned to a facility at site } j \\ 0 & \text{otherwise,} \end{cases}$$

the objective in the UFLM is to

$$\text{minimise} \quad Z'' = \sum_{j \in \mathcal{J}} f_j x_j + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{ij} z_{ij} \tag{8}$$

subject to the constraints

$$\sum_{j \in \mathcal{J}} z_{ij} = 1, \qquad i \in \mathcal{I}, \tag{9}$$

$$\sum_{j \in \mathcal{J}} x_j = p, \tag{10}$$

$$z_{ij} \leq x_j, \qquad i \in \mathcal{I}, \ j \in \mathcal{J} \tag{11}$$

$$z_{ij}, x_j \in \{0, 1\}, \qquad i \in \mathcal{I}, \ j \in \mathcal{J}. \tag{12}$$

The parameter f_j in (8) represents the fixed cost associated with locating a facility at site $j \in \mathcal{J}$. Constraint set (9) ensures that a demand point is covered by exactly one facility, while constraint (10) ensures that exactly p facilities are placed. Constraint set (11) performs a linking function ensuring that a demand point $i \in \mathcal{I}$ is only assigned to a facility at candidate site $j \in \mathcal{J}$ if, in fact, there is a facility at that site.

4 Analysis

We discretised the area of the TMNP into a grid of square cells (each approximately 650m$\times$ 650m in size), as shown in Figure 2. This resulted in 162, 161 and 371 candidate locations for firebases in the northern, central and southern sections of the park, respectively. The locations of the current three permanent firebases in the park are also indicated in Figure 2. Our first step was to evaluate the effectiveness of these existing firebases within each section of the park according to the five response radii shown in Figure 3. Measuring this effectiveness as the ratio of the number of cells within these neighbourhoods of each firebase to the total number of cells in the corresponding section of the park, we found the effectiveness values in Table 1(A) during a sensitivity analysis with respect to increasing firebase response radii. The small effectiveness percentages in Table 1(A) achieved by the existing firebases in the northern and central sections of the park may be attributed to the fact that these bases are situated on the park boundary (so as to provide easy access by staff to the bases from the urban fringe), while for the southern section one firebase is ineffective due to the sheer size of that section of the park.

The natural question arising from the effectiveness results of Table 1(A) is what the maximum effectiveness value is that can possibly be achieved according to the response radii of Figure 3 by placing p firebases in each each section of the park. Using the software suite Lingo 11 [6] to solve the MCLM for $p = 1$ and the response radius in Figure 3(d), for example, the answer to this question for the northern [central, resp.] section of the park is 57/162 = 35.2% [51/161=31.7%], which is much better than the effectiveness value of 17.3% [21.1%, resp.] achieved by the current firebase locations. For $p = 5$ and the same response radius, however, the answer to this question increases to 161/162 = 99.4% [157/161=97.5%] for the northern [central, resp.] section of the park. According to the MCLM, optimal locations for sets of one, three and five firebases in the northern section of the park corresponding to the response radius in Figure 3(e) are, for example, shown in Figure 4.

Another natural question is to seek the smallest number of firebases required to achieve 100% coverage effectiveness in each section of the park. Solving the SCLM for the five firebase response radii of Figure 3 revealed that the smallest numbers of firebases required to cover each section of the park entirely is as shown in Table 1(B). According to the SCLM, optimal firebase locations in the northern section of the park corresponding to the response radii in Figure 3(c)–(e) are, for example, shown in Figure 5.

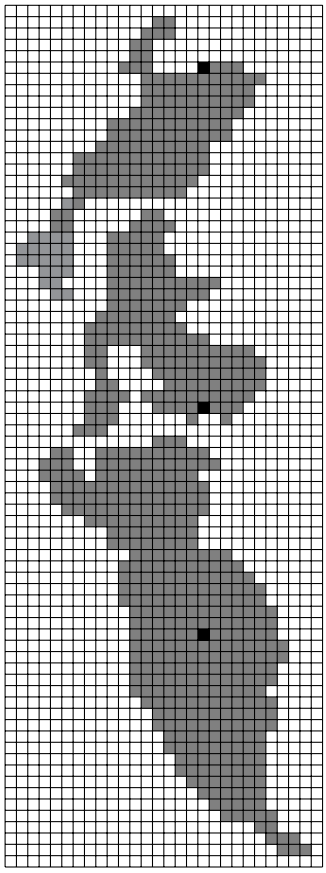


Figure 2: *Discretisation of the TMNP into 694 square cells. The locations of the current permanent firebases in the TMNP are indicated.*

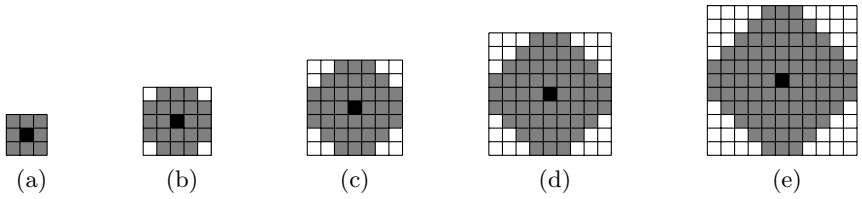
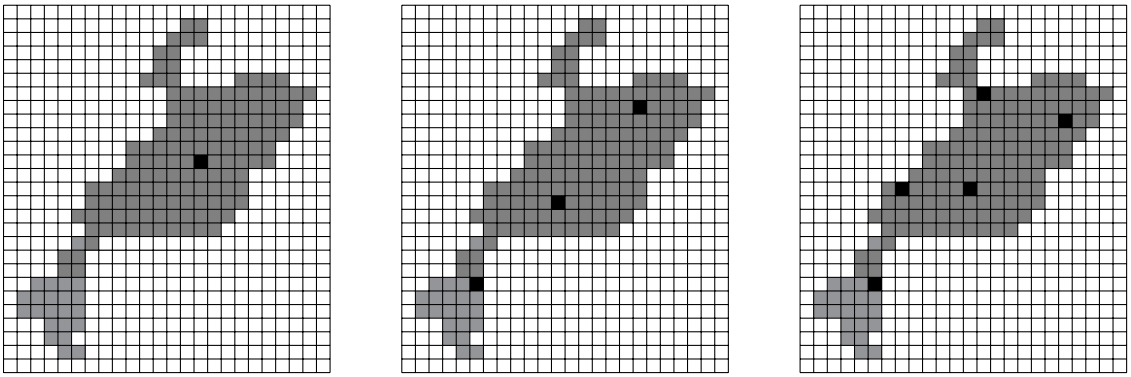


Figure 3: *Five neighbourhoods of a firebase (shown in black) induced by coverage radii of different sizes (shown in grey). It is assumed that the fire-fighting staff and equipment located at a firebase can sufficiently reach and quickly service any shaded demand cell.*

Since the UFLM (8)–(12) is deemed the most realistic of the basic facility location models described in §3, it makes sense to base preliminary decision support in terms of recommended firebase locations within the park on this model. The cost c_{ij} of servicing demand point i from firebase j in the UFLM was taken as the direct distance from i to j (measured as the Euclidean distance between the corresponding cells, in units of cells). According

	(A): Current firebase coverage % with coverage radius in Figure 3					(B): Min no of firebases required for full park coverage with radius in Figure 3				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
Northern section	3.1%	6.8%	11.7%	17.3%	25.9%	24	13	8	6	5
Central section	4.3%	8.7%	14.9%	21.1%	26.7%	25	13	8	6	5
Southern section	2.4%	5.7%	9.9%	15.4%	21.8%	50	25	15	11	8

Table 1: *(A) Effectiveness (coverage percentage) of existing permanent firebases in the TMNP and (B) the number of facilities in optimal solutions to the SCLM for the various sections of the TMNP, according to the facility coverage neighbourhoods of Figure 3.*



(a) (b) (c)

