– Proceedings –

40th Annual Conference of the Operations Research Society of South Africa

18–21 September 2011
Elephant Hills Hotel, Victoria Falls, Zimbabwe

ISBN: 978-0-7972-1351-7

Proceedings of the 40^{th} Annual Conference of the Operations Research Society of South Africa

ISBN: 978-0-7972-1351-7

Editorial Board

Review process

Thirty-four (34) manuscripts were submitted for possible inclusion in the *Proceedings of the 40^{th} Annual Conference of the Operations Research Society of South Africa, 2011.* All submitted papers were double-blind peer-reviewed by at least two independent reviewers. Papers were reviewed according follow criteria: technical quality, i.e. correct use of language, clarity of expression, quality and justification of arguments; and on the contribution to Operations Research, i.e. knowledge of field, quality and consistency of referencing, significance of contribution and suitability for conference proceedings. Of the thirty-four (34) submitted papers, fourteen (14) were ultimately, after consideration and incorporation of reviewer comments, judged to be suitable for inclusion in the proceedings of the conference. The proceedings will be published online at `http://www.orssa.org.za/XXXXX`.

Reviewers

The editorial would like to thank the following reviewers:

Wilna Bean	Council for Scientific and Industrial Research, South Africa
Wenwi Cao X	Georgia Tech, USA
Edward Chiyaka	National University of Science and Technology, Zimbabwe
Ian Durbach	University of Cape Town, South Africa
Dave Evans	Development bank of Southern Africa
Pieter Fourie	Swiss Federal Institute of Technology Zurich, Switzerland
Jan Greben	Council for Scientific and Industrial Research, South Africa
Jenny Halloway	Council for Scientific and Industrial Research, South Africa
Marthi Harmse	Sasol Technology, South Africa
Hans Ittmann	Council for Scientific and Industrial Research, South Africa
Johan Jansen van Rensburg	Sasol Technology, South Africa
Sibusiswe Khuluse	Council for Scientific and Industrial Research, South Africa
Renee Koen	Council for Scientific and Industrial Research
Philimon Nyamugure	National University of Science and Technology, Zimbabwe
Nadia Viljoen	Georgia Tech, USA
Elias Willemse	Council for Scientific and Industrial Research, South Africa

Elias J Willemse
(e) ejwillemse@gmail.com
(t) +27 71 890 2714
Editor-in-Chief: ORSSA 2011 Proceedings

Table of contents

Proceedings of the 2011 ORSSA Annual Conference
pp. 1–10
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Bankruptcy prediction in South Africa

A Maeteletsa[*] JW Kruger[†]

Abstract

Bankruptcy prediction modeling and studies are known to have existed since the 1960s. In this report a brief overview is given of the theory with reference to Altman's use of Multivariate Discrimination in the formulation of his Z-score model. This study looks at the impact of time in the classification or prediction accuracy. Two sets of 71 failed and non-failed companies listed in the Johannesburg Stock Exchange were used. The companies come from various sectors and range from the period of 1998 to 2007 excluding 2006. The data were subdivided into three parts according to periods and models were formulated and tested with five multivariate discriminate analysis functions and one classification tree algorithm. The results show that there is no impact in the classification or prediction accuracy due to the specific time period the model was used. It was further verified that the classification accuracy is independent of whether the model was chronologically closer or further than the test sample. The third test verified that the classification accuracy is independent of whether the time period is relatively narrow or wider. The overall result implies that the time period the model was created has no effect in the classification or prediction accuracy. This view is best when limited to a specific business cycle.

Key words: Bankruptcy, Financial Distress, Bankruptcy Prediction, Bankruptcy Modeling.

1 Introduction

A bankruptcy prediction model is one among various financial models that has been studied for almost eight decades. Ever since the studies in the subject evolved to the current state, they have been regarded as very important to various stakeholders within and outside; the corporate, political and social spheres surrounding businesses. There are various financial models in use today for corporate decision support applications. The common objective in the studies of bankruptcy prediction modeling is to establish the model (or modeling technique) that can classify or predict accurately and consistently whether the company will fail (i.e. cease to exist) or not.

[*] Corresponding author: UNISA SBL, South Africa, email: maeteletsar@yahoo.co.uk
[†] UNISA SBL, South Africa, email: jkruger@sbleds.ac.za

2 Literature review

The significant number of papers sought by the researcher in the subject of financial distress, insolvency and bankruptcy cite and reference the work of Edward Altman, and Beaver. Altman has written extensively in the field of corporate bankruptcy and financial distress [1, 2, 3, 5, 6, 10, 12]. The bankruptcy prediction studies include differences in the approach of modeling and the different set of parameters in formulating the model. An example of this is when for instance Klelinman and Anandarajan [7] used qualitative measures (non-financial cues) and Xu and Wang [17] used organisational efficiency as predictor. In the 1960s, Altman and Beaver used financial data from industrial, mostly manufacturing firms to develop their models [10].

In the 1960s Beaver used the univariate approach while establishing the financial distress prediction model. This approach, though it involved a use of many variables (ratios), each ratio was used distinctively to model and predict bankruptcy. This model was later expanded into multivariate framework by Altman [16].

Discriminant analysis became a prominent method for financial failure prediction till the 1980s when the logistic regression method was accentuated. There are slight and wide variations on the data used to develop the financial distress prediction models. The slight variations apply to a choice of different and total number of variables (financial ratios) used [13]. Wide variations apply to different data (i.e. either or both country and industry specific) to develop prediction models. This includes data that is non-financial in nature, and the emphasis tends to look at management attributes and economic effects [14, 7]

3 Multi-states of a failure process

The researchers; Lau [9], Laitinen [8] and Naidoo and Du Toit [11] have extended the distress prediction studies from the state of failed and healthy (i.e. dichotomous) to include other intermediary states. The approach of Naidoo and Du Toit [11] has four ordinal states; Healthy, Intermittent, Distressed and Severely Distressed, with the latter state equated to bankruptcy. Laitinen [8] also has four phases to failure and they are; Starting, Intervening, Final and Exit phases.

It appears; the main argument behind the extension of this theory is the basis on which failure is seen as a continuous process and not only as an instantaneous event. The support of this view has the researchers using financial data of companies from periods more than the one just before failure i.e. using couple of years of financial data before the occurrence of failure. This also tends to indicate the ability of the model to predict failure accurately a couple of years before it actually takes place [2, 4].

The theory of financial failure prediction does not attempt to establish or define the various causes of corporate failure, but it seem to stand on the premise that given sufficient time, the effect and impact of and leading to failure will be noticeable in financial data. The sufficiency of time is depended on how frequent the financial data is made publicly available and the integrity of the data.

The literature does not define the process of how to determine the boundary or cut-off between failed and non-failed explicitly for dichotomous models. But noticeable in the literature is the common objective to develop a model with highest classification/prediction accuracy. This makes the researcher conclude, this is as observed in Altman's work, that the cut-off point or boundary is inherently dependent on minimizing the classification/prediction error, this conversely implies maximizing the classification/prediction accuracy [3]. It is only in the studies that view failure as a process with multi-states that the boundaries are given definitions.

4 Research Problem

A predictive model is when a model is formulated prior (or with data from time-period prior) to its application whereas for classification purposes, this sequence of time period is irrelevant. Smith and Lioud [15] allude to the necessity to collect data of failed companies from a potentially lengthy time period to formulate industry-specific bankruptcy prediction model. The implied claims by various researchers as mentioned in Smith and Graves [14] that bankruptcy prediction models transcend the time period and industries other than those that were used to formulate it, lead the researcher to the question below with constituent sub-questions: Does the time period the model is formulated impact on the classification or predictive accuracy of the model? The sub-questions that aggregate to the above question are:

1. Does the specific time period used to formulate the model have an impact in the classification accuracy of the model?
2. Does the model formulated on the dataset with the time period adjacent to or further from the test sample impact the predictive and classification accuracy?
3. Does the time period length (longer/wider or shorter/narrower) from which the model is formulated have any impact in the classification accuracy of the model?

These questions are further assigned to hypothetical assertions to be tested in the research. The following are the hypotheses:

Impact of specific time period in the classification accuracy of the model

H_0 : The classification accuracy of the bankruptcy prediction model is not affected by the period the model was created in.

H_1 : The classification accuracy of the bankruptcy prediction model is affected by the period the model was created in.

Impact of closeness of time period of model to sample in the classification accuracy of the model

H_0 : The closer (or adjacent) the time period to formulate the model to test sample, the more accurate it is.

H_1 : The further the time period to formulate the model to test sample, the more accurate it is.

Impact of time period length of data to formulate model in the classification accuracy of the model

H_0 : The longer the time period to formulate the model, the more accurate it is.
H_1 : The shorter the time period to formulate the model, the more accurate it is.

5 Research Design

This research study is exploratory and quantitative in nature. The selection of failed or liquidated or winded-up companies used in this study were to be matched with healthy, operational and listed ones; this is a common practice in the studies of bankruptcy prediction. Then a set of specific ratios (see Appendix A) are evaluated from financial data extracted for each company in the distinct sets of failed and healthy companies (i.e. two sets).

6 Methodology

The research design aimed at answering the research question as well as testing the assertions made in the hypotheses led to data collection over a time period say T. Then this data was sub-divided into three distinct adjacent parts (T_1, T_2, T_3), and as illustrated in figure 1. The sub-division of dataset into these sub-periods had to consider data limitations and the critical statistical requirements to have the sample size quantitatively bigger than the variables to be observed in the sample.

The sub-periods T_1 and T_3 were then used to formulate the models which are then tested for classification/predictive accuracy on dataset from period T_1, T_2 and T_3. The sub-division of dataset into this period is aimed at answering the first and second sub-questions. This will then test the assertions made in the first and second hypotheses.

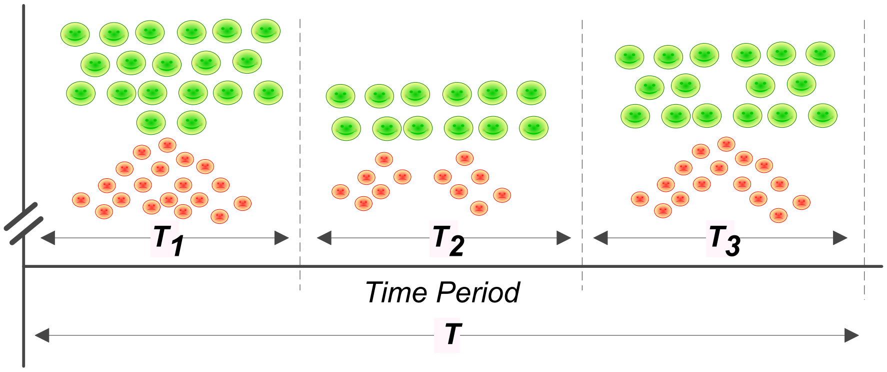


Figure 1: *Illustration of narrow time model*

The second part in the research design aimed at answering the third sub-question was designed such that, the same total data as above (illustrated in figure 1) over period T, was further stratified sampled. The Stratified sample was such that there was at minimum, one pair of companies (failed and non-failed) for each financial year in the chosen period. The total stratified sample was limited to be at least a third of the total initial sample. The model formulated out of this sample was designated T_m as the period over which the model was formulated was relatively wider compared to any of T_1, T_2 or T_3 as mentioned above (illustrated in figure 2).

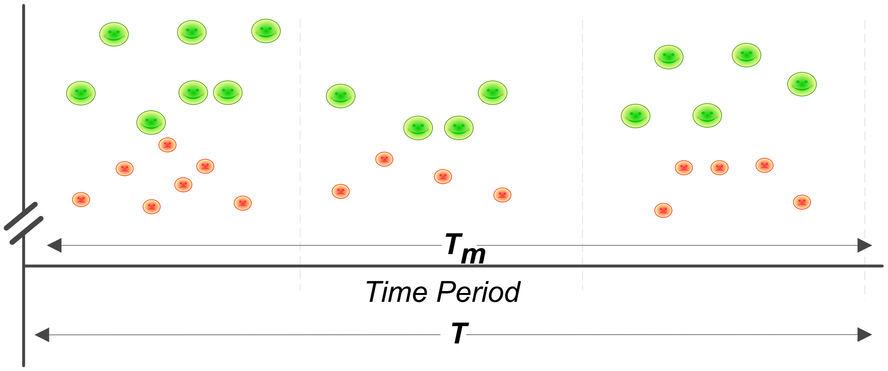


Figure 2: *Wide period model T_m (about a third of T)*

The remaining dataset (about two thirds) of the data was then to be used as the test sample for the model T_m as was designated as T_t. This test set is illustrated in figure 3. The model and the test sample were aimed at testing an assertion of the third hypotheses as well as classification accuracy in the test that is contemporary with the test set.

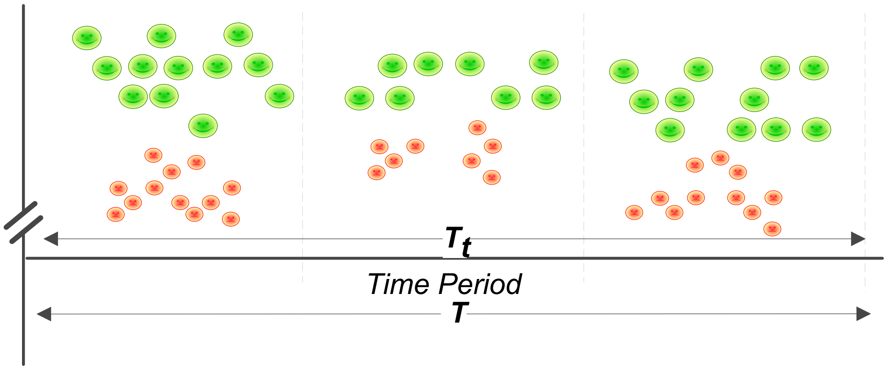


Figure 3: *Wide period test sample T_t (two thirds and remainder of T).*

7 Sampling and Data collection

This research study was limited to the companies that were initially listed in the Johannesburg Stock Exchange (JSE) and those that are currently listed. This study focuses on the companies that according to the JSE list; have been "*liquidated*" or "*voluntary liquidated*" and those that have "*winded up*" or have "*voluntary winded up*" in the period 1998 to 2007. The other JSE listed companies that were selected in the study, are also regarded as healthy and operational as at the time of the study.

The healthy companies were selected to match each failing company by industry (or sector) and not asset size as Altman [1] did. The other important requirement was to ensure that the healthy matching company had to be operational and listed for a minimum period of 5 years from the matching year. The last financial data of a failing company was regarded as the last year of operation; the matching healthy company financial data of the same year (as failing) were used. The quantity of both failed and non-failed companies per financial year is shown in table 1.

Table 1: *A tabulation of paired sample companies and the time period*

Financial Year	1998	1999	2000	2001	2002	2003	2004	2005	2007	Grand Total
Failed Companies	5	9	18	8	11	9	4	4	3	71
Non-Failed Companies	5	9	18	8	11	9	4	4	3	71
Total paired companies	10	18	36	16	22	18	8	8	6	142

The financial data of both failing and non-failing companies were drawn from the McGregor BH database. This was a total of 142 companies as depicted in table 1. The ratios in which either the numerator or denominator did not exist were set to zero. This was done so as to avoid the error of dividing by zero.

8 Analysis and Modeling

The data was analyzed and modeled using the Matlab^{TM} 2010a student version; The Matlab^{TM} has built-in statistical analysis tools with various Multivariate Discriminate Analysis (MDA) functions and Treebagger function. The latter function is part of the classification trees algorithms which does not require data to have normal distribution.

9 Modeling Results

In the results that follow in tables 2 and 3 for both the narrow and wide time period modeling, there were five MDA functions used and the Treebagger algorithm which forms part of the classification trees.