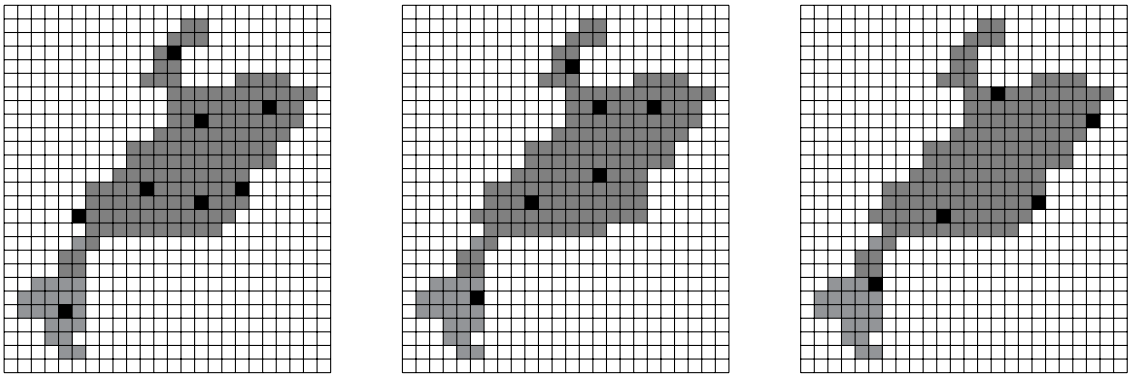
Figure 4: *Placement of (a) one, (b) three, and (c) five firebases in the northern section of the TMNP according to the MCLM with firebase response radius as in Figure 3(e).*

to the park management, the annual fire standby cost in the TMNP is approximately 200% of the costs incurred by actual fire-fighting activities, prescribed burning and aerial assistance [10]. The firebase location cost f_j at site j was derived from this figure. Since the average distance between all demand points in our discretisation of the TMNP is 10.6 cell units, a fixed placement cost $f_j = 21$ (approximately 200% of 10.6 units) was therefore assigned to all firebase candidate sites j. Because it would be costly to relocate the current three (permanent) bases, the UFLM was solved for $p = 4$ firebases in each section of the park, but with the restriction that the current firebase locations had to be respected (*i.e.* kept). The resulting recommended firebase locations in the northern and central sections of the park are shown in Figure 6.

5 Further work

The work reported in this paper emanated from a pilot project forming part of a larger, ongoing research project at Stellenbosch University. Future work will centre on improving four aspects related to the level of realism in the facility location models of §3.

The approximately circular firebase response radii in Figure 3 are currently measured in units of vertically projected distance, but should ideally rather be measured in units of



(a) (b) (c)

Figure 5: *Placement of firebases in the northern section of the TMNP according to the SCLM with firebase response radii as (a) in Figure 3(c), (b) in Figure 3(d), and (c) in Figure 3(e).*

time (*e.g.* a 10-minute response radius, a 20-minute response radius, and so on). Terrain features (such as ravines or cliffs), vegetation type and access road infrastructure play a significant role in the distance a fire-fighting team can cover within a fixed time frame. A fixed-time coverage area of a firebase is therefore expected to be non-circular. We plan to consult both the literature and fire-fighting experts to gain insight into how a fixed-time coverage area can be determined for a firebase and hence be incorporated into the facility location modelling approach.

Furthermore, in the results reported in §4, we took $h_j = 1$ for all $i = 1, \ldots, n$ in the MCLM objective (4). It is, however, envisaged that the possibility of allocating coverage importance ratings to certain cells in the park discretisation may be of importance to the park management. It may, for example, be more important to be able to respond rapidly to a fire close to the urban fringe of the park than to a fire in the park interior. We would like to engage with the park management in order to assign a suitable importance weighting h_j to each cell j in the park, and then re-solve the MCLM. In a similar vein, the fixed costs f_j and variable costs c_{ij} in the UFLM objective (8) were measured in terms of TMNP discretisation cell units in this paper. This was done merely for the sake of simplicity and is not considered a practical modelling approach. We would also like to engage with the park management in an attempt to elicit realistic cost values for these coefficients, and then re-solve the UFLM.

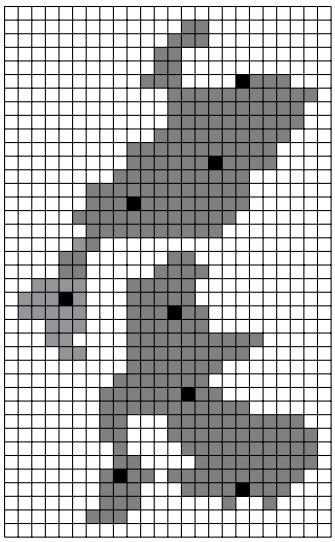


Figure 6: *Placement of firebases in the northern and central parts of the TMNP according to the UFLM with* $p = 4$.

Moreover, the performance measure of a firebase location was measured in this paper in terms of its ability to reach the potential fire ignition points within a specified coverage area, without taking into account the extent to which the fire may have spread before

a fire-fighting team reaches it. In reality, however, fire-fighting teams take action based on the potential threat of a wildfire — not on the current size or ignition point of the fire *per se* [10]. This shortcoming may be rectified by solving a facility location model in conjunction with a fire spread model. An example of such a fire spread model is the cellular automaton model of Berjak and Hearne [1]. There are, however, also analytic fire spread models which may be employed for this purpose. Frandsen [5], for example, showed that the quasi-steady rate of fire spread in the x-direction along a horizontal plane, and in the absence of wind, is given by

$$R = \frac{1}{\rho Q}\left(F_x + \int_{-\infty}^{0} \left.\frac{\partial I_x}{\partial y}\right|_{y=c} \mathrm{d}x\right),$$

where x and y are horizontal and vertical Cartesian coordinates, respectively, ρ is the effective bulk density (the amount of fuel per unit volume of the fuel bed raised to ignition ahead of the advancing fire), Q is the heat required to bring a unit weight of fuel to ignition, F_x is the horizontal heat flux absorbed by a unit volume of fuel at the time of ignition and $\partial I_x/\partial y$ is the vertical gradient of the fire intensity evaluated in a horizontal plane. Rothermel [8] generalised this result to hold for the rate of spread of a fire advancing along an incline. These spread rates will, of course, be influenced by the presence of wind.

Finally, the modelling approach adopted in this paper was deterministic in nature. A stochastic modelling approach may also be followed where fire ignition points are forecast, based on historical fire data, and firebase locations are planned according to likely fire ignition points instead of attempting to be able to cover *all* demand rapidly.

References

[1] Berjak SG & Hearne JW, 2002, *Spatial fire modeling in Mkuze game reserve: A case study*, ORiON, **18(2)**, pp. 37–57.

[2] Chow JYJ & Regan AC, 2011, *Resource location and relocation models with rolling horizon forecasting for wildland fire planning*, INFOR, **49**, pp. 31–43.

[3] Daskin MS, 2013, *Network and discrete location: Models, algorithms and applications*, Second edition, Wiley, Hoboken (NJ).

[4] Forsyth GG & Van Wilgen BW, 2008, *The recent fire history of the Table Mountain National Park and implications for fire management*, Unpublished Report, Table Mountain National Park, Cape Town.

[5] Frandsen WH, 1971, *Fire spread through porous fuels from the conservation of energy*, Combustion and Flame, **16**, pp. 9–16.

[6] Lindo Systems Incorporated, 2012, *Lingo 11.0*, [Online], [Cited September 30th,2012], Available from `http://www.lindo.com/`

[7] Liu N, Huang B & Chandramouli M, 2006, *Optimal siting of fire stations using GIS and ANT algorithm*, Journal of Computing in Civil Engineering, **20**, pp. 361–369.

[8] Rothermel RC, 1972, *A mathematical model for predicting fire spread in wildland fuels*, Research Paper INT-115, Intermountain Forest and Range Experiment Station, Ogden (UT).

[9] Schreuder JAM, 1980, *Application of a location model to fire stations in Rotterdam*, European Journal of Operational Research, **6**, pp. 212–219.

[10] Table Mountain National Park, 2004, *Fire management plan*, Unpublished Report ENV-S-C 2004-043.

Proceedings of the 2015 ORSSA Annual Conference
pp. 112–122
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Tri-objective generator maintenance scheduling for a national power utility

BG Lindner* J Eygelaar* DP Lötter* JH van Vuuren†

Abstract

Providing reliable energy 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 power system. The well-known *generator maintenance scheduling* (GMS) problem is the problem of finding a schedule for the planned maintenance outages of generating units in a power system. A novel tri-objective model formulation is proposed for the GMS problem in this paper. The first (and most commonly adopted) objective involves minimising the squared reserve levels, which serves to create an even ("reliable") margin of generating capacity over and above expected demand. The second objective involves the production cost associated with a maintenance plan for the generating units in a system, where planning maintenance of a power generating unit which is cheap to operate during a high demand period will incur a higher production cost. The third objective involves minimising the risk (on expectation) of generating units breaking down, where the longer the time period since the last maintenance service of a generating unit, the larger the risk of it breaking down.

Key words: Energy sector, Maintenance scheduling, Multiple objective optimisation, Reliability.

1 Introduction

Power outages in South Africa are mainly brought about by higher than expected demand, infrastructure failure, and a diminishing reserve capacity (available capacity over and above demand). The reserve margin for generating capacity has decreased in recent years from the desired 15% to less than 8% [11]. As a result, South African power stations have recently been forced to operate virtually continuously at high load factors (how near to maximum a plant is operating on a percentage basis). In addition, the generating units

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, emails: 15150526@sun.ac.za, 16516885@sun.ac.za & danielotter@sun.ac.za

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

of South African power stations are relatively old, which means that they require above-average levels of maintenance. These two aspects contribute significantly to the prevalence of unplanned outages [24]. Appropriate preventative maintenance planning is crucial to mitigate the above risks and is one of the key focus areas for a power utility [3, 5, 13, 18] — especially for a power utility such as the South African utility, which has been postponing maintenance plans on its already ageing stations [6].

A schedule for the planned maintenance outages of generating units in a power system is sought in the well-known *generator maintenance scheduling* (GMS) problem [22] and we propose a novel tri-objective model formulation for the GMS problem in this paper as well as how solutions to this problem may be included in a national power utility's decision support software.

The paper is organised as follows. The working of typical energy flow simulators is described in §2, after which a brief survey is conducted in §3 of existing single and multi-objective model formulations of the GMS problem. This is then followed by a methodology section (§4) in which our novel tri-objective approach towards formulating the GMS problem is described. Preliminary GMS results are presented in §5, and this is followed by a discussion on possible future work in §6.

2 Energy flow simulators

Power utilities often use decision support software tools in the form of energy flow simulators in which the entire energy supply chain is modelled "from fuel to fridge." The working of and various constituent components of such a software tool are elucidated in Figure 1. An energy flow simulator is typically designed to function as a what-if analysis tool in the context of possible different future scenarios. This allows for the accommodation of different *Energy Availability Factors* (EAFs) per power station, different weather patterns, a variety of *Gross Domestic Product* (GDP) levels, varying supply levels and qualities, *etc.* of generation fuel. The main simulation technique employed in such an energy flow simulator is typically Monte-Carlo simulation [8, 9, 16].

A simulation is typically initiated by a *consumption module* (Figure 1(a)) which forecasts the total energy demand per geographic region and customer type (residential, manufacturing, mining, *etc.*) according to some estimated level of GDP (high, medium or low) and weather scenario (hot, normal or cold). The demand thus forecast may then be used by a *production planning module* (Figure 1(b)) to schedule the planned energy production per power station (including coal, nuclear, gas-turbine, hydro-electric and renewable energy units) so as to minimise production cost. Demand must be met whilst taking into account production capacity. A *primary energy module* (Figure 1(c)) usually facilitates what-if analyses in terms of a variety of different plans and scenarios, including unplanned power station maintenance, and variation in the quality and quantity of generation fuel. The final main component is a *generation module* (Figure 1(d)). The production plan, supply reliability, and the quality and quantity of generation fuel are fed into a *generation module*, which then quantifies emissions, such as sulphur and nitrogen oxides. System losses are usually also incorporated into a typical energy flow simulator [9].

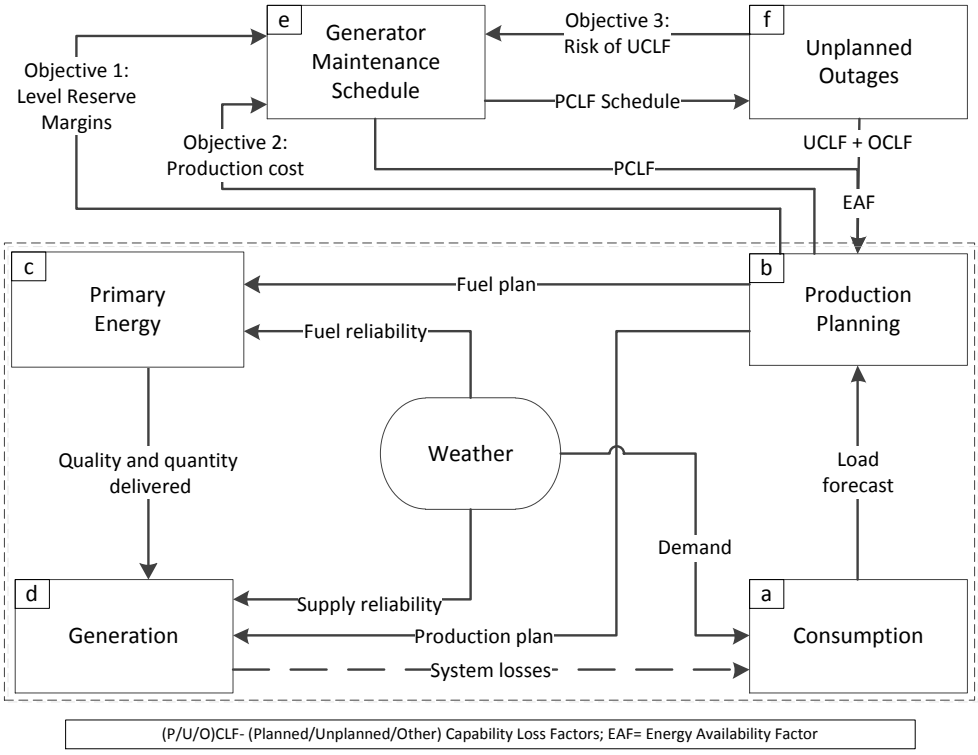


Figure 1: *High-level representation of a typical energy flow simulator (dashed area) and how it is anticipated that the proposed GMS modelling approach may form part of it (adapted from [9]).*

Energy flow simulators usually have no optimisation capacity (other than possibly in their energy production planning components (Figure 1 (b)). Many variable and parameter values employed within such a simulator are typically known to be sub-optimal, and hence there is a need to be able to optimise decision variables within such simulators [8]. The GMS modelling approach proposed in this paper has been designed specifically to allow for its incorporation into a typical energy flow simulator's decision support software framework (see Figure 1(e)).

3 Literature review

The GMS problem is well-known in the operations research literature. The most prevalent solution methods applied to solve instances of the GMS problem include heuristic rules, mathematical programming techniques, dynamic programming, expert systems, fuzzy systems, and metaheuristics [1, 24]. Three dominant criteria are usually incorporated in formulations of the GMS problem, namely economic criteria, reliability criteria and convenience criteria [13, 25].

The most commonly adopted economic criterion consists of minimising the total operating cost associated with a generator maintenance schedule, including both energy production cost and maintenance cost [3]. The most popular reliability-related objectives, on the

other hand, are to minimise the expected lack of peak net reserve, to minimise the expected energy not supplied and to minimise the loss of load probability [3]. Examples of convenience-related objectives include minimising soft constraint violations or minimising possible disruptions to the power generation schedule [24]. These three categories of objectives are conflicting, ultimately making the GMS problem multi-objective in nature. Both single and multi-objective formulations have, however, been proposed for the GMS problem in the literature [25].

3.1 Single-objective GMS formulations and solution methodologies

In single-objective GMS problem formulations the two dominant objectives usually involve economic or reliability criteria, with some authors including other objectives as constraints [13, 25].

Single-objective GMS formulations incorporating economic criteria (operating cost of some composition) are widespread [24], with Canto [2] adopting a 0/1 mixed integer linear programming approach and applying Bender's decomposition, Edwin & Curtius [7] following an integer linear programming model approach, and Mromlinski [19] preferring an integer programming model formulation and solving it with the branch-and-bound method.

Reliability criteria may either be modelled in a deterministic or stochastic fashion [18]. The most commonly adopted reliability-related objective is levelling the reserve load over all the time periods, which is generally achieved by minimising the sum of squares of the reserve [18, 26]. This approach has successfully been followed in [3, 4, 18], for example. In these studies, metaheuristics were employed as approximate solution techniques (typically a genetic algorithm, simulated annealing or a hybrid of the two). An alternative reliability-related objective is to maximise the minimum reserve during any time period.

We are not aware of any work in the literature in which single-objective formulations of the GMS problem involve only convenience criteria. The multi-objective formulations in [12, 14, 15], however, include it as an optimality criterion within a multi-objective modelling paradigm, as described in the following section.

3.2 Multi-objective GMS formulations and solution methodologies

Huang *et al.* [10] used fuzzy dynamic programming to solve instances of a bi-objective GMS model formulation in which the objectives were to level the reserve margins (a reliability criterion) and to minimise the lost opportunity production cost of generating units undergoing maintenance (an economic criterion). Goal programming was used by Munoz & Ramos [20] to solve instances of another bi-objective GMS model formulation involving thermal generating units under both economic and reliability criteria.

Leou [15] combined a genetic algorithm with the method of simulated annealing to solve a GMS model with objectives including convenience criteria (minimisation of reliability and cost constraint violations) and economic criteria (cost of operation and maintenance).

Krajl [12, 14] used the multi-objective branch-and-bound algorithm to solve instances of a tri-objective GMS model formulation including an economic objective (minimisation of

fuel costs), a reliability-related objective (minimisation of the expected unserved energy over time) and convenience-related objectives (minimisation of constraint violations).