10 Discussion of Results

The results of narrow time period depict the expected trend for holdout samples, in that the holdout sample will have on average, higher classification accuracy results than the classification and or prediction tests. The results from models which statistically require dataset to be normally distributed are distinctly different from that which does not. This is seen in reference to Treebagger classification tree algorithm; this has yielded results of 100% for all

Table 2: *Narrow time-period modeling results with overall classification/prediction accuracy*

Model	T_1 on T_1	T_1 on T_2	T_1 on T_3	T_3 on T_1	T_3 on T_2	T_3 on T_3	Average
Linear	57.80%	39.50%	50.00%	50.00%	52.60%	67.50%	52.90%
DiagonalLinear	48.40%	44.70%	40.00%	45.30%	57.90%	72.50%	51.50%
Diagquadratic	60.90%	50.00%	50.00%	48.40%	52.60%	62.50%	54.10%
Quadratic	67.20%	52.60%	50.00%	43.80%	47.40%	95.00%	59.30%
Mahalanobis	81.30%	50.00%	50.00%	48.40%	42.10%	90.00%	60.30%
Treebagger	100.00%	55.30%	62.50%	55.30%	47.40%	100.00%	70.10%
Average	69.30%	48.70%	50.40%	48.50%	50.00%	81.30%	

Table 3: *Wide time-period modeling results with overall classification/prediction accuracy*

Model	Tm on Tm	Tm on Tt	Tt on Tm	Tt on Tt	Average
Linear	67.30%	43.30%	48.10%	66.70%	56.30%
DiagonalLinear	67.30%	41.10%	50.00%	61.10%	54.90%
Diagquadratic	57.70%	53.30%	53.90%	61.10%	56.50%
Quadratic	73.10%	55.60%	44.20%	67.80%	60.20%
Mahalanobis	71.20%	60.00%	44.20%	65.60%	60.20%
Treebagger	100.00%	50.00%	53.90%	100.00%	76.00%
Average	72.80%	50.60%	49.00%	70.40%	

holdout sample tests, compared to other statistical algorithms which yielded results in the ranges 48.4% to 81.3% and 62.5% to 90.0% for T_1 on T_1 and T_3 on T_3 respectively. The overall average holdout accuracy is 69.3% and 81.3% for the tests T_1 on T_1 and T_3 on T_3 respectively.

The prediction results of narrow time period ranges in 39.5% to 55.3% with a mean of 48.7% and 40.0% to 62.5% with a mean of 50.4% for T_1 on T_2 and T_1 on T_3 respectively. The results from a Treebagger algorithm are the highest in the test set.

The holdout sample results of wider period tests ranges from 57.7% to 100% with a mean of 72.8%. The Treebagger algorithm yields 100% classification accuracy on this holdout sample. When the Treebagger algorithm is excluded the mean is still reasonably high at 67.3% compared to Table 5 where the average is 63.1% and 77.5% for T_1 on T_1 and T_3 on T_3 respectively. The overall classification results of wider period tests ranges from 41.1% to 60.0% with a mean of 50.6%.

11 Hypotheses Results

The hypotheses mentioned earlier are in the same order as the sub-questions that aim to test the assertions made.

a) Does the specific time period used to formulate the model have impact in the classification accuracy of the model?

The result shows that since the P (two-tail) value is 0.69 and is higher than the rejection threshold P(0.05), then there is a failure to reject the null hypothesis. This result implies that the classification accuracy of a bankruptcy prediction model is not affected by the period the model was created in.

b) Does the model formulated on the dataset with the time period adjacent to or further from the test sample impact the predictive and classification accuracy?

The result shows that since the P value (one-tail) is 0.0005, and is smaller than the rejection threshold P(0.05), then the null hypothesis is rejected in favor of the alternative. This result implies that the further the time period to formulate the model to test sample, the more accurate it is. The hypothesized difference in the test was 15%. Despite the hypothesis results this is inconclusive since the difference between the Means is less than 2%.

c) Does the time period length (longer/wider or shorter/narrower) from which the model is formulated have any impact in the classification accuracy of the model?

The result shows that since the P value (one-tail) is 0.003 and is smaller than the rejection threshold P(0.05), then the null hypothesis is rejected in favor of the alternative. The shorter the time period to formulate the model, the more accurate it is. The hypothesized difference in the test was 15%. Despite the hypothesis results this is questionable and is regarded as inconclusive since the difference between the Means is less than 2%.

12 Conclusions

The conclusions drawn from this study depict that in the bankruptcy prediction modeling done:

- The classification or prediction accuracy of bankruptcy prediction models in the study are not affected by the specific time period used to formulate the model relative to when it is applied.
- The classification or prediction accuracy results in the study are independent of whether the data used in the model was chronologically closer or further from the time it was applied.
- The classification or predictive modeling accuracy results are independent of whether the time period was narrow or wider in this study. The important statistical requirement is to ensure that the sample size was bigger than the number of variables observed. In this study the minimum ratio was 13 variables to 40 observations.

The summary of the conclusion points made above, leads to the answer of the main question of this study, that the time period the model was formulated has no effect or no impact in the bankruptcy prediction modeling accuracy. This conclusion is best when limited to the dataset (including the specific choice of ratios used) and modeling techniques used in this particular study. There has been no attempt in this study to make reference to the business cycle which is also known impact in the failure of businesses. This further extends the constraint of this conclusion that it is best when limited to a known business cycle period.

Bibliography

[1] Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):171–198.

[2] Altman, E. (1984). The success of business failure prediction models: An international survey. *Journal of Banking and Finance*, 8:171–198.

[3] Altman, E. (1993). *Corporate Financial Distress and Bankruptcy.* John Willey and Sons Inc.

[4] Beaver, W. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, Supplement:71–127.

[5] Grice, J. and Dugan, M. (2001). The limitations of bankruptcy prediction models: Some cautions for the researcher. *Review of Quantitative Finance & Accounting*, 17:151–166.

[6] Grice, J. and Dugan, M. (2003). Re-estimation of the zmijewski and ohlson bankruptcy prediction models. *Advances In Accounting*, 29:77–93.

[7] Klelinman, G. and Anandarajan, A. (1999). The usefulness of off-balance sheet variables as predictors of auditors going concern opinions: an empirical analysis. *Journal of Managerial Auditing*, 14(6):273–285.

[8] Laitinen, E. (1993). Financial predictors for different phases of the failure process. *International Journal of Management Science*, 21(2):215–228.

[9] Lau, H. (1987). A five state financial distress prediction model. *Journal of Accounting research*, 25(1):127–138.

[10] Lincoln, M. (1984). An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk. *Journal of Banking and Finance*, 8:321–340.

[11] Naidoo, S. and Du Toit, G. (2007). A predictive model of the states of financial health in south african businesses. *Southern African Business Review*, 11(3):33–55.

[12] Purnanandam, A. (2008). Financial distress and corporate risk management: Theory and evidence. *Financial distress and corporate risk management: Theory and evidence*, 87:706–739.

[13] Rushinek, A. and Rushinek, S. (2008). Financial distress and corporate risk management. *Journal of Financial Economics*, 89:706–739.

[14] Smith, M. and Graves, C. (2005). Corporate turnaround and financial distress. *Managerial Auditing Journal*, 20(3):304–320.

[15] Smith, M. and Lioud, D. (2007). Industrial sector and financial distress. *Managerial Auditing Journal*, 22(4):376–391.

[16] Ugurlum, M. and Aksoy, H. (2006). Prediction of corporate financial distress in an emerging market: the case of turkey. *Cross Cultural Management: An international Journal*, 13(4):277–295.

[17] Xu, X. and Wang, Y. (Article in press). Financial failure prediction using efficiency as a predictor. *Expert Systems with Applications.*

Appendix

List of Variables used in research

No	Ratio (Variable)	Ratio Description
1	EBIT/TA	Earnings Before Interest and Taxes / Total Tangible Assets
2	NATC/TC	Net available for total capital / Total Capital
3	Sales/TA	Sales / Total Tangible Assets
4	Sales/TC	Sales / Total Capital
5	EBIT/Sales	Earnings Before Interest and Taxes / Sales
6	Int Cov	Interest Coverage (= EBIT/Interest Expense)
7	WC/Cash Exp	Working Capital / Cash Expense
8	WC/TA	Working Capital / Total Tangible Assets
9	Current Ratio	Current Assets / Current Liabilities
10	WC/LTD	Working Capital / Long-term Debt
11	Ret E/TA	Retained Earnings / Total Tangible Assets
12	Book Eq/TC	Book Equity/ Total Capital
13	MV Eq/TC	Market Value equity / Total Capital

Proceedings of the 2011 ORSSA Annual Conference
pp. 11–18
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

The best posterior probability model could be different from the actual causal model

JW Kruger*

Abstract

An example is shown that the best posterior probability model, as found by the Cooper and Herskovits equation, is not always the causal model. This involves moving from a causal model, to a frequency distribution, to a correlation matrix and back to a model. The elimination heuristic, utilising independence relations, is introduced as a first step towards getting the actual causal model. At this stage this method is not developed enough to find the causal model.

Key words: Bayesian Belief Networks, Causal Model, Posterior probability

1 Introduction

This paper finds an example that the best posterior probability model as found by the Cooper and Herskovits equation is not the actual causal model. The paper starts by giving the Cooper and Herskovits equation. I then found an example to show that it does not always identify the correct causal model. Thereafter the elimination heuristic is introduced where a fully connected network is trimmed by identifying and eliminating independence relations. With a tree model this elimination heuristic builds the same model as the Cooper and Herskovits equation, but with more complicated causal models it can identify a different model. The main reason to revisit the Cooper and Herskovits equation is to show that an alternative approach is needed to find causal structure.

In Section 2 the graphical language used to reason with causality is explained. The Cooper and Herskovits [4] equation is given in Section 3 and in Section 4 an example is given that the best posterior probability causal model is fallible. In Section 5 the elimination heuristic is discussed as a first step to find the correct causal model, while in Section 6 guidelines for future research is given as well as some general conclusions and comments.

*Corresponding author: UNISA SBL, South Africa, email: jkruger@sbleds.ac.za

2 Graphical Language

Graphical representations offer a framework for representing and reasoning with probabilities and independencies. In order to introduce the graphical framework a number of definitions are needed.

A graphical model can be used to represent a set of probability distributions that satisfy the implied constraints. Graphical models give equivalent probability models if their corresponding sets of satisfying probability distributions are equivalent.

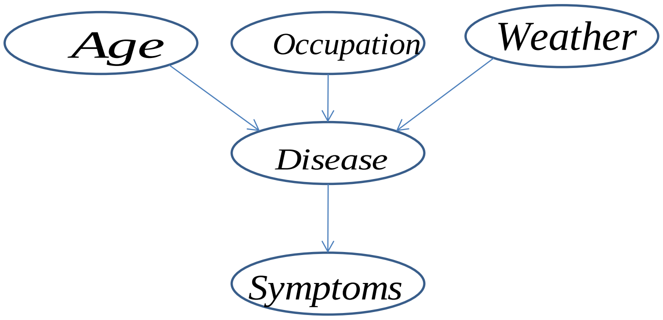


Figure 1: *Simplified Medical Model*

In Figure 1 an example of a graphical model of a simplified medical problem is given [2]. The variables Age, Occupation and Climate are variables that cause Disease. Disease on the other hand causes the Symptoms. This network model (factorisation) can be represented as:

$$P(Age, Occupation, Weather, Disease, Symptoms) =$$
$$P(Symptoms|Disease)P(Disease|Age, Occupation, Weather)P(Age)P(Occupation)P(Symptoms),$$

lets call this breakdown a factorisation of the joint probability distribution. The directed arcs show the causality.

2.1 Bayesian Belief Network

Definition of a Bayesian Belief Network (BBN) by Becker and Geiger [1]:

> *Let $P(i_1, i_2, \ldots, i_n)$ be a probability distribution where each variable i draws values from a finite set called the domain, $Dom()$, of i. A directed graph with no directed cycles is called a Bayesian Belief Network of P if there is a one to one mapping*

between $\{i_1, i_2, \ldots, i_n\}$ and variables in Dom, such that i and $P(i_1, i_2, \ldots, i_n) = \prod P(i|i_1, i_2, \ldots, i_{j(i)}$, where $i_1, i_2, \ldots, i_{j(i)}$ directed arcs to variable i in Dom. $j(i)$ indicates that the number of arcs directed to i is a function of i.

Note: A directed graph does not have to be a tree graph, but can have connections as shown in Figure 2 for example. Figure 1 is a BBN that is a tree model, where there are three arcs from variables directed to the variable *disease.*

2.2 Causality

In $P(A|B)$, there is a directed arc from B to A, and for the purposes of this paper B is considered the cause of A. Changes in the value(s) of a variable(s) (independent or causal variable(s)) can be the cause of changes in other variables (dependent variables), for instance, rainfall causes rivers to rise. This causation can be direct or indirect. Direct means that there are no intermediate variables between the causal and dependent variable, while indirect means that the causal variable first causes change in the value of some intermediate variable. The change in the intermediate variable causes the change in the dependent variable.

3 Cooper and Herskovits equation

The Cooper and Herskovits formula was designed to calculate the posterior probability model.

To learn about the model and probabilities of Bayesian belief Networks the posterior probabilities of different models can be compared [5]. Given observations, called data D, of the set of variables $\boldsymbol{X}$, representing the variables, the posterior probability of each model can be calculated. The uncertainty about the network model (can be seen in the joint probability distribution in Figure 1) the uncertainty can be encoded by defining a discrete variable S^h, whose states correspond to the possible network model and assessed by the probabilities $P(S^h)$.

To determine the posterior probability distribution for the network models, it is necessary to compute the unscaled marginal posterior probability of the data $P(D|S^h)$ for each possible model, as well as the normalizing constant $P(D)$.

When finding the most probable belief model S^h, from a database D, it is necessary to maximize $p(S^h|D)$. The number of possible models is super-exponential in the number of variables. There is no clear way of effectively pruning the search space as synergistic effects might occur between any set of variables.

Cooper and Herskovits [4] derived a formula to compare two models:

If $q = \sum\limits_{i=1}^{n} q_i$, and q_i is the number of states of parent i, [4]

$p(S^h, D) = p(S^h) \prod_{i=1}^{n} \prod_{j=1}^{q(i)} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!$. We can then select the largest of these (likelihood ratio test) to find the most likely model. If we assume that the prior probability $p(S^h)$ for all models to be the same and using the fact that $p(S^h|D) = \frac{p(S^h,D)}{p(D)}$. Heckerman then looked at this prior as frequencies α_{ijk} for each of the cells (Dirichlet distribution).

Hence, we arrive at the format of the Cooper and Herskovits equation used by Heckerman [5].

$$p(D|S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk}}{\Gamma(\alpha_{ijk})}.$$

Where S^h indicates the model h and k are the indices for the states of x_i. α_{ij} is the prior probability for each cell. N_{ijk} is the observed frequency of variable x_i in state k, given that the states of the parents are in combination j. r_i is the total number of states for variable i.

4 Best posterior probability does not mean casual model

To show that the model with the best posterior probability is not necessarily the causal model, I started with a model, thereafter expanded the model into a frequency matrix. An exhaustive search was carried out and the model with the best posterior probability was selected. In this example the best posterior probability model was not the model that created the data.

An exhaustive search is not feasible for large networks because the number of models are $3^{\frac{n(n-1)}{2}}$, where n is the number of variables in the model. The base is three because there can be an arc in either direction between any two variables, or the arc can be absent. The number of models is less if cyclical and infeasible models are eliminated.

I wrote a program to calculate the log posterior probability. To make sure that the program gives the correct answer, the program was run on the Sewell and Shah data of Wisconsin High School [8] and the program gave the same result as the result that Heckerman [5] found.

An example to show that the best posterior probability model is not the causal model:

I took $p(i = A, i = B, i = C, i = D, i = E) = p(A)p(B|A)p(C|B)p(D|B,C)p(E|D)$ and each variable i, with two states 0 and 1, with probabilities as indicated in Figure 2.