4 Proposed modelling approach

As we previously suggested in [16], a multi-objective modelling approach is required in the context of GMS because of the inherent trade-offs between the conflicting scheduling objectives. The three objectives we propose for inclusion in a GMS formulation are two common objectives found in the literature on GMS, namely levelling reserve margins and minimising production cost, together with the novel objective of minimising the risk of generating unit failure.

A maintenance schedule is defined as follows. Suppose there are n generating units in the power system and m time periods during the planning horizon. Let $\mathcal{I} = \{1, \ldots, n\}$ be the set of generating units and let $\mathcal{J} = \{1, \ldots, m\}$ be the set of time periods. Finally, define the binary decision variable x_{ij} to take the value 1 if maintenance of generating unit $i \in \mathcal{I}$ commences during time period $j \in \mathcal{J}$, or zero otherwise. Then a maintenance schedule is an assignment of zeros and ones to the $n \times m$ matrix $\boldsymbol{X} = [x_{ij}]$ of decision variables satisfying a variety of constraints, including maintenance window constraints, load constraints, resource constraints, and exclusion constraints amongst others [16].

We propose the use of simulated annealing for computing high-quality maintenance schedules as the method has been adopted successfully a number of times in the GMS literature [3, 22, 23, 25]. The simulated annealing hybrid developed by Schlünz and van Vuuren [25] has outperformed a genetic algorithm and genetic algorithm/simulated annealing hybrid, and has matched the best known solution found via ant colony optimisation in the context of documented case studies [25]. An innovative multi-objective simulated annealing algorithm, developed by Smith *et al.* [27, 28] may be used to find a maintenance schedule, as described above, which achieves acceptable trade-offs between the three objectives proposed here. In this approach, the conventional "energy" difference between a current and neighbouring candidate solution in standard single-objective simulated annealing algorithms is replaced by a measure of the difference in dominance (in terms of an archived non-dominated front).

4.1 Objective 1: Levelling of reserve margins

We endorse the standard objective of levelling the reserve energy over and above demand over all time periods by minimising the sum of squared reserve margins over time [25]. The reserve margin during a particular time period is the difference between the available capacity and the expected demand (see Figure 2(a)). Minimising this sum of squared reserve margins results in an even ("reliable") band of reserve margins. Let r_j represent the reserve margin during time period j. Then this objective involves minimising $\sum_{j=1}^{m} r_j^2$.

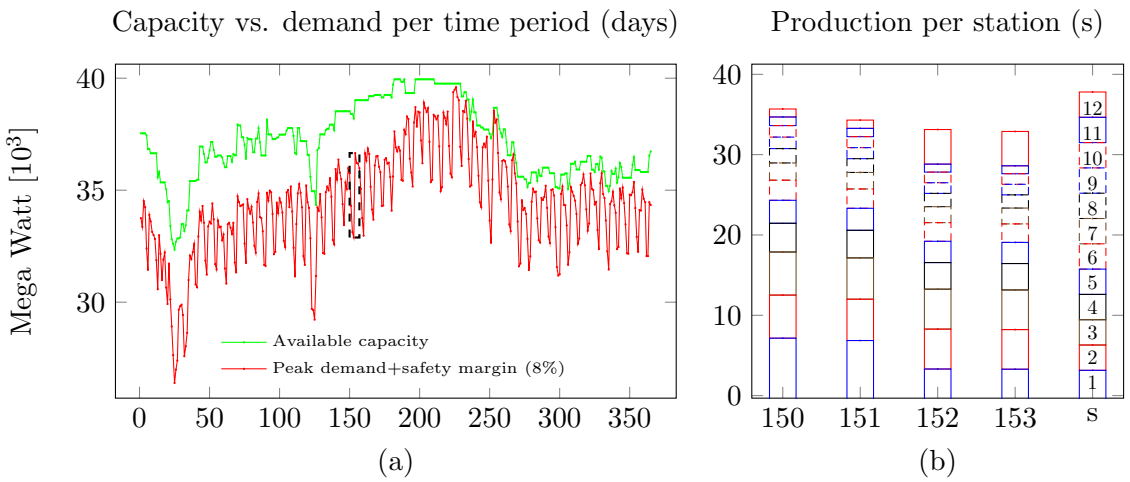


Figure 2: *(a) Available capacity versus peak demand (including an adequate safety margin) from a hypothetical South African case study [26]. The difference between these two is the reserve level (used in the first objective). (b) Output of a typical energy flow simulator's production planning module (used in the second objective). 1-Cheapest power station, 12-Most expensive power station (measured in R/MWh).*

4.2 Objective 2: Minimising production cost

As mentioned, the production planning module of an energy flow simulator (see Figure 1(b)) typically schedules the planned energy production by making use of available power generating units (including coal, nuclear, gas-turbine, hydro-electric and renewable generating units) with a view to minimise production cost. Power stations associated with cheaper production costs are typically scheduled first until demand is met, whilst taking into account production capacities of the various power stations [9].

The production cost minimisation may be achieved using a linear programming model. The decision variables in such a linear programming model should represent the amount of energy production planned per power station (MW). Important model parameters should include the associated energy production rate (measured in R/MWh) and the *Energy Availability Factor* (EAF) of power station s during time period j, denoted by

$$\text{EAF}_{sj} = 1 - (\text{PCLF}_{sj} + \text{UCLF}_{sj} + \text{OCLF}_{sj}), \tag{1}$$

where PCLF_{sj} refers to power generation losses specifically planned by the management of a power utility for maintenance purposes and other planned shutdowns at power station s during time period j. Furthermore, UCLF_{sj} refers to breakdowns (often as a result of a lack of planned maintenance) at power station s during time period j, while OCLF_{sj} refers to other losses due to extraordinary events outside the control of the management of the power utility at power station s during time period j, such as employee strikes or theft of transmission cables [8, 17].

To illustrate how a maintenance schedule may affect production cost, an example of a production plan is shown in Figure 2(b). If power station 1 (cheapest) has to undergo maintenance during days 152–153, this will increase its PCLF (as illustrated in Table 1), which will, in turn, decrease the station's EAF in (1). This latter value is an input

parameter to the production planning module (in the form of capacity constraints in the linear programming model), translating into less energy production being scheduled for the station (although it is the cheapest power station), which will increase the overall production cost. The GMS solution approach should attempt (if possible) to ensure that maintenance of cost-efficient power stations does not occur during high energy demand periods.

Table 1: *An example of how a maintenance schedule may affect a station's Planned Capability Loss Factor (PCLF) during time period j.*

Power Station	Energy Production Rate (R/MWh)	Generating Unit	Maintenance Schedule	PCLF_{sj}
$s = 1$	110	$i = 1$ $i = 2$ $i = 3$	$x_{1j} = 1$ $x_{2j} = 0$ $x_{3j} = 0$	33.33%
$s = 2$	150	$i = 4$ $i = 5$	$x_{4j} = 0$ $x_{5j} = 1$	50%
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s = 12$	2300	$i = 54$ $i = 55$	$x_{54j} = 0$ $x_{55j} = 0$	0%

4.3 Objective 3: Minimising risk of unit breakdown

The third objective proposed for inclusion in our GMS model formulation is a risk function. The objective quantifies the risk of generating unit breakdown associated with a specific maintenance schedule. Based on the schedule there will be an ever-increasing risk of generating unit breakdown per time period over which the unit has been operating continually without maintenance. The process of estimating the expected risk of breakdown of a power generating unit over time will involve obtaining and cleaning historical failure time series data for the unit. A trend test, such as the Laplace trend test, may then be performed on the data [21] in order to classify the power generating unit as either a repairable or a non-repairable system. After the units have all been classified in this manner, an appropriate lifetime distribution model may be chosen (such as an exponential or a Weibull model for non-repairable systems, or a *homogeneous poisson process* (HPP) or a *non-homogeneous poisson process* (NHPP) following an exponential or power law for repairable systems), based on the nature of the failure data of the power generating unit [29]. Next, the parameters for each model selected for each unit may be estimated by either the method of least squares or the maximum likelihood method.

It is advocated that the risk be weighted according to the importance of the power generating unit to the network as a whole, because some units may be more essential in respect of grid integrity, such as those that have the highest rated capacity. The risk of generating unit failure may, for example, be weighted according to the rated capacity of the unit.

In addition to the above objective, it is also proposed that the expected failure data of a unit should be used to constrain its maintenance window in the GMS problem. An earliest and a latest starting time are usually specified as input parameters (in the form

of maintenance window constraints) in traditional formulations of the GMS problem. We advocate that the decision maker should instead specify some maximum tolerable risk of generating unit failure and that this input parameter should rather be incorporated into the GMS formulation. Using these parameter estimates as described above, a time instant can be estimated at which the next failure of each unit is expected to occur. This time may then be used to identify a date before which the specific unit should be scheduled for planned maintenance so as to avoid UCLFs as far as possible.