A prior frequency of 5 and divide the frequency over the cells was taken (Heckerman used a prior frequency of 5). Therefore $\sum_k \alpha_{ik} = 5$, for all i, and equal for each state of i.

In Figure 2 a graphical presentation of a network model is given with the probabilities indicated on the arcs. This is then used to calculate the joint probability matrix that results from this network.

The joint probability matrix calculated from the data is Figure 2 is:

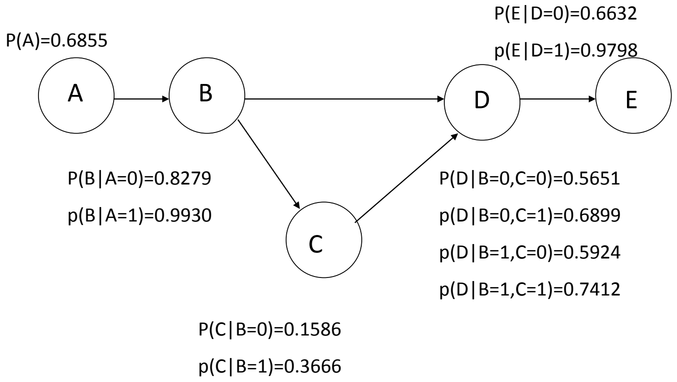


Figure 2: *The model selected*

$$P(0,0,0,0,0) = 0.0067, P(0,0,0,0,1) = 0.0131, P(0,0,0,1,0) = 0.0005, P(0,0,0,1,1) = 0.0252 \tag{1}$$

$$P(0,0,1,0,0) = 0.0009, P(0,0,1,0,1) = 0.0018, P(0,0,1,1,0) = 0.0001, P(0,0,1,1,1) = 0.0058 \tag{2}$$

$$P(0,1,0,0,0) = 0.0226, P(0,1,0,0,1) = 0.0446, P(0,1,0,1,0) = 0.0020, P(0,1,0,1,1) = 0.0957 \tag{3}$$

$$P(0,1,1,0,0) = 0.0083, P(0,1,1,0,1) = 0.0164, P(0,1,1,1,0) = 0.0014, P(0,1,1,1,1) = 0.0693 \tag{4}$$

$$P(1,0,0,0,0) = 0.0006, P(1,0,0,0,1) = 0.0012, P(1,0,0,1,0) = 0.0000, P(1,0,0,1,1) = 0.0022 \tag{5}$$

$$P(1,0,1,0,0) = 0.0001, P(1,0,1,0,1) = 0.0002, P(1,0,1,1,0) = 0.0000, P(1,0,1,1,1) = 0.0005 \tag{6}$$

$$P(1,1,0,0,0) = 0.0592, P(1,1,0,0,1) = 0.1165, P(1,1,0,1,0) = 0.0052, P(1,1,0,1,1) = 0.2503 \tag{7}$$

$$P(1,1,1,0,0) = 0.0218, P(1,1,1,0,1) = 0.0428, P(1,1,1,1,0) = 0.0037, P(1,1,1,1,1) = 0.1812 \tag{8}$$

This matrix was then multiplied by 10 000 to get a frequency matrix, and the prior frequencies added to get a frequency matrix that can be used to find the log of the posterior probability.

The log posterior probability for the correct model (the model in Figure 2) is -23750. The models with the best log posterior probability of -23755 are A→B→C→D→E, A←B→C→D→E, A←B←C→D→E, A←B←C←D→E and A←B←C←D←E. This gives a posterior probability for models that is e^5 times larger than the posterior probability of the correct model (the correct model is not a tree model and is given in Figure 2). These are the only models that have a better posterior probability than the correct model. I started with a prior frequency of 5 distributed over all the cells in the probability matrix, but changing this value did not make any significant changes to the results (did not change the selected model). I found a number of multiply connected models where the Cooper and Herskovits did not find the correct causal model.

5 The elimination heuristic

As an alternative to the Pearl [7] and the Cooper and Herskovits [4] algorithms, that start building the network model by adding arcs to an unconnected network, I present the Elimination heuristic that starts with a fully connected network and eliminates the arcs that are redundant. This produces the same model for the tree models that Pearl investigated. A norm is also introduced to formalize the elimination decisions.

5.1 Investigating the path between two variables

The following definition by Castillo et al. [3] defines the path between two variables:

Definition 5.1: *A path from variable A to variable C is an ordered set of variables $i_1, \ldots, i_m$ starting in $i_1 = A$ and ending in $i_m = C$ such that there is a link from i_k to $i_{k+1}, k = 1, \ldots, m-1$, that is, $i_{k+1} \in Adj(i_k), k = 1, \ldots m-1$, where Adj is the set of variables with links to i_k. The length of the path is $(m-1)$, the number of links it contains.*

A directed path is indicated with arrows, for instance, the path A→B→C, and an undirected path as A-B-C. If the intermediate variables between A and C are not known, the path is still indicated as A-C, meaning that it is the path from A to C, with the possibility that there might be intermediate variables. Because it is not known beforehand whether there are intermediate variables, it is difficult to use a different notation. The intermediate variables are often only discovered at a later stage.

When the actual causal model is a tree model, Pearl [7] uses $\rho_{\text{AB}}\rho_{\text{BC}} = \rho_{\text{AC}}$ to identify variables to connect to the tree model, from this I define the norm $\frac{\rho_{\text{AB}}\rho_{\text{BC}}}{\rho_{\text{AC}}}$ which is a more convenient format. The norm equals one when the Markov Condition holds, i.e. A is independent of C given B. The star decomposition restrictions hold as is expected of a Bayesian belief network model with a tree model.

The norm of the path A-B-C will be one if and only if all paths connecting A and C have to go through B. Only one path means that A-B-C forms a tree model. In a tree model, if the norm is one, then there is one and only one path connecting A and C.

The Pearl algorithm specifies which paths must be in the network and, by systematically adding variables, the algorithm reconstructs the network. For small networks the model can be identified by inspection from the star decomposition restrictions (the paths identified), but for larger models the more formal Pearl algorithm must be used. In the next section an alternative is given identify the model.

5.2 The elimination heuristic

Castillo et al. [3] defines conditional independence as follows:

Definition 5.1 Conditional dependence and independence [3]: *"Let X, Y and Z be three disjoint sets of variables, then X is said to be conditionally independent of Y given Z, if and only if $P(x|z,y) = P(x|z)$, for all possible values x, y and z of X, Y and Z; otherwise X and Y are said to be conditionally dependent given Z. When X and Y are conditionally independent given Z we write $I(X,Y|Z)$. The statement $I(X,Y|Z)$ is referred to as a conditional independence statement (CIS)."*

Given variables A, B and C and assume only linear relationships, then the norm of the path(A-B-C) = 1 if and only if A and C are conditionally independent given B or I(A, C| B)

Proof:

If the norm of the path A-B-C equals one, then $\frac{\rho_{\mathrm{AB}}\rho_{\mathrm{BC}}}{\rho_{\mathrm{AC}}} = 1$ or $\rho_{\mathrm{AB}}\rho_{\mathrm{BC}} = \rho_{\mathrm{AC}}$.

Because only linear relationships are considered, the correlation between A and C given B is: $\rho_{\mathrm{AC}.B} = \frac{\rho_{\mathrm{AC}}-\rho_{\mathrm{AB}}\rho_{\mathrm{BC}}}{\sqrt{(1-\rho_{\mathrm{AB}}^2)(1-\rho_{\mathrm{BC}}^2)}} = 0$ [6] implies that A and C are conditionally independent given B or I(A,C|B).

The converse: If A and C are conditionally independent given B, I(A, C | B) then $\rho_{\mathrm{AB}}\rho_{\mathrm{BC}} = \rho_{\mathrm{AC}}$.

The implication of Lemma 5.1 is that if a path A-B-C is identified as a valid path, then the arc A-C can be eliminated, because the conditional independence statement I(A, C | B) has been identified. It is possible to return to the initial topology of the Bayesian belief network model, with undirected arcs from the theoretical frequencies.

6 Conclusions

The Bayesian belief Network model with the highest posterior probability is not necessarily the causal model. The best posterior probability models could have posterior probabilities that are higher than the correct causal model. This makes a different approach to finding the correct causal model important.

To summarise the elimination heuristic:

1. Start with a complete graph containing all the variables.
2. Eliminate arcs between variables that are independent (with a correlation coefficient not significantly different from zero).
3. Eliminate the arcs where Conditional Independence Statements have been identified (where the norm equals one).

This gives us the causal tree model, and I tested that it is also the model with the highest posterior probability.

For further research, researchers should distinguish between the causal model and the model with the highest posterior probability. Algorithms should be developed to identify the causal model from data. I have already identified and proven a number of theorems based on this norm, which will be included in later papers. At this stage there are still too many unproven theorems to find the actual causal model.

Bibliography

[1] Becker, A. and Geiger, D. (1996). Pearls method of conditioning and greedy-like approximation algorithms for the vertex feedback set problem. *Artificial Intelligence*, 83:167–188.

[2] Buntine, W. (1994). Graphical methods for discovering knowledge. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

[3] Castillo, E., Guitirrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Springer.

[4] Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

[5] Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1:79–119.

[6] Morrison, D. (1990). *Multivariate statistical methods*. London, McGraw Hill.

[7] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrica*, 82:699–710.

[8] Sewell, W. and Shah, V. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 73:59–572.

Proceedings of the 2011 ORSSA Annual Conference
pp. 19–28
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Constructive heuristics for the Residential Waste Collection Problem

EJ Willemse[*] JW joubert[†]

Abstract

The Residential Waste Collection Problem (RWCP) is a realistic extension of the classical Capacitated Arc Routing Problem (CARP), with application in municipal waste collection. Surprisingly, the problem with its extensions have not been solved in literature. This paper presents two heuristics that are capable of solving the RWCP. The heuristics are based on modifications of the classical Path-Scanning and Augment-Merge heuristics for the CARP. The modified heuristics are tested on new benchmark problems for the RWCP, and results show that the algorithms are capable of quickly solving the problem.

Key words: Capacitated Arc Routing Probelm, Intermediate Facilities, Mixed Network, Constructive heuristics

1 Introduction

Solid waste collection and transportation consists of the collection, transportation and disposal of waste at landfill sites, usually through waste collection vehicles. It is well recognised as being the most costly component of the waste management function and can account for between 50-80% of a municipality's solid waste management budget [14, 21, 22]. Of the different types of waste collected by municipalities, residential waste forms the biggest percentage [15]. As such it consumes the largest part of the municipal budget, making it a promising area to target for cost reductions.

The state of practice method of residential waste collection is through curb side collection. Municipalities have to collect the waste of each household at least once a week. Households place their generated waste, which are stored in either bins or bags, on the designated days in front of their properties where waste collection vehicles can then collect the waste. This process is highly repetitive and performed throughout the year, therefore even a small improvement in waste collection and transfer operations can lead to significant savings in costs.

[*] Corresponding author: CSIR Built Environment, South Africa, email: ejwillemse@gmail.com

[†] University of Pretoria, South Africa, email: johan.joubert@up.ac.za

A promising improvement area is to design better waste collection routes. In this paper we show that designing collection routes can be modelled as an extension of the classical Capacitated Arc Routing Problem (CARP). Two constructive heuristics that are capable of solving the problem are developed and tested on benchmark problems, created to mimic waste collection in residential areas.

The remainder of this paper is organised as follows. The next section gives a brief overview of previous work, related to the RWCP. The two heuristics are presented in Section 3 with computational results provided in Section 4. The paper concludes with Section 5 in which we summarize our main findings.

2 Related work

Residential waste collection requires waste to be collected on a street-by-street basis. As such, the problem of designing collection routes can be modelled as an Arc Routing Problem (ARP). Its aim is defined by Eiselt et al. [6] as determining a least-cost traversal of a specified subset of a graph, with or without constraints. Other ARP application areas include, for example, winter gritting [5, 17], postal delivery [13], security guard routing [23], street sweeping [2] and railway maintenance [12]. For a comprehensive review of ARPs the reader is referred to [3, 4, 6, 7].

As most practical routing applications contain capacity restrictions, the Capacitated Arc Routing Problem, first proposed by Golden and Wong [11], is probably the most important problem in the area of arc routing [7]. With the CARP a fleet of homogeneous vehicles are based at a depot and are tasked with serving all the street segments with waste. The problem consists of designing vehicle routes for the fleet of total minimal length so that each route starts and ends at the depot, each road segment with demand is serviced exactly once by a single vehicle, and the sum of demand on any route does not exceed vehicle capacity.

Two CARP extensions inspired by waste collection have been proposed in literature. The Mixed Capacitated Arc Routing Problem, studied by Lacomme et al. [16], Mourão and Amado [18], Mourão et al. [19] and Belenguer et al. [1], allows the modelling of more realistic street networks, containing one and two way streets. The extension further allows for streets that require each side to be serviced separately. The second extension, referred to as the Arc Routing Problem with Intermediate Facilities under Capacity and Length Restrictions (CLARPIF), accounts for intermediate facilities and dumpsites. The extension, first proposed by Ghiani et al. [8], allows for the collection vehicle to unload its cargo at a nearest Intermediate Facility (IF), including dumpsites, and then resume its collection route. A length (or cost) restriction, independent from capacity and typically presenting the number of working hours in a day, is also placed on each vehicle's route.

Both extensions contain elements fundamental to residential waste collection. In response we combine the CARP and CLARPIF into a new problem termed the Residential Waste Collection Problem (RWCP). The RWCP entails designing vehicle routes of total minimum length, so that each route starts and ends at the depot, each road segment with a demand (waste) is serviced exactly once by a vehicle, and the total cost of a vehicle route does not exceed a maximum allowed trip length or cost. Each route contains visits to IFs, which may or may not include the depot, and the sum of demand on the sub-route between dumpsite visits does not exceed vehicle capacity. Lastly, at the end of each route each vehicle must visit a dumpsite before returning to the depot.

To our knowledge the RWCP is new and has not been formally studied or solved in literaure. Various solution approaches have, however, been developed for the MCARP and CLARPIF. Since both problems, and by extension the RWCP, are $\mathscr{N}\mathscr{P}$-hard, the most effective methods for the solving the problems are based on heuristic techniques. The main studies on the problems are by Belenguer et al. [1], Lacomme et al. [16], Mourão and Amado [18] and Mourão et al. [19], all whom develop heuristics for the MCARP; and the works of Ghiani et al. [8], Ghiani et al. [9], and Polacek et al. [20], whom develop heuristics for the CARP with IFs. In all the studies on the MCARP, except the work of Mourão and Amado [18], one of two classical CARP heuristics are modified, namely Path-Scanning and Augment-Merge. The focus of this paper is to further modify the two heuristics to deal with IFs, thus making them capable of solving the RWCP.

3 Constructive heuristics for the RWCP

Formally defined, the RWCP consists of a mixed graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E} \cup \boldsymbol{A})$ where $\boldsymbol{V}$ represents the set of vertices, $\boldsymbol{E}$ represents the set of edges and $\boldsymbol{A}$ represents the set of arcs. A subset of required edges and arcs, $\boldsymbol{E}_r \subseteq \boldsymbol{E}$ and $\boldsymbol{A}_r \subseteq \boldsymbol{A}$, must be serviced by a fleet of K homogenous vehicles with limited capacity Q that are based at the depot vertex v_1. The fleet size, K, may be fixed or left as a decision variable. The vehicles are allowed to unload their waste at any IF at a cost of λ. The set of IFs are modelled by $\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} \subset \boldsymbol{V}$. Unless $v_1 \in \boldsymbol{\Gamma}$, a vehicle has to visit an IF before returning to the depot.

Before presenting the modified heuristics for the RWCP, we first show how the mixed graph $\boldsymbol{G}$ can be transformed into a directed graph and we introduce the algorithm encoding scheme used in the remainder of this section.

3.1 Graph transformation and encoding scheme

Consistent with the work of Belenguer et al. [1] and Lacomme et al. [16], the mixed graph $\boldsymbol{G}$ is transformed into a fully directed graph $\boldsymbol{G}^* = (\boldsymbol{V}, \boldsymbol{A}^*)$ by replacing each edge $(v_i, v_j) \in \boldsymbol{E}$ by two opposite arcs $\{(v_i, v_j), (v_j, v_i)\} \in \boldsymbol{A}^*$. Arcs in $\boldsymbol{A}^*$ are indentified by indices from 1 to m where $m = |\boldsymbol{A}^*|$. Each arc $u \in \boldsymbol{A}^*$ has a beginning vertex $b(u)$ and an end vertex $e(u)$. If arc u represents (v_i, v_j) then $b(u) = v_i$ and $e(u) = v_j$. Lastly, each arc u has a deadheading cost $c(u)$. Lastly, the total cost of any route must not exceed the maximum allowed trip cost, L.

The required arcs, $\boldsymbol{A}_r$, and edges, $\boldsymbol{E}_r$, of $\boldsymbol{G}$ correspond in $\boldsymbol{G}^*$ to a subset $\boldsymbol{R} \subseteq \boldsymbol{A}^*$ of required arcs, such that $|\boldsymbol{R}| = 2|\boldsymbol{E}_r| + |\boldsymbol{A}_r|$. Each arc $u \in \boldsymbol{R}$ has a demand $q(u)$, a collection cost $w(u)$ and a pointer $inv(u)$. Each required arc in the original graph $\boldsymbol{G}$ is coded in $\boldsymbol{R}$ by one arc u with $inv(u) = 0$, while each required edge is encoded as two opposite arcs u and v, such that $inv(u) = v$, $inv(v) = u$, $q(u) = q(v)$, $c(u) = c(v)$ and $w(u) = w(v)$. An arc task u represents an edge if $inv(u) \neq 0$. The depot is modelled by including in $\boldsymbol{A}^*$ a fictitious loop $\sigma = (v_1, v_1)$, with $b(\sigma) = e(\sigma) = v_1$, $inv(\sigma) = 0$ and $q(\sigma) = w(\sigma) = c(\sigma) = 0$. Similarly, the set of IFs are modelled in $\boldsymbol{A}^*$ as a set of dummy arcs $\boldsymbol{I}$, such that each IF in $\boldsymbol{\Gamma}$ is modelled as a fictitious loop $\Phi_i \in \boldsymbol{I}$, and Φ_i has the same start and end vertex, and zero cost and demand.

For the RWCP a solution $\boldsymbol{T}$ is a list $(\boldsymbol{T}_1, \dots, \boldsymbol{T}_K)$ of K vehicle trips. Each trip, $\boldsymbol{T}_i$, is a list of subtrips $(\boldsymbol{T}_{i1}, \boldsymbol{T}_{i2}, \dots, \boldsymbol{T}_{i,|\boldsymbol{T}_i|})$ which then consists of a list of tasks $(\boldsymbol{T}_{ij1}, \boldsymbol{T}_{ij2}, \dots, \boldsymbol{T}_{ij,|\boldsymbol{T}_{ij}|})$. The first subtrip, $\boldsymbol{T}_{i1}$, starts at the depot, and the last subtrip, $\boldsymbol{T}_{i,|\boldsymbol{T}_i|}$, ends with an intermediate facility and depot visit. All other subtrips start and end with IF visits whilst taking care that the starting IF of a subtrip coincides with the end IF of the previous subtrip. It is assumed that the shortest path, which can be efficiently calculated using a modified version of as Dijkstras' algorithm [16], is always followed between consecutive tasks. Lastly, denote by $\boldsymbol{D}(u, v)$ the cost of the shortest path from arc u to arc v, excluding the costs of deadheading u and v.

The best IF to visit after servicing arc u and before servicing arc v can be easily pre-calculated, as shown by Ghiani et al. [8]. The best IF to visit is given by $\mathit{dump}(u, v)$ and the cost of the visit, including unload and deadheading costs, is given by $\mathit{term}(u, v)$. Lastly we define $\mathit{load}(\boldsymbol{T}_i)$ as the total demand of trip $\boldsymbol{T}_i$, and $\mathit{cost}(\boldsymbol{T}_i)$ as the total cost of the trip. The same terms are also used to define the demand and cost of subtrips.

3.2 Path-Scanning

The first heuristic that we modify for the RWCP is the Path-Scanning heuristic developed by Golden et al. [10]. Our modification is based on the algorithm of Belenguer et al. [1] for the MCARP. Path-Scanning systematically builds a vehicle trip by adding the closest unserviced arc to the end of the trip. If there are multiple closest arcs one of seven rules (Table 1) is used to break the tie. The algorithm adds the closest unserviced arc to a vehicle trip until the vehicle is full, at which point the trip is closed by adding a depot arc visit. A new empty trip is created and the process repeats until all required arcs are serviced.

Table 1: *Tie-break rules used with the* Path-Scanning *heuristic to choose arc $u \in \boldsymbol{A}_c$, where $\boldsymbol{A}_c$ is the set of closest arcs.*

Rule	Description
1	Maximise the distance to the depot; $\max\{\boldsymbol{D}(u, \sigma) : u \in \boldsymbol{A}_c\}$
2	Minimise the distance to the depot; $\min\{\boldsymbol{D}(u, \sigma) : u \in \boldsymbol{A}_c\}$
3	Maximise the arc yield; $\max\{q(u)/w(u) : u \in \boldsymbol{A}_c\}$
4	Minimise the arc yield; $\min\{q(u)/w(u) : u \in \boldsymbol{A}_c\}$
5	Use Rule 1 if the vehicle is less than half-full, else use Rule 2
6	Randomly use any of the five rules
7	Do not use any rule and randomly choose u from $\boldsymbol{A}_c$

Our RWCP version of Path-Scanning sees the introduction of subtrips resulting from IF visits. Instead of starting with a trip, the algorithm starts with a subtrip. The first subtrip starts at the depot and the algorithm adds the closest unserviced arc to the end of the subtrip. When the subtrip is full the algorithm closes it by adding an IF visit, and a new subtrip is then created starting at this IF. The process repeats until no unserviced arcs can be added without exceeding the maximum allowed trip length. At this point the subtrip is closed by adding a visit to an IF and a depot. The combined subtrips now form a single vehicle trip whose cost does not exceed the maximum allowed trip length. The process is repeated until all arcs are serviced. Algorithm 1 summarises our RWCP version of the PATH-SCANNING algorithm.

Algorithm 1: PATH-SCANNING for the MCARP

Input : The directed transformed graph $\boldsymbol{G}^*$.
Output: A feasible solution $\boldsymbol{T}$ for the RWCP.

Step 0: Set $\boldsymbol{G}' \leftarrow \boldsymbol{G}^*$, $i \leftarrow 1$ and $j \leftarrow 1$.

Step 1: Let the first task in $\boldsymbol{T}_{i1}$ be the depot arc σ and set $\boldsymbol{G}'' \leftarrow \boldsymbol{G}'$.

Step 2: Let u be the last task in subtrip $\boldsymbol{T}_{ij}$. Remove from $\boldsymbol{G}''$ all arcs v with $costIF(\boldsymbol{T}_i) + \boldsymbol{D}(u,v) + term(v,\sigma) - \lambda > L$. If $\boldsymbol{G}''$ is empty, close $\boldsymbol{T}_{ij}$ by adding $dump(u,\sigma)$ and σ to the end of the trip; let $i \leftarrow i+1$ and return to Step 1. If $\boldsymbol{G}''$ is not empty go to Step 3.

Step 3: Set $\boldsymbol{G}''' \leftarrow \boldsymbol{G}''$ and remove from $\boldsymbol{G}'''$ all arcs v with $load(\boldsymbol{T}_{ij}) + q(v) > Q$. If $\boldsymbol{G}'''$ is empty find the arc v in $\boldsymbol{G}'$ that minimises $term(u,v)$; close the subtrip by adding a visit to $dump(u,v)$ to the end of the subtrip; set $j \leftarrow j+1$; create a new subtrip $\boldsymbol{T}_{ij}$ that starts at $dump(u,v)$ and return to Step 2. If $\boldsymbol{G}'''$ is not empty go to Step 4.

Step 4: From the last subtrip task u, find the closest arc v in $\boldsymbol{G}'''$. If there are more than one closest arc, choose an arc according to a tie-break rule. Add the closest chosen arc v to the end of $\boldsymbol{T}_{ij}$ and remove v and $inv(v)$, if it exists, from $\boldsymbol{G}'$ and $\boldsymbol{G}''$. If all edges in $\boldsymbol{G}$ are covered, thus $\boldsymbol{G}'$ is empty, go to Step 5, else return to Step 2.

Step 5: Close $\boldsymbol{T}_{ij}$ with a trip to $dump(v,\sigma)$ and the σ and stop the algorithm. The solution $\boldsymbol{T}$ is a feasible solution for the RWCP.

Traditionally the PATH-SCANNING algorithm is executed five times using one of Rules 1-5 (Table 1) in each execution to break closest arc ties. The best of the five solutions is then chosen for implementation. A similar approach can be followed using Rule 6 or 7 of the table with the advantage that an arbitrary number of solutions can be generated, owing to randomness of the rules. We test both the deterministic scheme (Rules 1 to 5) and random schemes (Rules 6 and 7) on the RWCP. The first scheme is the default scheme and is just referred to as PATH-SCANNING, or PS for short. The second and third schemes, using Rule 6 and 7, are referred to as PS RANDOM RULE 200 and PS RANDOM ARC 200, respectively, where 200 refers to the number of solutions generated.

3.3 Augment-Merge

The second algorithm that we modify for the RWCP is the Merge algorithm of Belenguer et al. [1] for the MCARP. The algorithm starts by creating a solution that contains a vehicle trip for each required arc. The trips are then systematically reduced through mergers. Initially trips containing only one arc task are merged, but as the process continues, trips with multiple tasks start to develop and these are also merged. As the algorithm progresses the trips become fuller. The algorithm stops when all the trips are full, or near full, and no more mergers are possible.

The merger of two trips, $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$, is performed through the Merge-Trips-Procedure. Before merging the trips the procedure first checks that the combined load of the two trips does not exceed vehicle capacity. The procedure then merges the trips by splicing them together. The merge procedure also considers adding $\boldsymbol{T}_i$ to the end of $\boldsymbol{T}_j$. The procedure also considers the reversal and then merger of trips, resulting in an additional six possible mergers. Merge-Trips-Procedure evaluates all eight possible mergers between $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ and returns the best merge as $\tilde{\boldsymbol{S}}_{ij}$, and the best saving as $\Delta\tilde{S}_{ij}$.

The Merge algorithm performs one merger per iteration. During each iteration the algorithm evaluates all feasible trip mergers and the one with the best saving is implemented. The two original trips are replaced with the merged trip and the algorithm proceeds to the next iteration. This process is repeated until no more mergers are possible. Our implementation of Merge for the MCARP is described in Algorithm 2.

```
Algorithm 2: MERGE

Input  : The directed transformed graph $\boldsymbol{G}^*$
Output: A feasible solution $\boldsymbol{T}$ for the MCARP.

Step 0: Set $\boldsymbol{R}' \leftarrow \boldsymbol{R}$. For each arc v $\in \boldsymbol{R}'$ create a vehicle trip servicing only that arc. If
        $inv(v) \neq 0$ then remove $inv(v)$ from $\boldsymbol{R}'$, thus $inv(v)$ will not be assigned to a trip
        in a latter iteration.

Step 1: For the unique trips $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ in $\boldsymbol{T}$ check that their merger does not violate
        the capacity constraint, i.e., $load(\boldsymbol{T}_i) + load(\boldsymbol{T}_j) \leq Q$. If the merger is possible
        then determine the best merger $\tilde{\boldsymbol{S}}_{ij}$ and cost saving $\Delta\tilde{S}_{ij}$ through MERGE-TRIP-
        PROCEDURE. If no mergers are possible then go to Step 3, else go to Step 2.

Step 2: Using $\min \Delta\boldsymbol{S}_{ij}$, find the best merge $\tilde{\boldsymbol{S}}_{kl}$ and replace trips $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ in $\boldsymbol{T}$ with
        $\tilde{\boldsymbol{S}}_{kl}$. If $\boldsymbol{T}$ now consists of only one trip then stop, else return to Step 1.

Step 3: Since no more mergers are possible stop the algorithm and return $\boldsymbol{T}$ as a solution
        for the MCARP.
```

To apply Merge to the RWCP we execute Merge twice, with minor modifications to the algorithm in each execution. During the first execution each vehicle trip consists of only one vehicle subtrip, with an IF visit before returning to the depot. A merger between two trips is then performed as with the MCARP. Let $\boldsymbol{T}_i = (\sigma, a_1, a_2, \ldots, a_m, \Phi_i, \sigma)$ and let $\boldsymbol{T}_j = (\sigma, b_1, b_2, \ldots, b_n, \Phi_j, \sigma)$, where Φ_i and Φ_j are calculated as $\Phi_i = \mathit{dump}(a_m, \sigma)$ and $\Phi_j = \mathit{dump}(b_n, \sigma)$. A new merged trip $\boldsymbol{S}_{ij}$ is created such that

$$\boldsymbol{S}_{ij} = (\sigma, a_1, a_2, \ldots, a_m, b_1, b_2, \ldots, b_n, \Phi_j, \sigma).$$

Before merging the trips the procedure first checks that the combined load of the two trips does not exceed vehicle capacity. The algorithm evaluates all feasible trip mergers and the one with the best saving is implemented. As with the MCARP the algorithm checks all eight possible merge orientations. The first Merge execution terminates when no more mergers are possible without violating the vehicle capacity restrictions.

The algorithm then proceeds with the second Merge execution in which vehicle subtrips are merged into vehicle trips. Again let $\boldsymbol{T}_i = (\sigma, a_1, a_2, \ldots, a_m, \Phi_i, \sigma)$ and let $\boldsymbol{T}_j = (\sigma, b_1, b_2, \ldots, b_n, \Phi_j, \sigma)$. A new merged trip, $\boldsymbol{S}_{ij}$, is now created such that

$$\boldsymbol{S}_{ij} = \big((\sigma, a_1, a_2, \ldots, a_m, \Phi_s), (\Phi_s, b_1, b_2, \ldots, b_n, \Phi_j, \sigma)\big).$$

Before merging the trips the procedure checks that the combined cost of the two trips does not exceed the maximum trip length restriction. The algorithm evaluates all feasible trip mergers and the one with the best saving is implemented. The algorithm again checks all eight possible merge orientations. The second Merge execution terminates when no more mergers are possible without violating the maximum trip length restriction. Since the general structure of the two executions of Merge for the RWCP is the same as with the MCARP, we do not give a formal description of our implementation.

4 Computational results

To evaluate and compare the heuristics, both are tested on a new set of benchmark problems. The new problems were created by extending the *lpr* problem set of Belenguer et al. [1] by including IFs at vertices $\lfloor |\boldsymbol{V}|/2 \rfloor$ and $2\lfloor |\boldsymbol{V}|/2 \rfloor$. For each problem we introduce a max trip length of 28 800 seconds, which corresponds to an eight hour working day, and the vehicle capacity remains fixed at 10 000kg. Lastly, the number of vehicles required to service the area is not fixed but left as a decision variable. The new RWCP benchmark problems are referred to as the *lpr-IF* problem set.

The heuristics are evaluated on three criteria: total fleet travel time minimisation, vehicle fleet size minimisation and running time. All the algorithms were coded in Python version 2.6 and run on a 3Ghz Intel(R) Core(TM)2 Duo CPU with 3.25GB of RAM.

The results of the three Path Scanning heuristics and the Merge heuristics on the *lpr-IF* problem set are shown in Table 2. The table shows the total cost of each solution generated, the number of vehicles that the solution requires and the total time required to generate the solution. Results show the random versions of Path Scanning to produce slightly better solutions than the deterministic versions, which is expected since they generate more solutions from which to choose the best. The results further show that the Path-Scanning heuristics outperforms Merge on all but four of the problems. Merge faired particularly poorly in minimising the vehicle fleet size and required an extra route with eight of the fifteen solutions.