5 Preliminary results

In 2012, Schlünz and van Vuuren [26] solved a hypothetical South African case study instance of the GMS problem with the single objective of levelling the reserve energy margins using the method of simulated annealing. The data included in their GMS problem instance do not represent the exact South African generation system due to confidentiality concerns, but the case study nevertheless represents a realistic GMS scenario. Constraints in the scenario were restricted to the adherence to maintenance windows, the system meeting the load demand together with a safety margin (Figure 2(a)), and respecting simultaneous generating unit maintenance exclusion constraints. The case study consists of a GMS problem instance containing 157 generating units requiring maintenance over a 365-day planning horizon. These dimensions are considerably larger than other test systems in the literature.

We used the same data set, but instead solved a bi-objective version of the GMS problem in which two of the three objectives proposed in §4 were pursued: minimising the sum of square levels of reserve margins and minimising the energy production cost associated with a maintenance schedule. Each maintenance schedule has an associated available capacity of the entire system of power stations (Figure 2(a)), which may be used to calculate the reserve margins for the first objective (the horizontal axis in Figure 3). In addition, the maintenance schedule per power generating unit may be translated into the PCLF for each station (as illustrated in Table 1) and may then be used by the linear programming model within the production planning module of the energy flow simulator (see Figure 1(c)) to construct an energy production plan and its associated cost (the vertical axis in Figure 3). These two outputs are the objectives we minimised. As may be seen in Figure 3, the bi-objective simulated annealing algorithm of Smith *et al.* [27, 28] converges towards the final non-dominated front over the course of 1 000 iterations.

6 Conclusion and future work

A novel tri-objective GMS modelling approach was proposed in this paper, which may easily be incorporated into a national power utility's decision support software. Preliminary results found by solving a problem instance including two of the proposed three objectives seem to indicate that the modelling approach and proposed solution methodology are capable of producing a sensible non-dominated front of solutions in objective space.

The work reported in this paper forms part of a larger, ongoing research project at Stel-

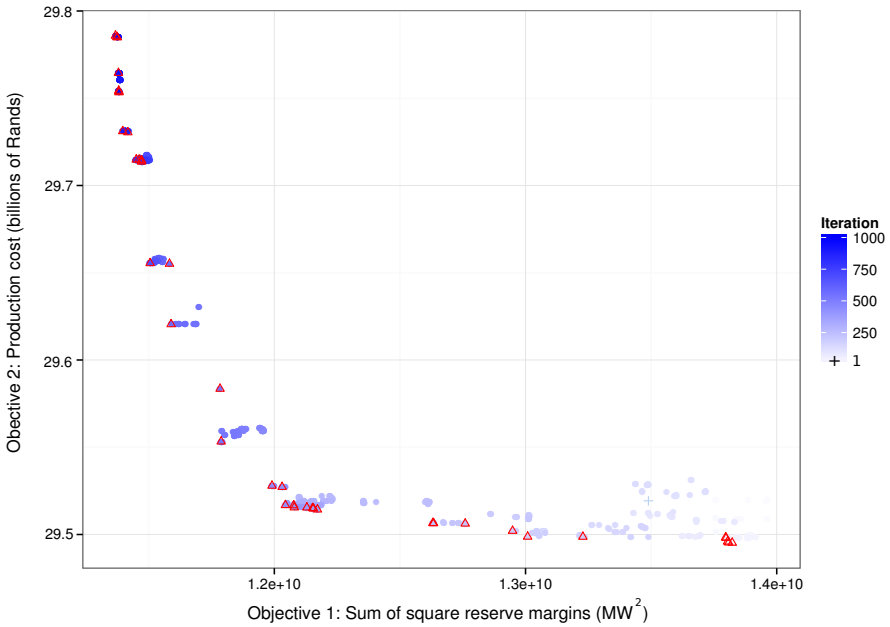


Figure 3: *Optimisation results for two (of the three) GMS objectives proposed in this paper, namely the minimisation of the sum of squared reserve margins and the minimisation of energy production cost.* ∘ — *Dominated solutions,* △ — *final non-dominated solutions.*

lenbosch University aimed at providing support for the complex planning decisions of a power utility. The next step will be to conduct further experiments (by varying parameters and cooling schedules) of the multi-objective simulated annealing algorithm employed, in addition to collecting further data so as to be able to construct further real-life instances of the GMS problem for testing purposes. Also, the authors will begin to incorporate the third objective proposed. It is envisaged that the results of this GMS modelling approach may be incorporated into the energy flow simulation framework of Figure 1 as a GMS decision support software tool to be used by the managers of a national power utility.

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] Canto SP, 2008, *Application of Benders' decomposition to power plant preventive maintenance scheduling*, European Journal of Operational Research, **184(2)**, pp. 759–777.

[3] 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.

[4] Dahal K, Aldridge C & McDonald J, 1999, *Generator maintenance scheduling using a genetic algorithm with a fuzzy evaluation function*, Fuzzy Sets and Systems, **102**, pp. 21–29.

[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] Department of Energy, 2011, *Integrated resource plan for electricity 2010–2030*, [Goverment Report], Goverment Publications, Pretoria.

[7] Edwin K & Curtius F, 1990, *New maintenance-scheduling method with production cost minimization via integer linear programming*, International Journal of Electrical Power & Energy Systems, **1(3)**, pp. 2–7.

[8] Hatton M, 2015, *Requirements specification for the optimisation function of an electric utility's energy flow simulator*, MEng Thesis, Stellenbosch University, Stellenbosch.

[9] Hatton M & Bekker J, 2014, *Development of an optimiser for a simulator of an electric utility: Challenges and approach*, Proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa, pp. 18–26.

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

[11] Kolb J, 2009, *Eskom 2000–2008: Our recent past*, [Online], [Cited May 15th, 2015], Available from `http://heritage.eskom.co.za/eskom_2000.htm`.

[12] Kralj B & Rajaković N, 1994, *Multiobjective programming in power system optimization: New approach to generator maintenance scheduling*, International Journal of Electrical Power & Energy Systems, **16(4)**, pp. 211–220.

[13] 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.

[14] 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.

[15] Leou RC, 2006, *A new method for unit maintenance scheduling considering reliability and operation expense*, International Journal of Electrical Power & Energy Systems, **28(7)**, pp. 471–481.

[16] Lindner B & Van Vuuren JH, 2014, *Maintenance scheduling for the generating units of a national power utility*, Proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa, pp. 36–44.

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

[18] 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.

[19] Mromlinski L, 1985, *Transportation problem as a model for optimal schedule of maintenance outages in power systems*, International Journal of Electrical Power & Energy Systems, **7(3)**, pp. 161–164.

[20] Munoz Moro L & Ramos A, 1999, *Goal programming approach to maintenance scheduling of generating units in large scale power systems*, IEEE Transactions on Power Systems, **14(3)**, pp. 1021–1028.

[21] Natrella M, 2013, *NIST/SEMATECH e-handbook of statistical methods*, [Online], [Cited March 2nd, 2015], Available from `http://www.itl.nist.gov/div898/handbook/`.

[22] 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.

[23] Satoh T & Nara K, 1991, *Maintenance scheduling by using simulated annealing method [for power plants]*, IEEE Transactions on Power Systems, **6(2)**, pp. 850–857.

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

[25] 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.

[26] 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.

[27] Smith KI, Everson RM, Member JEF, Murphy C & Misra R, 2008, *Dominance-based multi-objective simulated annealing*, IEEE Transactions on Evolutionary Computation, **12(3)**, pp. 323–342.

[28] Smith KI, Everson R & Fieldsend J, 2004, *Dominance measures for multi-objective simulated annealing*, Proceedings of the 3rd Congress on Evolutionary Computation, pp. 23–30.

[29] Tobias PA & Trindade D, 2011, *Applied reliability*, 2nd Edition, CRC Press, New York (NY).

Proceedings of the 2015 ORSSA Annual Conference
pp. 123–133
www.orssa.org.za/wiki/uploads/Conf/2015ORSSAConferenceProceedings.pdf

ORSSA Proceedings
ISBN
978-1-86822-666-5

Value-based methods for threat value fusion within a ground-based air defense environment

ML Truter* JH van Vuuren†

Abstract

The ability to estimate accurately the threat levels posed by aircraft is a key factor in ensuring success when countering hostile aerial threats in a ground-based air defense environment. Threat evaluation, in this context, is the process whereby aerial threats are prioritised according to the estimated level of danger they pose to the defended system. This process is a high-level information fusion problem aimed at enhancing decision making by fire control officers who are responsible for solving the threat evaluation problem in real time, in order to counter threats effectively. Some context to the process of threat evaluation is provided in this paper, after which a novel data fusion process is proposed which employs two value-based fusion methods in conjunction with one another — a multi-attribute utility function method and an additive aggregation method.