In terms of the computational time, Path-Scanning significantly outperformed Merge on all the problem instances. On all the *lpr-IF* problems the Path-Scanning variants are capable of generating up to 200 solutions in less than one minute, making them very efficient, even on large problems. Based on the tests for the *lpr-IF*, we conclude that Path-Scanning is better suited than Merge to solve the RWCP.

Table 2: *Computational results for Path-Scanning and Merge on the lpr-IF problem set, with the best solutions for each problem underlined. All time values are given in seconds.*

	PS			PS RA 200			PS RR 200			Merge		
File	*K*	Cost	Time	*K*	Cost	Time	*K*	Cost	Time	*K*	Cost	Time
a-01	1	13783	0.1	1	13678	0.3	1	13659	0.4	1	13954	0.9
a-02	2	29687	0.1	1	28647	1.1	1	28605	1.1	1	28791	6.9
a-03	3	80261	0.2	3	79545	9.0	3	79748	9.1	4	78695	173.4
a-04	5	134027	0.6	5	133434	26.2	5	133779	26.4	6	141678	802.8
a-05	8	213707	1.6	8	212208	66.2	8	212489	66.8	9	218128	3339.1
b-01	1	15049	0.1	1	14911	0.3	1	14868	0.3	1	15261	0.8
b-02	2	29955	0.1	2	29684	1.1	2	29715	1.1	2	29119	6.3
b-03	3	81708	0.2	3	80779	8.7	3	80912	8.7	4	80302	174.9
b-04	5	134580	0.6	5	133351	22.8	5	133567	23.0	6	136897	793.0
b-05	8	222874	1.5	8	221952	63.2	8	222502	62.5	9	228057	3227.5
c-01	1	18837	0.1	1	18783	0.5	1	18814	0.5	1	18973	0.9
c-02	2	36903	0.1	2	36796	1.6	2	36808	1.6	2	37345	8.2
c-03	4	114763	0.4	4	114593	16.3	4	114179	16.5	5	116422	219.4
c-04	7	177011	1.0	7	175819	41.2	7	176366	41.1	7	172506	916.4
c-05	10	276976	2.1	10	276876	89.2	10	276239	89.7	11	276447	3370.1

Acronyms: PS - Path Scanning, RR - Random Rule, RA - Random Arc, *K* - Number of routes

5 Conclusion

The RWCP is a new problem combining key elements of the MCARP and CLARPIF consistent with actual waste collection and transportation activities. Two heuristics were developed and tested to solve the RWCP, of which Path-Scanning performed the best on fifteen newly developed benchmark problems. The heuristic also proved efficient and is capable of generating multiple solutions within minutes on large problems. The developed constructive heuristics forms an important first step in studying the RWCP as its starting solutions can be used with developing and testing improvement heuristics and metaheuristics for RWCP.

Bibliography

[1] Belenguer, J.-M., Benavent, E., Lacomme, P., and Prins, C. (2006). Lower and upper bounds for the mixed capacitated arc routing problem. *Computers & Operations Research*, 33(12):3363–3383.

[2] Bodin, L. D. and Kursh, S. J. (1979). A detailed description of a computer system for the routing and scheduling of street sweepers. *Computers & Operations Research*, 6(4):181–198.

[3] Corbern, A. and Prins, C. (2010). Recent results on arc routing problems: An annotated bibliography. *Networks*, 56(1):50–69.

[4] Dror, M., editor (2000). *Arc Routing: Theory, Solutions, and Applications.* Boston: Kluwer Academic Publishers.

[5] Eglese, R. and Li, L. Y. O. (1992). Efficient routeing for winter gritting. *The Journal of the Operational Research Society*, 43(11):1031–1034.

[6] Eiselt, H. A., Gendreau, M., and Laporte, G. (1995a). Arc routing problems, part I: The Chinese Postman Problem. *Operations Research*, 43(2):231–242.

[7] Eiselt, H. A., Gendreau, M., and Laporte, G. (1995b). Arc routing problems, part II: The Rural Postman Problem. *Operations Research*, 43(3):399–414.

[8] Ghiani, G., Guerriero, F., Laporte, G., and Musmanno, R. (2004). Tabu search heuristics for the arc routing problem with intermediate facilities under capacity and length restrictions. *Journal of Mathematical Modelling and Algorithms*, 3(3):209–223.

[9] Ghiani, G., Improta, G., and Laporte, G. (2001). The capacitated arc routing problem with intermediate facilities. *Networks*, 37(3):134–143.

[10] Golden, B. L., DeArmon, J. S., and Baker, E. K. (1983). Computational experiments with algorithms for a class of routing problems. *Computers & Industrial Engineering*, 10(1):47–59.

[11] Golden, B. L. and Wong, R. T. (1981). Capacitated arc routing problems. *Networks*, 11(3):305–315.

[12] Groves, G., Le Roux, J., and Van Vuuren, J. (2004). Network service scheduling and routing. *International transaction in operational research*, 11(6):613–643.

[13] Irnich, S. (2008). Solution of real-world postman problems. *European Journal of Operational Research*, 190(1):52–67.

[14] Karadimas, N. V., Papatzelou, K., and Loumos, V. G. (2007). Optimal solid waste collection routes identified by the ant colony system algorithm. *Waste Management & Research*, 25(6):139–147.

[15] Korfmacher, K. S. (1997). Solid waste collection systems in developing urban areas of South Africa: an overview and case study. *Waste Management Research*, 15(5):477–494.

[16] Lacomme, P., Prins, C., and Ramdane-Chérif, W. (2004). Competitive memetic algorithms for arc routing problems. *Annals of Operations Research*, 131(4):159–185.

[17] Li, L. Y. O. and Eglese, R. W. (1996). An interactive algorithm for vehicle routeing for winter - gritting. *The Journal of the Operational Research Society*, 47(2):2.

[18] Mourão, M. C. and Amado, L. (2005). Heuristic method for a mixed capacitated arc routing problem: A refuse collection application. *European Journal of Operational Research*, 160(1):139–153.

[19] Mourão, M. C., Nunes, A. C., and Prins, C. (2009). Heuristic methods for the sectoring arc routing problem. *European Journal of Operational Research*, 196(3):856–868.

[20] Polacek, M., Doerner, Karl F. Hartl, R. F., and Maniezzo, V. (2008). A variable neighborhood search for the capacitated arc routing problem with intermediate facilities. *Journal of Heuristics*, 14(5):405–423.

[21] Tchobanoglous, G., Theisen, H., and Vigil, S. (1993). *Integrated solid waste management: Engineering principles and management issues.* McGraw-Hill New York.

[22] Viotti, P., Pelettini, A., Pomi, R., and Innocetti, C. (2003). Genetic algorithms as a promising tool for optimisation of the MSW collection routes. *Waste Management Research*, 21(4):292–298.

[23] Willemse, E. J. and Joubert, J. W. (In Press). Applying min-max k postmen problems to the routing of security guards. *Journal of the Operational Research Society.* Advance online publication, doi:10.1057/jors.2011.26.

Proceedings of the 2011 ORSSA Annual Conference
pp. 29–37
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

An econometric study of currency crisis in a hyperinflationary economy: a case study

C Sigauke*

Abstract

The paper discusses the modelling of currency crisis in a hyperinflationary economy with Zimbabwe as a case study for the period 1991 to 2004 using logistic regression. The robustness and resilience of binary choice models in predicting currency crisis in an unstable and hyperinflationary economic environment is tested. A comparative analysis is done with a probit model. The findings from using the logit and probit models are very similar and both statistically reliable. The results from this study show that lax monetary policy and deterioration in economic fundamentals can contribute to currency crisis; large devaluation can be perceived as currency crisis; macroeconomic and financial variables are important determinants of currency crisis. These results are useful in modelling currency crisis in an unstable and hyperinflationary economic environment.
JEL Classification Numbers: C23, C25, E40, E44, F30, F31

Key words: Logistic regression, currency crisis, emerging markets, exchange rates.

1 Introduction

Currency crisis also known as balance-of-payments crisis has been studied for over fifty years and to the best of our knowledge modelling of currency crisis in a hyperinflationary and unstable economic environment such as the one experienced by Zimbabwe between 2000 and 2008 has not been covered in literature. An updated review of literature on currency crisis can be found in [2, 3, 4, 10, 12, 14, 13]. First generation models of speculative attacks were initiated by [6]. These models stress the fact that excessive expansionary fiscal and monetary policies result in a persistent loss of international reserves, ultimately forcing the authorities to abandon a fixed exchange rate. This happens when a government is running

*Corresponding author: University of Limpopo, South Africa, email: caston.sigauke@ul.ac.za

an excessive budget deficit causing it to run short of liquid assets or foreign currency which it can sell to support its currency at the fixed rate [6]. The crisis occurs because a country finances its fiscal deficit by printing money to the extent that excessive credit growth leads to the eventual collapse of the fixed exchange-rate regime [12]. Second generation models were pioneered by Obstfeld [7], who argues that currency crisis can be self-fulfilling if agents expect the government to switch to an inflationary monetary policy in the presence of a speculative attack. An important distinction between the first and second generation models is that first generation models suggest that when cross-country currency ties are strengthened, exchange rates should be stabilized whilst second generation models suggest otherwise [5]. Third-generation models have explored how problems in the banking and financial system interact with currency crisis, and how the crisis can have real effects on the rest of the economy [2]. A detailed literature survey on first generation models is found in [1] and second generation models are reviewed in [5].

In this paper we develop a logistic regression model which considers how various macroeconomic data affect the probability of a currency crisis. A comparative analysis is done with a probit model. The study is based on data from Zimbabwe over the period 1991 to 2004 with an emphasis on variables which are consistent with literature on speculative attacks and currency crisis. Currency crisis during peacetime is usually caused by banking instability [4]. Financial disasters in the financial and banking sector in Zimbabwe in 2004 saw the collapse of nine financial institutions being placed under curatorship, five banks were eventually placed under liquidation during the same year and five Asset Management Companies had their licences cancelled. This was due to probably speculative attacks on the Zimbabwean currency by agents who perceived that the government was pursuing lax financial policies. This is consistent with first generation models of currency crisis. The financial markets and institutions in Zimbabwe include among others, the Reserve Bank of Zimbabwe (RBZ), commercial banks, merchant banks, discount houses, building societies, post office savings bank, insurance companies and the Zimbabwe Stock Exchange (ZSE). It should be noted that after 2004, it was no longer possible to develop meaningful financial models for the Zimbabwean economy due to the rate at which inflation was increasing. For instance during the period 2000 to 2006 year on year inflation surged from 55% in December 2000 to 1281% in December 2006 [11].

The rest of the paper is organized as follows. In Section 2 we discuss briefly the macroeconomic background in Zimbabwe for the sampling period 1991-2004. The logistic regression model used in this paper is discussed in Section 3. Empirical results are covered in Section 4 and Section 5 concludes.

2 Macroeconomic Background

This section outlines the developments in the economy during the sampling period, 1991–2004, using the following economic indicators: inflation, exchange rate, gross domestic product and budget deficit.

2.1 Inflation

One of the challenges to the Zimbabwean economy at the inception of the Economic Structural Adjustment Programme (ESAP) during the period 1990/1991 was high inflation. In this paper we use the consumer price index (CPI) as a proxy for inflation. Inflation tends to promote speculative rather than productive investment. Major drivers of inflation have been: lax policies, drought, money supply growth caused by monetary financing of the budget deficits and droughts, widening foreign exchange parallel market, high monetary growth, supply bottlenecks, escalation of equity and property market prices, discretionary pricing behaviour, etc. After 1997, inflation was on an upward path. Inflation reached a peak of 598.7% in 2003 before crushing to 132.8 % in 2004 [14]. However, this decline never lasted long and from 2005 inflation surged upwards. This was possibly due to lax policies. It was 585.8% in December 2005 and it continued with this upward trend until end of 2008[1] [9].

2.2 Exchange Rate

Exchange rate was fixed for the greater part of the period under review 1991 to 2004. Under a fixed exchange rate, the excessive rate of domestic credit expansion is fully reflected in a depletion of foreign exchange reserves. There was a moderate devaluation of the local currency until 1997. After 1997 there was a shift from a fixed to a floating exchange rate. Figure 1 illustrates the devaluation of the Zimbabwean dollar over the period 1991–2004.

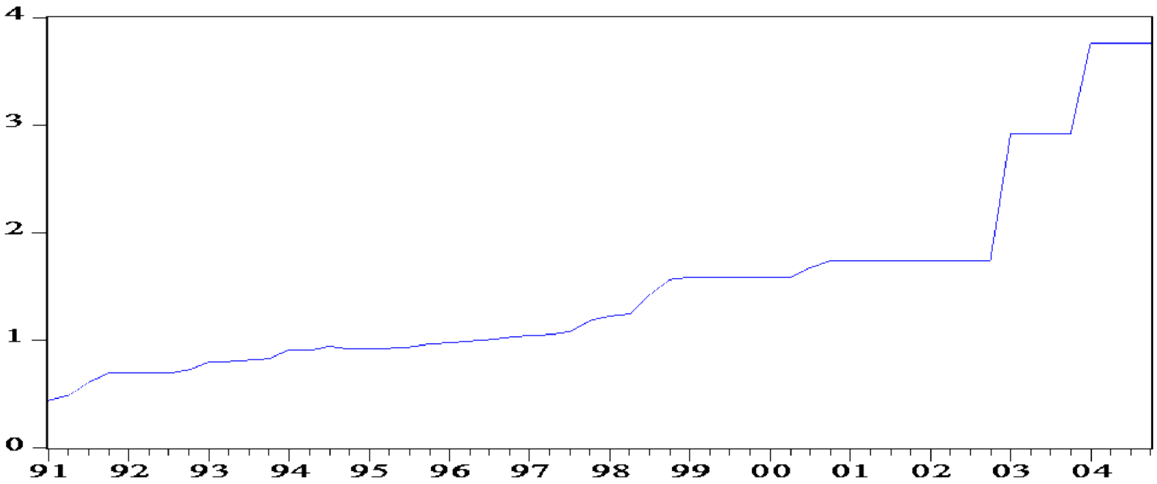


Figure 1: *Exchange rate for the period 1991-2004. The vertical axis shows the logarithm of the exchange rate*

[1] In 2008 the hyper inflation rate surged to 500 billion percent

2.3 Gross Domestic Product

Growth pattern of the economy was sluggish during the sampling period 1991-2004. Some contributing factors to the decline of the economy were drought, adverse effects of the reform programme, world recession particularly in the year 1992, high inflation, high interest rates, tight liquidity and resultant low investment. Real GDP declined by about 30% since 1999 [9].

2.4 Budget Deficit

During the period under which reforms were implemented, high budget deficit was the main cause of macroeconomic instability. It was an outcome of the growth in frequent and magnitude of unbudgeted expenditures. The budget was affected by decreased taxes, tax reforms and retrenchment that lowered the revenue base, high increment of wages and salaries and large subsidy allocation.

The budget deficit level was also affected by the erratic nature of the disbursement of the external support, that is, suspension of balance of payments support by the International Monetary Fund. The financing of the government budget deficit was mainly from domestic banking system which generated inflationary pressures. Droughts experienced during the reform period worsened the deficit situation due to food imports.

2.5 Summary of Macroeconomic Background

The period 1991 - 2004 was characterized by high inflation, unstable exchange rate; persistently, high budget deficits, poor export performance and social unrest. Because of the macroeconomic instability the performance of the economy could have had a negative impact on the performance of the financial sector. In the banking sector the period 1991 - 2004 was characterized by poor corporate governance, poor risk management systems, high levels of inside abuse via indecent loans to related parties, overindulgence in speculative non-core banking activities, inadequate capitalization, lax prudential supervision and regulatory forbearance. Consequently the year 2004 saw nine financial institutions being placed under curatorship. Five banks were eventually placed under liquidation during the same year and five Asset Management Companies had their licences cancelled [8].