Key words: Threat evaluation, High-level data fusion, Ground-based air defense, Decision support, Aggregation.

1 Introduction

In a *Ground-Based Air Defense* (GBAD) scenario, numerous *Defended Assets* (DAs) are distributed over a geographical region and are afforded protection by a variety of weapon systems which are, in turn, used to counter hostile aircraft [8]. In order to protect the DAs, it is important to be able to quantify the threat level posed by each aircraft to the defended system. This is required so as to understand the risk associated with each aircraft and, consequently, facilitate selection of a suitable countering strategy.

A so-called *Threat Evaluation* (TE) process is an important input process to the weapon assignment process in order to ensure that available ground-based resources (weapons and

*Stellenbosch Unit for Operations Research and Engineering, Department of Industrial Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa, email: louw.truter@gmail.com

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

ammunition) are effectively utilised to defend the DAs. The threat value fusion problem consists of acquiring a comprehensive overview of the level of threat posed by each aircraft in respect of the entire defended system of assets.

TE is achieved by employing aircraft-related data obtained from sensor systems and associated information obtained from pre-programmed data bases (*e.g.* threat-specific information and doctrinal procedures) in order to quantify the threat level posed by detected aerial threats in respect of the defended system. A major concern during this TE process, when different TE models are employed in conjunction with one another, is the fusion of the different threat values returned by these models into a single representative *system threat value*[1].

This paper is structured as follows: An overview of the processes of TE and weapon assignment is provided in §2 within the context of decision support. After introducing the type of *Decision Support System* (DSS) typically employed in this context and various levels of data fusion, the TE process typically implemented is described in §3. This is followed in §4 by an explanation of the proposed data fusion process, with an emphasis on the construction of the utility function and the method of fusion of the different types of threat values. A hypothetical example is used in §5 to illustrate these concepts. The paper closes in §6 with some ideas for future work.

2 Threat evaluation in context

The tasks of analysing aerial threats and assigning ground-based weapon resources to counter these threats in a GBAD environment are the responsibility of a *Fire Control Officer* (FCO). Generally, a *Threat Evaluation and Weapon Assignment* (TEWA) DSS is employed by the FCO to aid him with the processes of TE and weapon assignment. Such a TEWA DSS is, in essence, typically a hybrid DSS and expert system, because it usually employs both computerised analytical methods and heuristic rules[2] to aid the decision making processes of the operator. The functional architecture of such a DSS is depicted in Figure 1. The figure is interpreted in this section with reference to the various levels of information in the well-known (updated) JDL II data fusion model [10].

In Figure 1 it may be seen that the input information required by a TEWA DSS is a combination of real-time sensor data and pre-programmed domain expert knowledge. The real-time sensor data usually include the positional and kinematic data of the detected threats and normally have to be fused together from various sensor sources. This level 0 (source preprocessing) data fusion process, is generally performed within the sensors [10]. The knowledge base, on the other hand, typically includes pre-deployment data on enemy arsenals, threat types and electro-magnetic signatures of known threats.

The data processor unit is a level 1 (object assessment) data fusion process. This process is concerned with the estimation and prediction of relations between the entities (threats,

[1] A *system threat value* provides a holistic view of the level of threat that each aerial threat poses to the defended system as a whole and is, in turn, used within a weapon assignment objective function in an attempt to optimise the assignment of weapon systems.

[2] Heuristic rules are typically developed through experience, intuition and judgment [4].

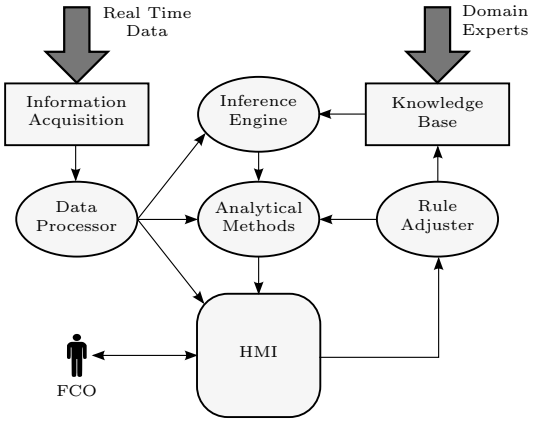


Figure 1: *Functional architecture of a typical TEWA DSS (adapted from [4]).*

DAs) and their relation to the environment, in order to develop the current situational picture from which further impact assessments can be performed. Typical functionality of this component may include triangulation and aircraft track extraction. The inference engine, on the other hand, is a level 2 (situation assessment) data fusion process. This process utilises a combination of the data from the data processor and pre-deployment data from the knowledge base to infer certain characteristics of the threats (threat type, weapon envelope, *etc.*) by using a combination of context-based reasoning, pattern recognition techniques and heuristic rules.

The analytical methods process is the focus area of this paper. This process includes the TE and weapon assignment processes; both are level 3 (impact assessment) data fusion processes. During these processes, the levels of threat of the different aircraft are determined and counter attacks devised [2]. Finally, the results of the TE and weapon assignment processes are displayed on the *Human Machine Interface* (HMI) and the FCO can interact with these solutions in order to select a suitable course of action.

The FCO can usually interact with the HMI through the rule adjuster component so as to configure certain analytical methods for the TE and weapon assignment processes, or update and modify the existing information in the knowledge base. This will ensure that the DSS complements the FCO's analysis style and enhances the FCO's confidence in the DSS.

3 The threat evaluation process

Roux and Van Vuuren [8] proposed three levels of TE models of varied complexity. In order of increasing complexity, they are flagging models, *Deterministic Models* (DMs) and stochastic models. Flagging models are binary in nature and are activated when certain threshold violations occur (*e.g.* sudden increases in altitude or dropping of paratroopers). Stochastic models are probability-based and require detailed information on enemy arsenals, threat types and doctrine. The focus in this paper is only on DMs.

DMs utilise the measured kinematic data from sensors, and calculate derived attributes

which are collectively used to estimate an aircraft's threat value. The estimation criteria used by DMs may include the time to weapon release, or any course, heading or distance-related measure. For the implementation of these models, basic pre-deployment information, such as DA positions, importance values of the different DAs and, in some cases, their orientations as well as the maximum turn radii of attacking aircraft, are required [9]. As a result, the exact input information required depends on the specific DM implemented. Heyns and Van Vuuren [3], as well as Roux and Van Vuuren [9], developed four very specific DMs with the help of domain experts. The principles on which these models rely are described in some detail in [8].

4 Proposed data fusion process

The purpose of the TE fusion component within the TEWA cycle is to combine the results from the various DMs. This fusion must be achieved in a manner that is not only mathematically tractable, but also practical for use in real-world military applications.

All DMs produce a threat value on the real interval $[0, 1]$. The aim of the fusion process should therefore be to construct a value-based prioritised list of threats, based on their system threat values. Value-based in this context refers to a cardinal prioritised list, as opposed to an ordinal list. Multiple techniques for this purpose exist in the field of *Multi-Attribute Utility Theory* (MAUT) [5]. These different techniques may be classified as *value measurement models*, *goal aspiration models* or *outranking models*.

Value-based measurement models are the only models in which a numerical preference score is calculated pertaining the degree to which a certain alternative may be preferred above another. The results are therefore quantitative in nature and the level of preference of one alternative to another is retained during fusion.

When utilising the DMs referred to in the previous section, a threat value is determined for each threat-DA-DM triple. If there are, for example, three incoming threats, two DAs to protect, and four different DMs, then a total of 24 different threat values will therefore be calculated. The output of the TE subsystem should, however, be a single system threat value for each threat in order for the weapon assignment subsystem to function effectively.

The purpose of our proposed data fusion process is to fuse together these different threat values — which are distinguished according to threat, DA and DM — so as to obtain a single system threat value per threat. Different hierarchies of threat values are shown in Figure 2. Since the DMs are typically configured differently, the first step should be to obtain a threat value with respect to each threat-DA pair (*i.e.* to fuse together the threat values obtained by the different DMs). After obtaining the theat-DA threat list thus fused, the importance weights of the DAs may finally be used to fuse together a single system threat value for each threat.