3 The Econometric Model

In this section we briefly discuss the logistic econometric model used in this study.

Let $Pr(CC_t = 1|x)$ be the probability of a currency crisis. Then

$$Pr(CC_t = 1|x) = \frac{1}{1 + e - CC_t} \tag{1}$$

where

$$CC_t = c + X_t\beta + \varepsilon_t, t = 1, \ldots, T \tag{2}$$

and

$$\pi_t = Pr(e_t + 1 > \bar{e}) \tag{3}$$

with

$$CC_t = \begin{cases} 1 & \text{if } \pi_t \geq 0.30 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where CC_t takes value one if a currency crisis occurs at time t and zero otherwise, X_t is a vector of predictor variables, β is a vector of coefficients, ε_t is a normally distributed stochastic disturbance term, e_t is nominal exchange rate at time t, $\bar{e}$ is a fixed exchange rate, i.e. $e_t = \bar{e}$ and π_t is a one-step-ahead probability of a regime change which is approximated by the probability of a speculative attack, which is assumed to take place when speculators operating in the current period expect the shadow rate to exceed the actual fixed rate e in the next period [9]. Equation (2) can therefore be written as

$$\begin{aligned} CC_t =& c + \beta_1 \ln TBR + \beta_2 \ln ER + \beta_3 \ln M1 + \beta_4 BD + \beta \frac{5BA}{GDP} + \beta_6 \ln GDP \\ &+ \beta_7 \ln CPI + \beta_8 \frac{BD}{GDP} + \beta_9 \ln BA + \beta_{10} \ln MI + \beta_{11} \ln II \varepsilon_t, \end{aligned} \tag{5}$$

where GDP is the gross domestic product which represents real activity (national output), TBR, the treasury bill rate represents interest rate, [2] represents money supply, ER represents the exchange rate, BD represents budget deficit, CPI represents inflation, BA represents bank assets, MI is the mining index and II is the industrial index and ln in all cases represents the natural logarithm. The parameters, $\beta j, j = 1, 2, \ldots, 11$ and c are estimated using the maximum likelihood method. The estimates were obtained by [3] algorithm using numerical derivatives. The log-likelihood function, $L(\beta|x_i)$ is given by

$$\ln L(\beta|x_i) = \sum_{i=1}^{n} [(1 - CC_t) \ln Pr(CC_t = 0|x_i; \beta) + CC_t \ln Pr(CC_t = 1|x_i; \beta)]. \tag{6}$$

For the logit model we have

$$Pr(CC_t = 1|x_i; \beta) = \Lambda(x_i'\beta) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}; Pr(CC_t = 0|x_i; \beta) = 1 - \Lambda(x_i'\beta) = \frac{1}{1 + \exp(x_i'\beta)}, \tag{7}$$

to give the log-likelihood of the form

$$\ln L(\beta|x_i) = \sum_{i=1}^{n} [(1 - CC_t) \ln[1 - \Lambda(x_i'\beta)] + CC_t \ln \Lambda(x_i'\beta)], \tag{8}$$

while for the probit model we have

$$\Pr(CC_t = 1|x_i; \beta) = \Phi(x'i\beta); \Pr(CC_t = 0|x_i; \beta) = 1 - \Phi(x'i\beta), \tag{9}$$

to give the log-likelihood of the form

$$\ln L(\beta|x_i) = \sum_{i=1}^{n} [(1 - CC_t) \ln[1 - \Phi(x_i'\beta)] + CC_t \ln \Phi(x_i'\beta)], \tag{10}$$

[2] M1(narrow money) is notes and coins in circulation plus demand deposits in the banking system.

4 Results and Discussion

In this section we present results from logistic regression and a comparative analysis is done with a probit model.

4.1 Logit Model Results

The data consist of various quarterly macroeconomic variables from 1991 to 2004. The choice of variables was largely dictated by theoretical considerations and variables used in existing surveys. The variables are inflation, exchange rates, money supply growth, interest rates, gross domestic product, bank assets, ratio of budget deficit to GDP, ratio of bank assets to GDP, budget deficit, mining and industrial indices.

$$\begin{aligned} CC_t = c + \beta_1 \ln TBR + \beta_2 \frac{BA}{GDP} + \beta_3 \ln ER + \beta_4 \ln M1 + \beta_5 BD + \beta_6 \ln BA \\ + \beta_7 \ln GDP + \beta_8 \ln CPI + \beta_9 \frac{BDG}{DP} + \beta_{10} \ln MI + \beta_{11} \ln II + \varepsilon_t \end{aligned} \tag{11}$$

and $Pr(CC_t = 1|x)$ as defined in equation (1).

Table 1: *Empirical results of the logit model*

Parameter	c	BD	BD/GDP	lnCPI	lnER	BA/GDP
Coefficient	-163.24	0.00099	-1.59	36.14	-55.64	0.38

Parameter	lnGDP	lnM1	lnTBR	lnII	lnMI	lnBA
Coefficient	48.65	-4.28	-13.84	53.48	-89.47	4.71

McFadden R-squared is 0.995222 and the p-value of the LR statistic is 0.00043. Substituting the values of the coefficients in equation (4) we get

$$\begin{aligned} CC_t = -163.24 + 0.00099BD - 1.59\frac{BD}{GDP} + 36.14 \ln CPI - 55.64 \ln ER + 48.65 \ln GDP \\ - 4.28 \ln M1 - 13.84 \ln TBR + 53.48 \ln II - 89.47 \ln MI + 0.38BAGDP + 4.71 \ln BA \end{aligned} \tag{12}$$

4.2 Probit Model Results

Table 2: *Empirical results of the Probit model*

Parameter	c	BD	BD/GDP	lnCPI	lnER	BA/GDP
Coefficient	-70.14	0.00075	-2.58	12.88	-20.89	0.12

Parameter	lnGDP	lnM1	lnTBR	lnII	lnMI	lnBA
Coefficient	19.18	1.55	-5.19	13.41	-27.19	2.25

McFadden R-squared is 0.974957 and the p-value of the LR statistic is. After substituting values of coefficients in equation (4) we get

$$\begin{aligned} CC_t &= -70.14 + 0.00075BD - 2.58\frac{BD}{GDP} + 12.88\ln CPI - 20.89\ln ER + 19.18\ln GDP \\ &\quad + 1.55\ln M1 - 5.19\ln TBR + 13.41\ln II - 27.19\ln MI + 0.12BAGDP + 2.25\ln BA \end{aligned} \tag{13}$$

In both logit and probit estimations the slopes are varying and the marginal effects will be conditional upon a particular value of the explanatory variable. The observed responsiveness for any given observation will be conditional upon the magnitude of the explanatory variable.

Table 3: *Goodness of fit measures*

	Probit Model	Logit Model
McFadden R-squared	0.9749	0.9952
LR statistic (11df)	32.85	33.54
P-value of LR statistic	0.00055	0.00043
ROC area	0.6513	0.6742

Table 3 shows the goodness of fit measures. Comparing the two models the logit model is better than the probit model.

Table 4: *Forecasts of currency crisis probabilities*

Year	$CC_t = 1$	Probit Model: $Pr(CC_t = 1\|x)$	Logit Model: $Pr(CC_t = 1\|x)$
1991 Q2	1	0.8389	0.971221975
1998 Q2	1	0.83891	0.991497083
2002 Q4	1	0.97725	0.996209105
2003 Q4	1	1	1
2004 Q4	1	1	1

Table 4 shows the predictive power of the probit and logit models. The logit model outperformed the probit model in correctly predicting currency crisis in 1991, 1998 and 2002. However from 2003 onwards when currency crisis worsened both models were able to correctly predict a currency crisis. The findings from using the logit and probit models are very similar and both statistically reliable.

A graphical plot of the logit and probit model results is shown in Figure 2.

5 Conclusion

In this paper, a logistic regression model was developed to predict currency crisis in emerging markets with Zimbabwe as a case study. The findings from using the logit and probit models are very similar and both statistically reliable. The major findings in this study are: lax monetary policy can contribute to currency crisis; deterioration in economic fundamentals can also contribute to currency crisis; large devaluation can be perceived as currency crisis;

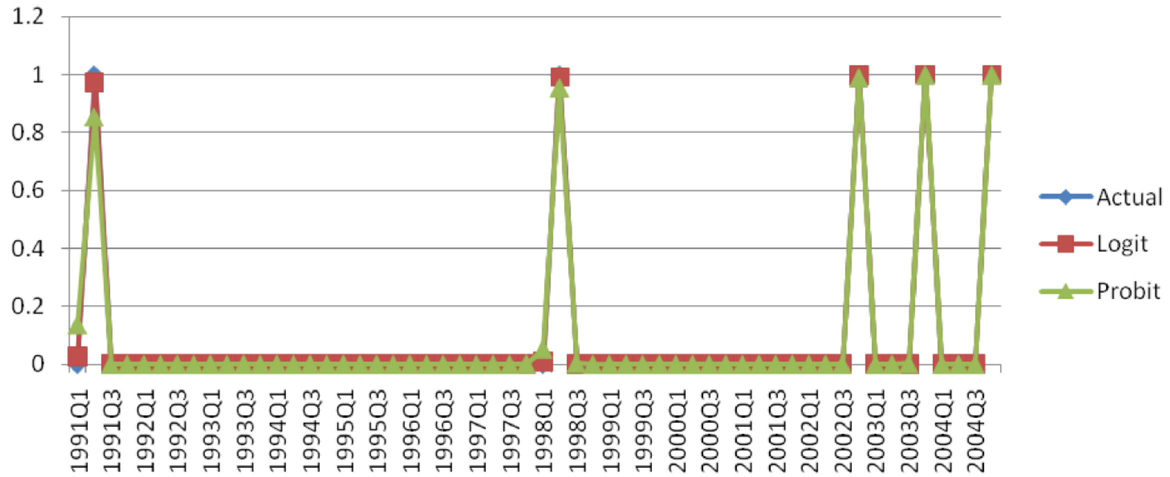


Figure 2: *Graphical plot of fitted values using the probit and logit models*

macroeconomic and financial variables are important determinants of currency crisis. The possible remedies to these challenges are that fiscal and monetary tightening is critical for a successful stabilization; price and exchange liberalization should be a high priority for stabilization; establishing a strong money anchor to reduce inflation and inflation expectations; the need to realize that outside intervention is crucial for stabilization.

Areas for further research would include timing of crisis, contagion effect, political risk, corruption, financial derivatives, direct investment and carrying out sensitivity analysis. Another interesting area of study would involve the development of a hybrid logistic regression- multivariate adaptive regression splines (LRMARS) model and use of binary quantile regression modelling. These areas will be studied elsewhere.

Acknowledgements

The author is grateful to Central Statistical Office of Zimbabwe, Reserve Bank of Zimbabwe, Old Mutual Asset Managers Zimbabwe and the Zimbabwe Stock Exchange officials for support with their respective company data and information, and to the anonymous reviewers for their valuable comments.

Bibliography

[1] Agenor, P. and Robert, P. (1994). *Macroeconomics Policy, Speculative Attacks, and Balance of Payments Crises.* Cambridge, Massachussetts, and Oxford, England: Blackwell.

[2] Chang, R. and Velasco, A. (2000). Financial fragility and the exchange rate regime. *Journal of Economic Theory*, 92:1–34.

[3] Chang, R. and Velasco, A. (2001). A model of currency crises in emerging markets. *Quarterly Journal of Economics*, 2(2):489–517.

[4] Fontaine, T. (2005). Currency crises in developed and emerging market economies: a comparative empirical treatment. IMF Working Paper WP/05/13.

[5] Jeanne, O. (1997). Are currency crises self-fulfilling? a test. *Journal of International Economics*, 36:413–430.

[6] Krugman, P. (1979). A model of balance-of-payment crises. *Journal of Money, Credit and Bank*, 11:311–325.

[7] Obstfeld, M. (1986). Rational and self-fulfilling balance-of-payments crises. *American Economic Review*, 76(1):72–81.

[8] of Zimbabwe, R. B. Annual report, 2004. http://www.rbz.co.zw/annual/2004contents.asp.

[9] of Zimbabwe, R. B. Annual report, 2005. http://www.rbz.co.zw/annual/2005contents.asp.

[10] Peltonen, T. (2006). Are emerging market currency crises predictable? No. 571 / January http://www.ecb.int or from the Social Science Research Network electronic library at http://ssrn.com/abstract_id=872529.

[11] Siguake, C., Mapose, D., Mudimu, E., and Nyamugure, P. (2010). Volatility modeling using arima-garch models in a hyperinflationary economic environment: The zimbabwean experience. In *Peer-reviewed Proceedings of the Annual Conference of the South African Statistical Association.* [Online] available: https://sites.google.com/site/sasaconference2010/.

[12] Su, C., Chang, H., M-N, Z., and Z, Q. (2010). An evaluation of leading indicators of currency crises. *African Journal of Business Management*, 4(15):3321–3331.

[13] Wikipedia (2011). http://en.wikipedia.org/wiki/financial_crisis_of_2007%e2%80%932010.

[14] Yu, L., Lai, K., and Wang, S. (2006). Currency crisis forecasting with general regression neural networks. *International Journal of Information Technology & Decision Making*, 5(3):437–454.

Proceedings of the 2011 ORSSA Annual Conference
pp. 38–50
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Estimating the threat of water scarcity in the Breede Water Management Area: a forecast-based analysis

SE Bester[*] TE Lane-Visser[†]

Abstract

Water supply and distribution systems in the Western Cape are under strain in terms of capacity. A number of studies have found that the pressure to supply enough water of suitable quality in the Western Cape is constantly increasing. In these studies, urban development and the effects of global warming are listed as key factors affecting the problem. Many of the studies found that the current water infrastructure will not be able to cope with growing demand unless new investments are materialised and water is consumed more judiciously.

The availability of water is directly dependent on the balance between supply and demand. In terms of supply, weather patterns and infrastructure design, condition and availability are some of the key factors to take into consideration. Water demand is influenced by population growth, agricultural activities as well as weather patterns.

In this study components that influences supply and demand in the Breede Water Management Area (WMA) was determined. Multiple regression analysis was used to establish the influence that these components have on the supply and demand. Several scenarios depicting the influences that climate change may have, were used to forecast the supply and demand for the next 20 years. The severity and timeframes related to the onset of water scarcity in the Breede WMA was predicted.

The main goal of the study was to investigate the expected simultaneous changes in the supply and demand as well as the resulting net impact on water availability. It was found that all scenarios have shortfalls within the forecasting period. In the best case scenario, which is already severe, water shortages will occur 12 years from now. The water shortage by 2031 will be 22%. In the worst case scenario water shortages will occur within 6 years.

Key words: Forecasting, regression, scenarios, water resources

[*] Corresponding author: Stellenbosch University, South Africa, email: 15400476@sun.ac.za
[†] Stellenbosch University, South Africa, email: tanyav@sun.ac.za

1 Introduction

South Africa is situated in a region with increasing levels of water scarcity and water quality problems; compounded by population growth and issues of social and economic development [9]. Water resources in the country are limited and additional stresses on water resources, such as those arising from climate change, could exacerbate water scarcity over much of the country and the Southern African region. For this reason the proper management of water resources in South Africa is vital. Droughts of varying extent are a regular occurrence in South Africa.

The impacts of climate change are predicted to raise global and regional temperatures and cause changes in other climate variables that drive the terrestrial hydrological cycle – most notably precipitation and potential evaporation [9]. An increase in temperatures is predicted to result in increased water demand from the domestic, agricultural and industrial sectors. Supply capacity, on the other hand, is shown to decrease non-linearly as either precipitation decreases or potential evapotranspiration increases. Both of these changes in precipitation and evapotranspiration are predicted for the Breede WMA [9]. The Breede Overberg Catchment Management Agency [3] has recently stated that the effects of climate change, usage requirements and the seasonal water scarcity has to be considered when catchment management strategies are formulated. This notion is supported by Steynor et al. [11], of Water Affairs and Africa [10] and New [9].