4.1 Computation of threat-DA pair threat values

Evaluation of threat values according to the DMs described in §3 is updated each time new information is received from the sensor systems, the duration of this update-cycle is

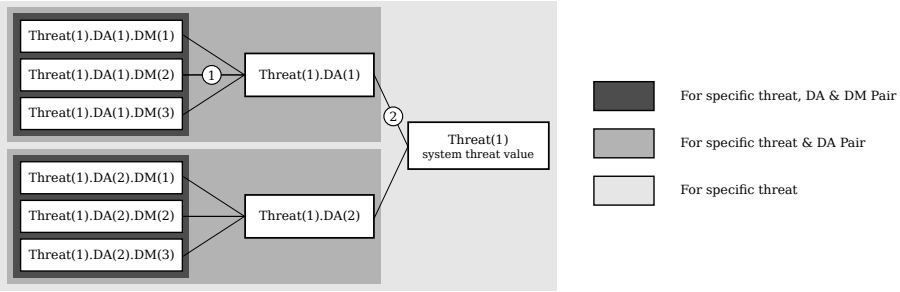


Figure 2: *Relationships between different threat values. Fusion process (1) involves multi-attribute utility functions within an aggregated value function tree in order to obtain an individual threat value per DA. Fusion process (2) involves the additive weighting method, where the weighting coefficients are the normalized importance values of the DAs.*

known as the TEWA *cycle time*. Consequently, the deterministic threat values are updated throughout the engagement as the threats approach the DAs.

This calls for a fusion model which is dynamic in nature. Furthermore, as threats approach the DAs, certain DMs become more accurate or relevant in estimating an aircraft's threat level. For fixed-wing aircraft, for instance, when threats are far from the DAs, the time-to-weapon-release may be a better predictor of threat level than distance-related threat measures. To illustrate, at long ranges when two threats are executing the approaching phases of their attack profiles at the same distance from the DAs, time should be a better predictor of threat level. Although they are the same distance from the DAs, the threat with the higher velocity poses a more imminent danger to the DAs, since it would be able to release its weapons earlier. In contrast, if threats are entering the manuevre phases of their attack profiles (in anticipation of weapon release), distance-related measures ought to be a better predictor of threat level, since time-related measures become increasingly difficult to predict accurately during these final phases of the attack profile.

It is therefore clear that a multi-attribute utility function is required in which the attributes are the DMs. The utility function should provide a single threat value for each threat-DM pair, where the fused threat value depends on the threat values of the DMs combined. For illustrative purposes, three spatial DMs are considered — a slant distance model, a passing distance model[3] and an altitude-related model. These three models were specifically chosen in order to adhere to the requirements of utility- and preferential independence when constructing a multi-attribute utility function.

In order to construct the utility function, it is first required to quantify the preferences of the end-users. To this end, the threat-value interval may be subdivided into three intervals; the midpoints of these intervals may then represent respectively high (0.83), medium (0.5) and low (0.16) threat values. This concept is illustrated in Figure 3.

The intervals in Figure 3 are suggested in order to facilitate effective elicitation of end-user

[3]The implemented passing distance model is generally referred to as the *Closest Point of Approach* (CPA) model in the literature. Passing distance is used here in order to aid understanding to a non-military readership.

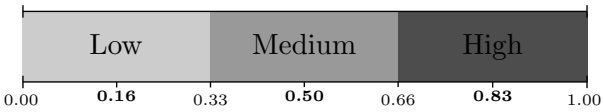


Figure 3: *Threat value intervals.*

preferences. It is anticipated that domain experts should be able to provide an appropriate or expected combined threat value when the threat value of an altitude DM and a passing distance DM are low, for instance, but that of a slant distance DM is high. This elicitation process may be repeated for all the different combinations of the DM threat levels in order to obtain a combined or characteristic threat value for each triple. This is required in order to serve as input for the construction of an accurate and robust utility function for the fusion process.

In an attempt to limit the amount of personal subjectivities of domain experts and to address the difficulties when aggregating individual preferences, it is advocated that a group of military experts attend a workshop and discuss the different alternatives for each of the criteria, with the goal of reaching group consensus in respect of input values for construction of the utility function.

Hypothetical threat values for the aforementioned combinations were selected for use in a proof-of-concept example. MATLAB was used to fit various functions through the resulting 27 data points. It was determined that a third degree, three-variable polynomial provides the best fit. The resulting function,

$$\Gamma(\gamma, \rho, \eta) = \sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{k=0}^{3} a_{ijk} \gamma^i \rho^j \eta^k, \tag{1}$$

returns a fused characteristic threat value in the real interval $[0, 1]$ for a specific threat-DA pair. In (1), the symbols γ, ρ and η denote slant distance, passing distance and altitude DM threat values, respectively. The values of the coefficients a_{ijk} are shown in Table 1. All values not in the table assume a value of zero.

Table 1: *Coefficients for the multi-attribute utility function (1).*

$a_{000} = -0.12$	$a_{100} = 1.3$	$a_{200} = -1.2$	$a_{300} = -0.48$	$a_{110} = -0.74$	$a_{210} = 0.52$	$a_{101} = -1.8$
$a_{201} = 1.5$	$a_{010} = 0.45$	$a_{020} = 0.31$	$a_{030} = -0.21$	$a_{001} = 0.42$	$a_{002} = 0.8$	$a_{003} = -0.54$

4.2 Threat-DA threat value scaling

Another consideration in the fusion process, not described above, is the scaling of threat values. TE is typically conducted on all threats within an *Area of Responsibility* (AOR) which are identified as hostile or unknown. Doctrine often requires that if a threat enters a prespecified distance from a DA — here referred to as the *Keep-Out Boundary* (KOB) — the threat must be classified as highly threatening with respect to the DA considered. If the assumption is made that threats are classified according to three priority categories,

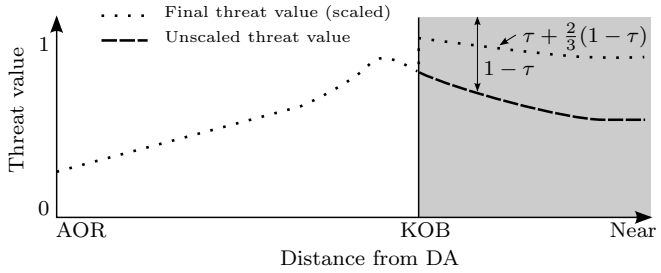


Figure 4: *Suggested threat value scaling method when threats cross the KOB.*

namely low, medium and high (as illustrated in Figure 3), then the scaling should ensure that any threat is classified as highly threating when it enters the KOB.

The scaling should therefore make provision for increasing the threat priority of low and medium threats to high priority when a threat crosses the KOB. It is advocated that the threat value τ of a threat crossing the KOB should be increased by $\frac{2}{3}(1-\tau)$, as illustrated in Figure 4. This scaling of the threat values will ensure that the low and medium threats are always classified as highly threatening threats when inside the KOB. This is true if the threat values τ of the classes of low, medium and high priority threats occupy the ranges $0-\frac{1}{3}$, $\frac{1}{3}-\frac{2}{3}$ and $\frac{2}{3}-1$, respectively. In practice, however, the scaling value (suggested here to be $\frac{2}{3}$) should be agreed upon by domain experts. The result of this process is a scaled threat-DA threat list.

4.3 Computation of system threat values

After calculating a single threat value for each threat-DA pair, it is required to fuse these values together in order to obtain a system threat value for each threat. One way to achieve this is to use the importance value of each DA as a linear weight applied within the additive weighted method. A linear weighting may be applied to all the criteria independently in order to obtain a system threat value for each threat.

The relative DA importance weights should be determined prior to system implementation and should therefore be stored in the knowledge base. Each DA usually has an importance value assigned to it by the defending force. This importance value quantifies the relative importance to the defending force of protecting the DA in question. Several variables may influence the importance of a specific DA, such as its repairability, vulnerability and vital importance [7].

Repairability of a DA refers to its ability to recover from damage inflicted to it, and is usually determined based on the manpower, equipment and time required to repair the asset to a functioning state. *Vulnerability*, on the other hand, refers to the extent to which an asset is susceptible to damage and surveillance during an attack. Armour, position, countermeasures and camouflage are factors which influence a DA's vulnerability. Finally, *vital importance* is the degree to which the mission's success relies on a specific DA. Assessing the impact that the destruction of an asset will have on the mission's success is one way of determining its vital importance value. A command centre, for example, is more important in respect of ensuring mission success (for maintaining command and

control superiority) than a redundant (backup) sensor system.

The DA importance values, together with the threat-DA threat values, may be fused together by an additive weighting function to obtain the system threat value

$$V_a = \sum_{k=1}^{n_D} v_{k,a} \cdot \psi_k$$

associated with threat $a \in \{1, \ldots, n_T\}$, where ψ_k denotes the normalised importance value of DA $k \in \{1, \ldots, n_D\}$. Furthermore, $v_{k,a}$ represents the (fused) threat value of threat a and DA k pair. The total number of DAs is denoted by n_D and the total number of threats by n_T.

5 Worked example

A worked example is provided in this section in order to illustrate and clarify the concepts of the preceding sections. A hypothetical ground-based air defense scenario is illustrated in Figure 5. In this figure the threat paths of three threats are indicated by the three lines, and the positions of two DAs are depicted by black dots. The differently colored domes represent the single-shot hit probability distribution volumes of two ground-based weapon systems which are indicated by the two crosses. The weapon systems are only included for the sake of completeness and their assignment does not form part of this worked example.

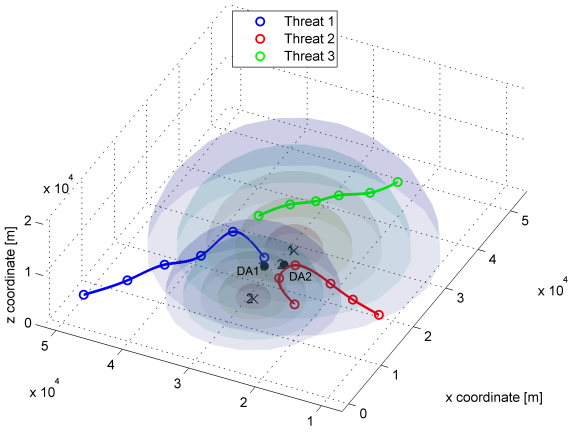


Figure 5: *Hypothetical ground-based air defense scenario.*

The threat tracks in Figure 5 correspond to a scenario spanning 120 seconds. For the purpose of this example, the TEWA cycle is repeated every second (*i.e.* the real-time data updates are assumed every second). Furthermore, only three spatial DMs are implemented — a slant distance model, a passing distance-related model and an altitude-related model. The resulting threat values are shown in Figure 6 as a function of time.

Although the DMs are unaware of this, Threat 1 is executing a pitch-and-dive attack manoeuvre in respect of DA 2. Threat 2, on the other hand, is executing a toss-bomb

attack manoeuvre in respect of DA 1. Finally, Threat 3 is a passenger aircraft passing over the conflict zone and therefore poses no real threat. Threat 3 is only present to ascertain the response of the TE algorithms in the case where an aircraft is not attacking the DAs.

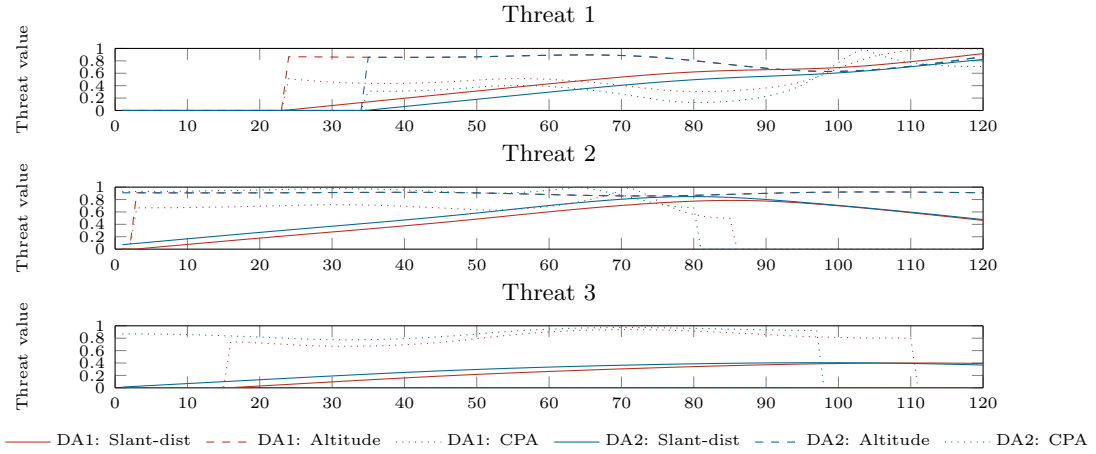


Figure 6: *Original threat values; distinguished in terms of threat, DA and DM.*

The threat values shown in Figure 6 were fused together in order to obtain a single threat value for each threat-DA pair. This fusion process was achieved using the aggregated value function approach described in §4.1. The $\frac{2}{3}$–KOB scaling referred to at the end of §4.3 was subsequently applied to the unscaled threat-DA threat list.

After obtaining the scaled threat-DA threat value list, these threat values were fused together using the normalised importance weights of the DAs. The additive weighting method described in §4.3 was used for obtaining the system threat values shown in Figure 7. These threat values may be used for ground-based weapon assignment purposes.

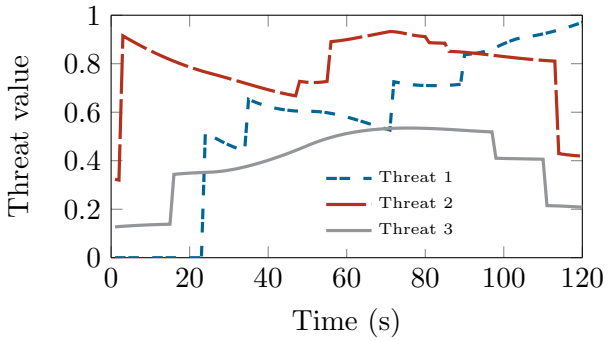


Figure 7: *System threat values as a function of time.*

From Figure 7 it is clear that the threat values provide realistic treat estimates of the various threats. The sudden rises and drops in threat values occur when certain models are "switched on" or "switched off." For example, a threat must be within the prespecified AOR radius before TE is conducted in respect of the threat. Similarly, the passing distance-related DM is also only active if the threat is heading towards a DA. It is worth mentioning that Threats 1 and 2 are scheduled to release their weapons at approximately times 90 and 60, respectively. It is therefore encouraging to note that the

fused threat values of these threats are at their highest levels close to the weapon release stage. Finally, it is heartening that the threat values associated with Threat 3 which is, in fact, not exhibiting threatening behaviour, is significantly lower than those of Threats 1 and 2.

Rapid changes in the threat values may be of concern to the effective functioning of the TEWA system, as described by Lötter and Van Vuuren [6]. The rapid changes in threat values, observed at times 3, 22, 37, 58, 72, 91 and 116 in Figure 7, may result in switching of weapon assignment recommendations. This switching is a typical emergent property of TEWA systems and is something that must be fully understood before implementing such systems. These switching recommendations may cause confusion on the part of the operator. An analysis of the extent of this switching behaviour is something that is an aspect of our current research. Several mitigation strategies are being investigated to alleviate this problem. Allouche [1] did similar work by implementing Kohonen's self-organising maps for the stabilisation of threat values. Threat value stabilisation was achieved by smoothing the observed real-time trajectory of the threats (anti-ship missiles in his case).

6 Conclusion and future work

The work reported in this paper forms part of a larger project which entails the performance evaluation of TEWA algorithms developed during the period 2006–2010 at Stellenbosch University, as described in [11]. Although the focus here is on higher level data fusion processes (levels 2–3 in the JDL fusion model), it is nevertheless important also to consider the effects that inaccurate, noisy input information may have on the output values of the TEWA algorithms. The performance evaluation of the system will therefore include an investigation into the sensitivity of various system parameters as well as the effects of uncertain input information on the system.

References

[1] Allouche MK, 2005, *Real-time use of Kohonen's self-organizing maps for threat stabilization*, Information Fusion, **6(2)**, pp. 153–163.

[2] Falzon L, 2006, *Using Bayesian network analysis to support centre of gravity analysis in military planning*, European Journal of Operational Research, **170(2)**, pp. 629–643.

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

[4] Ignizio JP, 1991, *An introduction to expert systems*, 1st Edition, McGraw-Hill, New York (NY).

[5] Keeney RL & Raiffa H, 1993, *Decisions with multiple objectives: Preferences and value tradeoffs*, 1st Edition, Cambridge University Press, Cambridge.

[6] Lötter DP & Van Vuuren JH, 2014, *Implementation challenges associated with a threat evaluation and weapon assignment system*, Proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa, pp. 27–35.

[7] Roux JN, 2010, *Design of a threat evaluation subsystem in a ground-based air defence environment*, PhD Thesis, Stellenbosch University, Stellenbosch.

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

[9] 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.

[10] Steinberg AN & Bowman CL, 2004, *Rethinking the JDL data fusion levels*, National Symposium on Sensor and Data Fusion, **38**, pp. 39.

[11] Truter ML & Van Vuuren JH, 2014, *Prerequisites for the design of a threat evaluation and weapon assignment system evaluator*, Proceedings of the 43rd Annual Conference of the Operations Research Society of South Africa, pp. 54–61.