Many factors affect the achievable levels of water supply and the volumes of water demanded in a water distribution system. These factors are constantly changing, yet water distribution managers need reliable information to base their management strategies on.

The objective of this study is to inform managers on the extent of the threat of water scarcity occurring in the Breede Water Management Area (WMA), cognisant of the expected impact of climate change and other changes external to the system. The water scarcity threat is estimated both in terms of the potential magnitude and the expected timeframes of the water scarcity onset.

An overview of the current state of water supply and demand in the Breede WMA is provided in section 2. Sections 3 to 6 inform on the forecasting approaches used to project future supply and demand in the Breede River WMA. This is followed by an exposition of the combined impact of both these changes on water provision in section 7. The paper concludes with a summary of the most important findings reached.

2 The Breede Water Management Area

The Breede WMA is situated in the South-Western region of South Africa. It derives its name from the largest river within its boundaries, namely the Breede River. The Breede WMA is bounded by the Indian Ocean to the south, the Olifants/Doorn WMA in the northwest, the Berg WMA in the west and the Gouritz WMA in the east (Figure 1). The Breede WMA can be divided into two sub-management areas, namely the Breede and the Overberg sub-management areas. The Breede sub-management area is further divided into the Upper Breede, Lower Breede and Riviersonderend sub-areas [10].

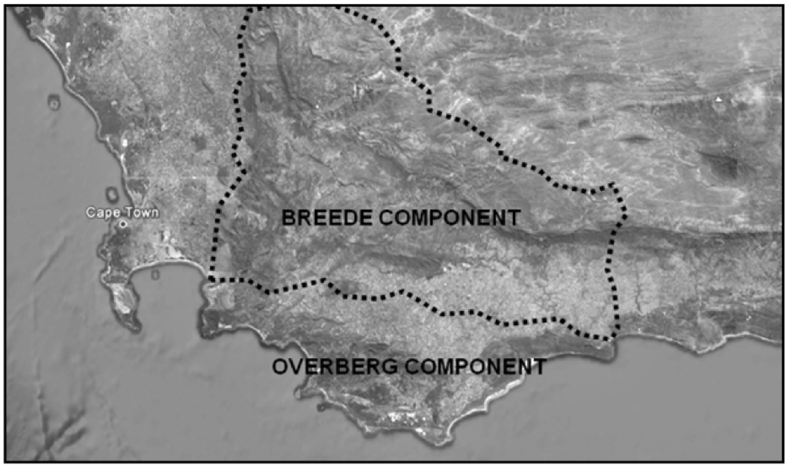


Figure 1: *Breede Component of Breede WMA*

The topography of the Breede WMA is characterised by mountain ranges in the north and west, the wide Breede River Valley, and the rolling hills of the Overberg. Rainfall is highest in the mountainous regions in the southwest where the mean annual precipitation is as high as 3000 mm per annum, whilst the central and north-eastern areas receive as little as 250 mm per annum. Most of the WMA has a winter rainfall pattern while an all year rainfall pattern prevails only in the far south-east. The average potential mean annual evaporation (S-Pan) ranges from 1200 mm in the south to 1700 mm in the north of the WMA. The mean annual temperature equals 17C for the whole WMA. Frost and occasional snowfall occur in the winter [10]..

The Breede River and its largest tributary, the Riviersonderend River, are the two main rivers in the Breede River component, and drains an area of approximately 12 600 km^2. A number of water supply schemes have been developed in the Breede WMA, and with the exception of the Lower Breede sub-area, the WMA is generally crisscrossed by canals and pipelines supplying water for irrigation of commercial crops. The Greater Brandvlei Dam, with a capacity of 475 million m^3, is filled mainly during the winter months with water from the Smalblaar River and the Holsloot River [3]. During the summer irrigation period, water is released from the Brandvlei dam into the Breede River to supplement river flows for use by a number of water usage associations. The Greater Brandvlei Dam and its inflow and outflow canals can be seen in Figure 2.

Water is also supplied through pumping schemes directly from the dam to nearby irrigation districts. The Theewaterskloof Dam (434 million m^3) is the source reservoir for the Riviersonderend-Berg-Eerste River Government Water Scheme, an inter-catchment water transfer project owned and managed by the DWAF that also supplies water to the Berg WMA (BOCMA, 2010). Stettynskloof Dam is the only dam of significant size that is owned by a local authority. Its primary purpose is to supply water for domestic use. Of the dams supplying water for irrigation, the Greater Brandvlei Dam (yield of 155 million m^3/a) is the largest. It has spare storage capacity of 133 million m^3 citepdwaf2004. This offers potential

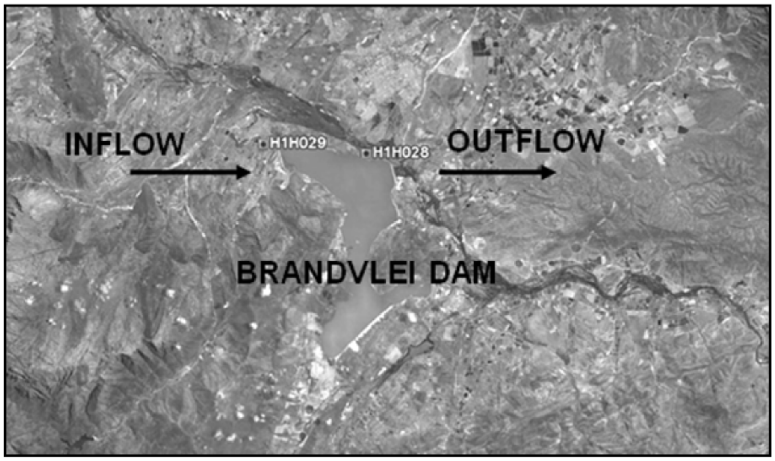


Figure 2: *Greater Brandvlei Dam*

for increasing the yield via pumping out of the Breede River. Other large dams used for irrigation include the Lakenvallei and Roode Elsberg Dams of Sanddrift Government Water Scheme (yield of 9 million m^3/a), the Keerom Dam (yield of 3.8 million m^3/a), the Elandskloof Dam (yield of 12 million m^3/a), the Buffeljags Dam (yield of 11 million m^3/a) and farm dams collectively providing about 83 million m^3 of storage [10].

The total volume of water (made up of surface water, ground water, return flows and transfers) available for supply in the Breede River component is listed in Table 1.

Table 1: *Sources of water supply in the Breede River component (million m^3/a) (Source: citetdwaf2004.)*

Resource Category	**Upper Breede**	**Riviersonderend**	**Lower Breede**	**Total**
Gross Surface Water Resource Yield	428	262	59	**749**
Less Impact on Yield of:				
Preliminary Ecological Reserve	16	0	0	**16**
Invasive Alien Plants	25	13	7	**45**
River Losses	5	0	0	**5**
Net surface Water Resource	**382**	**249**	**52**	**683**
Plus Groundwater	94	5	4	**103**
Plus Return Flows	85	10	7	**102**
Total Local Yield	**561**	**264**	**63**	**888**
Transfers In	0	0	14	**1**
Total	**561**	**264**	**77**	**889**

The main elements comprising total water demand in the area are irrigation (95% of total demand) and urban settlements (approximately 5%) [3, 10]. The total demand volumes are provided in Table 2. Water for domestic use is primarily supplied out of schemes owned and operated by local authorities.

Table 2: *Sources of water supply in the Breede River component (million m^3/a (Source: citetdwaf2004)*

Category	Upper Breede	Riviersonderend	Lower Breede	Total
Irrigation	495	91	72	658
Urban	23	2	1	26
Rural	4	2	1	7
Impact of Afforestation on Yield	0	1	0	1
Total Requirements	**522**	**96**	**74**	**692**
Transfers Out	22	168	0	177
Total	**544**	**264**	**74**	**869**

Water is transferred from the Upper Breede area to the Berg WMA (9 million m^3/a) and the Olifants/Doorn WMA (2.5 million m^3/a). A further 10 million m^3/a is released to maintain acceptable water quality levels. Transfers are made to the Overberg component (4 million m^3/a) and the Lower Breede area (2.5 million m^3/a) [3, 10].

When comparing the supply and demand figures, it was determined that a surplus of 20 million m^3/a exists (17 million m^3/a in the Upper Breede area). This surplus lies in the Koekedouw Dam (3 million m^3/a), the Stettynskloof Dam (14 million m^3/a) and the Buffeljags Dam (3 million m^3/a). The former two dams are not owned by the DWAF [3].

It can be concluded that the Breede Component's water supply system is very close to its capacity. In fact, shortfalls are already occur during the dry summer seasons ([3]. New [9] states that the system is already unable to provide a 1:50 year yield, and will only be able to do so in the future if projected increases in demand (due to socio-economic factors) are matched by efficient demand management practice. "If the climate of the region evolves as suggested by recent general circulation model predictions, the resultant supply decreases and demand increases will exacerbate the existing water resource problems in Cape Metropolitan Region". [9]

From the supply and demand data discussed, it is evident that the Brandvlei Dam is the main source of water for the Breede Component of the Breede WMA. The rest of this study is therefore limited to supply from the Brandvlei Dam and demand in the Breede Component.

3 Methodology

Forecasting methods can generally be classified into two groups: extrapolation and causal forecasting. Extrapolation methods forecast future values of a time series from past values of a time series. Here it is assumed that past patterns and trends will continue in the future. Causal forecasting methods attempt to forecast future values of a dependent variable using historical data to estimate the relationship between the dependent and one or more independent variables [12, 2, 8] The first step in this study was to identify which forecasting methods are most appropriate for use in both the supply and demand forecasts, respectively.

The demand and supply from the Brandvlei Dam are both influenced by many variables. For instance, the precipitation, evapotranspiration and levels of the supplementary dams in the region have an influence on the demand from the Brandvlei Dam. Supply from the Brandvlei Dam is also dependent on a many different variables. It was concluded that a causal forecasting method like regression is the most suitable approach.

Historic data from 1987 until the end of 2010 was used. The data was obtained from the Department of Water Affairs' website [1].

4 Demand Forecasting

There are many possible factors that could contribute to the volume of water demanded in a region. Some of these include demographics such as population size, economic prosperity, extent of agriculture, access to external water sources and meteorological factors. The of Water Affairs and Africa [10] states that no significant increase in population is expected in the WMA over the next few years; hence this variable was excluded from the analysis. Moreover, the region's contribution to national gross domestic product (GDP) is less than 1% [10]. This, combined with the slow growth rate of GDP, warrants the omission of GDP as a causal variable in the study.

The extent of agriculture in the region is assumed to remain relatively constant over the next twenty years. As discussed in section 2, no transfers are made into the Breede River component of the WMA. Water is only transferred from one sub-area to another. There is however a number of supplementary dams in the region and it is expected that those dams can either alleviate or exacerbate the water demand from the Brandvlei Dam. The independent variables that can be used to predict the volume of water demanded from the Brandvlei Dam over the next twenty years are the water available from supplementary dams, precipitation and evapotranspiration.

When regression was applied to the above mentioned data, it resulted in the following equation:

$$\text{Water demand } = \alpha_0\text{Supplemetary damns} + \alpha_1\text{Precipitation} + \alpha_2\text{Evapotranspiration} + \epsilon_1$$

When doing multiple regression analysis, some key assumptions must be tested. Violation of any of these assumptions could render the study results void. These assumptions are: variables and errors must be normally distributed, homoscedasticity is present (indicated by the Durbin-Watson statistic) and that errors are independent of each other (test for autocorrelation) [12, 2, 8]. It was found that both heteroscedasticity and positive autocorrelation was present in the data. The model was adjusted for heteroscedasticity by taking the square root of the water demand variable. Each of the independent variables was adjusted to remove the autocorrelation. The original residual plots of the data can be seen in Figure 3. The adjusted and final data plots can be seen in Figure 4.

The results from the adapted model were a great improvement on that of the first model and are sufficient to allow acceptance of the model. The final R^2 value is 0.83, the standard error 4.5 and the Durbin-Watson statistic 1.67. Although all the p-values of the coefficients

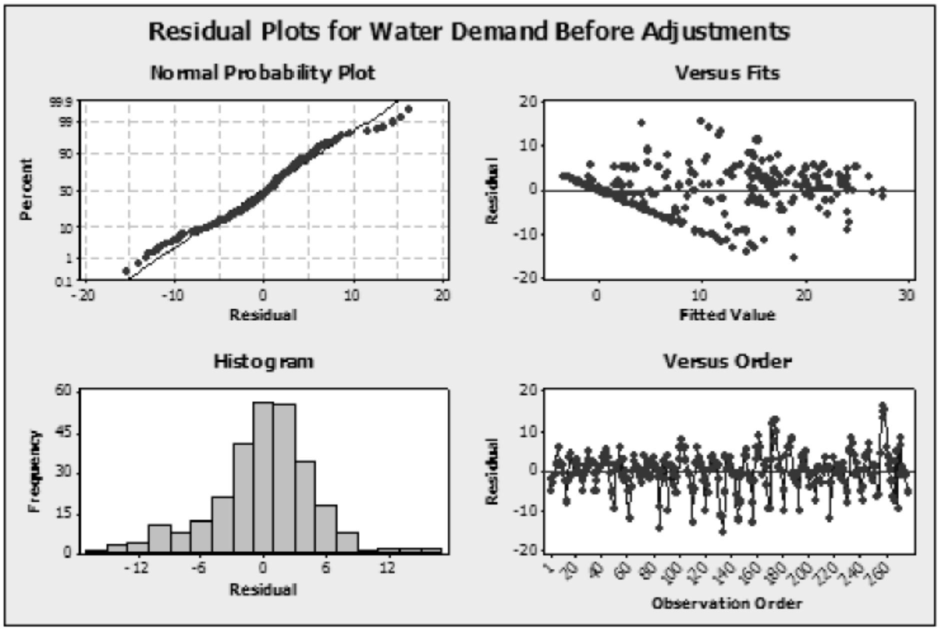


Figure 3: *Residual plots for water demand before adjustments was made to the data.*

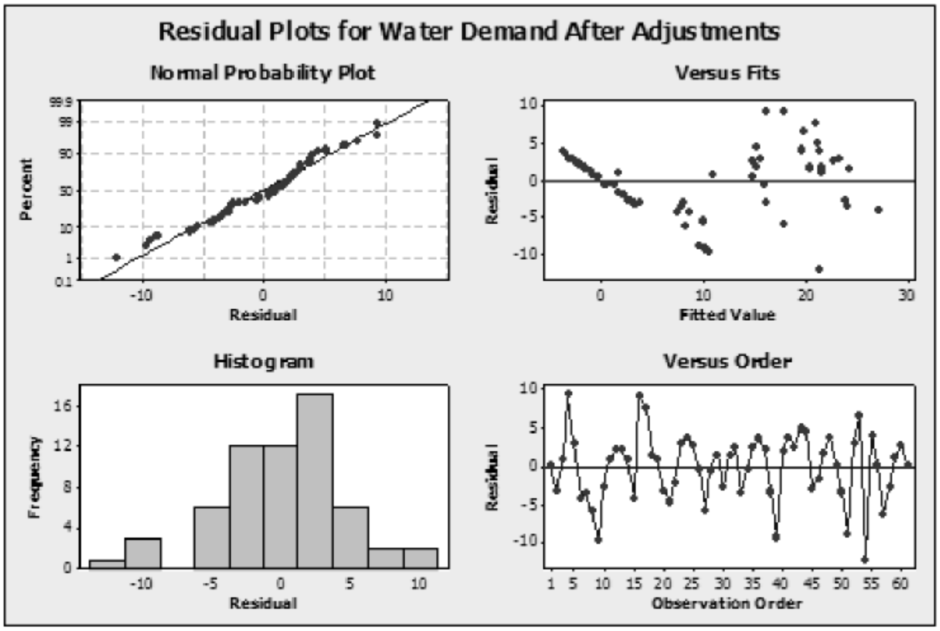


Figure 4: *Residual plots of the final multiple regression model.*

are not smaller than 0.05, none of them are greater than 0.25. Evapotranspiration is the most influential variable and is negatively correlated to water demand. The impacts of the supplementary dam levels and precipitation on demand are similar in size and positively correlated to demand.

The resulting regression model can now be used to forecast water demand levels, based on expected changes in the precipitation, evapotranspiration and supplementary dams' variables. Uncertainties in estimating future water requirements (such as the impact of climate change, the impact of changes in land-use and the impact of water conservation and demand management) are abound [10].

5 Supply Forecasting

The total water available for supply (i.e. the Brandvlei Dam level plus the outflow canal) is used as the dependent variable. This can be seen as the total supply in the system. The dependent variable is again influenced by many independent variables. These independent variables are the precipitation, evapotranspiration, the inflow canal and the previous dam level. The previous dam level was determined by the following equation:

$$\text{Dam}_{t-1} = \text{Dam}_{t-2} + \text{Inflow}_{t-1} + \text{Precipitation}_{t-1} - \text{Outflow}_{t-1} - \text{Evaporation}_{t-1}$$

The precipitation and evaporation used in this equation was calculated only for the area of the dam. The dam level and therefore the dam's area was calculated for each period to determine the amount of precipitation and evaporation applicable to the dam for that period.

When regression was applied to the above mentioned data, it resulted in the corresponding regression equation:

$$\text{Water supply } = \beta_0\text{Dam}_{t-1} + \beta_1\text{Precipitation} + \beta_2\text{Evapotranspiration} + \beta_3\text{Inflow} + \epsilon_1$$

Similar to the demand, key assumptions had to be tested. All the assumptions were satisfied. The Durbin-Watson statistic's value was 1.8, which is acceptable. However, because a variable is lagged, the Durbin h statistic must also be calculated. The Durbin h statistic is a test used to determine if autocorrelation is present if one or more independent variables are lagged. In this case, Dam_{t-1} was lagged by one period. The Durbin h statistic was 1.91, which indicates that no autocorrelation was present. The plots for the residuals can be seen in Figure 5.

The final R^2 value is 0.74 and the standard error is 29.19. The p-values of precipitation, Inflow and Dam_{t-1} were all smaller than 0.05. The p-value of evapotranspiration was smaller than 0.1.

6 Forecasting Scenarios

There is consensus in the literature that there is no concrete causal relationship between changes in temperatures (expected to be the main effect from climate change) and changes in rainfall [4, 5, 6, 7]. The only expected impact is that weather phenomena will become more

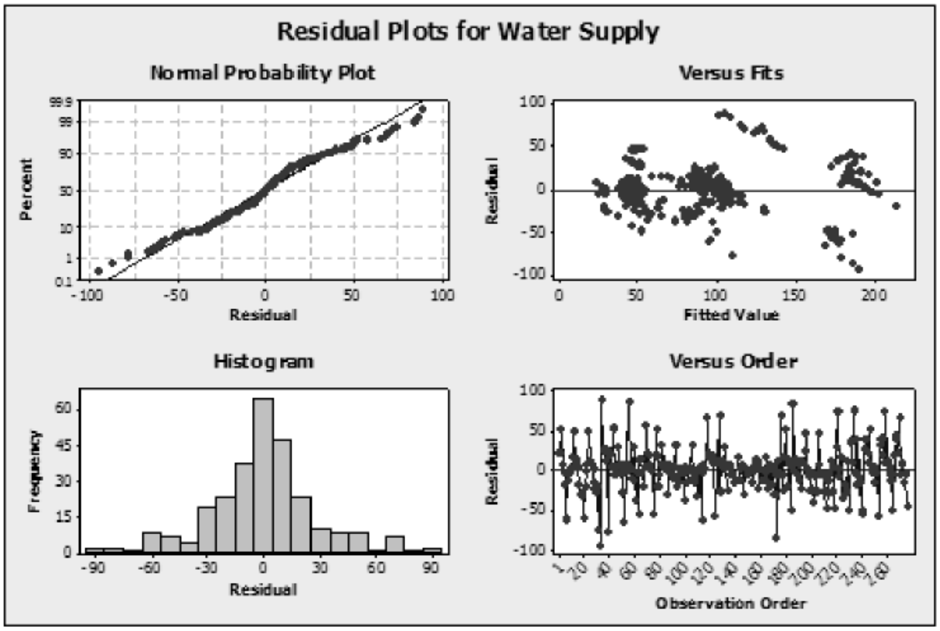


Figure 5: *Residual plots for the water supply.*

extreme (in terms of rainfall or dry spells) as temperatures increase, although not necessarily more frequent. If one of the extremes occurs more frequently in a region, it could have a resulting positive or negative impact on the availability of water supply in that region. New [9] suggests that climate change will result in an annual reduction in stream flow in the South-Western Cape over the next twenty years. This supports the notion that it is reasonable to add a "low" supply scenario to the study. It is possible to increase potential yield in catchment systems without the addition of new infrastructure [10]. A "high" supply scenario will reflect the implementation of such practices.

Further, New [9] suggested that a 5% decrease in precipitation can be expected by 2020. Assuming this is a linear decrease [9], a 2.5% decrease in average annual precipitation will have occurred by 2010. Evapotranspiration is calculated based on a formula that relies on both temperature and humidity data. In general, as temperature increases, humidity tend to decrease (an inverse relationship is present between these variables). A linear 1C increase in average global temperatures is expected by 2020 [9]. The above mentioned factors were used to develop the normal demand and supply scenarios.

6.1 Demand

In addition to the forecast above for the normal scenario (section 6), a forecast must be done for the supplementary dams. A trend based on historic data was used to forecast the mean dam levels over the next twenty years.

To develop the high demand scenario precipitation forecast, a more severe 5% decrease in average annual rainfall was used. The corresponding evaporation forecast was based on an expected increase in average temperature to 1C in twenty years. Furthermore, empty supplementary dams would increase demand for water from the Brandvlei Dam. To calculate this, a linearly decreasing trend in dam water levels to the minimum recorded historic dam levels was used.

Correspondingly, the low demand scenario was based on an expected increase in current precipitation levels by 2.5% over twenty years and no expected increase in temperature due to climate change. To determine the dam levels suited to this scenario, dam levels were trended to reach maximum capacity levels by 2031.

The high, normal and low scenarios are summarized in Table 3.

Table 3: *Demand Scenarios*

	Demand Scenarios		
	High	Normal	Low
Precipitation	-5%	-2.50%	2.50%
Supplementary dams	Trend to minimum capacity	Trend on historic data	Trend to maximum capacity
Temperature	+1C	+0.5C	0

6.2 Supply

In section 6, the normal scenario was already discussed, but the stream flow and Dam$_{t-1}$ also have to be considered for the normal supply scenario. New [9] suggested that stream flow will decrease with 6.4% by 2020. Assuming a linier decrease, a 3.2% decrease of stream flow would have occurred by 2010. In this study it is assumed that the linier decrease will continue for the next twenty years. Dam$_{t-1}$ was forecasted using the same factors, but only applicable to the area of the dam. The above mentioned factors were used to develop the normal demand scenario.

For the high supply scenario precipitation increases with 2.5%, temperature has no change, the stream flow decreases with 1%, Dam$_{t-1}$ was predicted using the same factors, but only data applicable to the area of the dam was used.

For the low supply scenario precipitation decreases with 5%, temperature increases with 1C, the stream flow decreases with 5%, Dam$_{t-1}$ was predicted using the same factors, but only data applicable to the area of the dam was used. Table 4 summarises the supply scenarios.

Table 4: *Supply scenarios*

	Supply Scenarios		
	High	Normal	Low
Precipitation	2.50%	-2.50%	-5%
Temperature	0	+0.5C	+1C
Stream flow	-1%	-3.20%	-5%
Dam$_{t-1}$	Calculated the same as above		

7 Discussion

Figure 6 shows a graphic illustration of the simultaneous forecasts of water supply and demand in the Upper Breede sub-area of the Breede WMA. The graph illustrates both the severity and timeframes of expected shortfalls in water supply within the next twenty years. Regardless of the scenario, the demand is shown to steadily increase while the supply steadily decreases for the forecasting period.

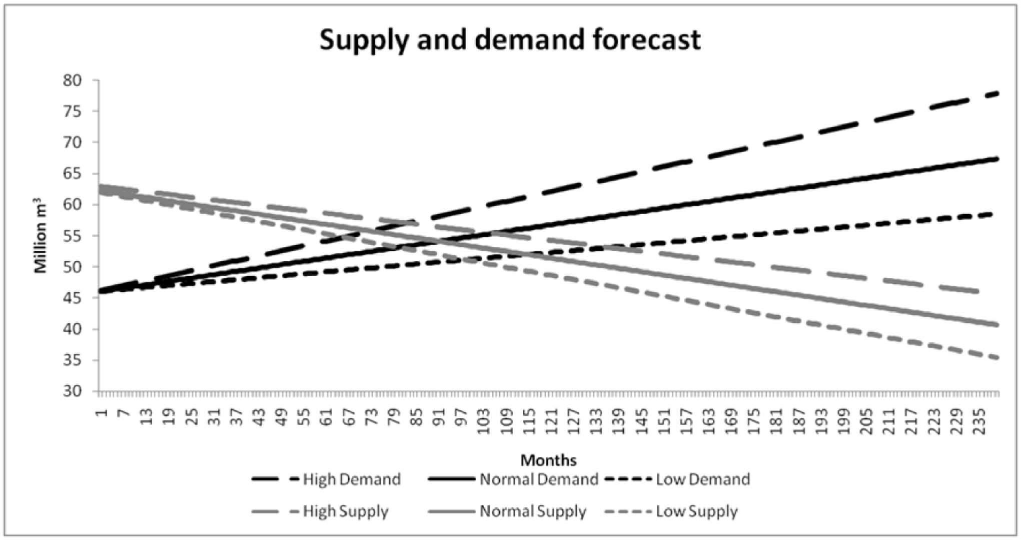


Figure 6: *Combined graph of supply and demand twenty year forecast scenarios*

A summary of the expected magnitude of the shortfall (given as percentage of demand that will not be met by 2031) and the expected timing of the shortfall is provided in Table 5.

Table 5: *Summary table of supply shortfall threats*

			Water Demand		
			High	Normal	Low
	High	Expected Shortfall by 2031	41%	32%	22%
		Expected Timeframe (years)	7	9	12
Water supply	Normal	Expected Shortfall by 2031	48%	40%	31%
		Expected Timeframe (years)	7	8	10
	Low	Expected Shortfall by 2031	55%	47%	39%
		Expected Timeframe (years)	6	7	9

The estimated worst case scenario according to this study, will occur within 6 years from now are between high demand and low supply. The water shortage by 2031 will be 55%; which means that about half of the people in the Breede WMA will not be able to use water as they previously did.

The estimated best case scenario according to this study is when water shortage only occurs 12 years from now. This will happen when there is low demand and high supply. The water shortage by 2031 will be 22%; which means that about one fifth of the people in the Breede WMA will not be able to use water as they previously did.

There may be ways to delay or prevent the onset of water shortage. Methods that can be implemented to increase water supply capacity have been proposed. The first and most obvious method is to fully exploit any spare storage capacity that is underutilised in the system. Secondly, existing infrastructure can be expanded (by raising dam levels for example). Thirdly, new infrastructure can be developed. Other methods include the verification of existing lawful use and management of water, water conservation and demand management interventions.

The main opportunities to save water lie in the maintenance and upgrading of water conveyance and distribution systems as well as improved management of releases from the Greater Brandvlei Dam [10]. Trading of existing authorisations is a way of shifting water towards more beneficial use or higher paying use, without increasing total volume demand. Clearing of invasive alien plants can result in water gains. Priority areas include the upper reaches of the Riviersonderend and Upper Breede sub-area [10]. The use of bio-control presents a cost-effective and sustainable form of control. Further to this, abstracting groundwater (which only has weak links to surface water and can be abstracted without significantly impacting on surface water yields) can prove to be a solution to the problem. However, within the Ceres catchment (Upper Breede sub-area), groundwater abstraction currently exceeds what is considered to be sustainable abstraction [10].

8 Conclusions

Assuming that [9] and other researchers' predictions of climate change is correct, all scenarios have shortfalls within the forecasting period. In the best case scenario, which is already severe, water shortages will occur 12 years from now. This will happen when there is a low demand and high supply. The water shortage by 2031 will be 22%. In the worst case scenario water shortages will occur within 6 years.

The values on the y-axis of Figure indicate the amount of water (in million m^3/month) when the forecasted demand surpasses the forecasted supply. The amount of water for the worst case and best case scenarios are 55 million m^3/month and 53 million m^3/month respectively. These values are in line with the demand values of Table which, if given as a monthly amount, is 45 million m^3. The data from Table dates out of 2000 and it can be expected that the forecasted demand amounts will be more than that of Table .

It must be stressed again that these results are only estimated values. Nonetheless, the results show that planners and developers are afforded a lead time of approximately 6 years to effect the required interventions. Although there are a number of methods for increasing water supply capacity to choose from, there is not a lot of guidance on the method that would be best suited to the Breede WMA. Studies that inform on the best decisions are recommended for future research.

Bibliography

[1] (2011). Hydrological services surface water (data, dams, floods, flows). [Online]. Available: http://www.dwa.gov.za/Hydrology/. [2011, October 3].

[2] Anderson, D., Sweeney, D., and Williams, T. (2008). *Statistics for business and economics, 10th edition*. CENGAGE Learning, Mason (OH).

[3] (BOCMA), B.-O. C. M. A. (2010). Overberg catchment management strategy. [Cited June 28th, 2011], Available from http://bocma.co.za.

[4] Buishand, T. and A.M.G., K. (1996). Regression model for generating time series of daily precipitation amounts for climate change impact studies. *Stochastic Hydrology and Hydraulics*, 10:89–106.

[5] Buishand, T. and Brandsma, T. (1999). Dependence of precipitation on temperature at florence and livorno (italy). *Climate Research*, 12:53–63.

[6] Buishand, T. and Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the rhine basin by nearest-neighbor resampling. *Water Resources Research*, 37(11):2761–2776.

[7] Liu, S., Fu, C., Shiu, C., Chen, J., and Wu, F. (2009). Temperature dependence of global precipitation extremes. *Geophysical Research Letters*, 36:L17702.

[8] Makridakis, S., Wheelwright, S., and McGee, V. (1983). *Forecasting: Methods and applications (2nd edition)*. Wiley & Sons, Inc. Hong Kong.

[9] New, M. (2002). Climate change and water resources in the southwestern cape, south africa. *South African Journal of Science*, 98.

[10] of Water Affairs, D. and Africa, F. D. S. (2004). Breede water management area: Internal strategic perspective. [Government Report No P WMA18/000/00/0304], Government Publications, Pretoria.

[11] Steynor, A., Hewitson, B., and Tadross, M. (2009). Projected future runoff of the breede river under climate change. *Projected future runoff of the Breede River under climate change*, 35(4):433–440.

[12] W.L., W. (2004). *Operations research: applications and algorithms, 4th Edition*. CENGAGE Learning, Belmont (CA).

Proceedings of the 2011 ORSSA Annual Conference
pp. 51–58
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Estimation of multivariate proportional hazards model parameters with Artificial Bee Colony optimization

W Carstens* TE Lane-Visser† PJ Vlok‡

Abstract

Over years industry has seen an increased interest in physical asset management to improve competitiveness. As part of physical asset management, Asset Care Plans (ACP) consist of maintenance strategies used to maximize asset utilization and performance. Many companies perform condition monitoring on physical assets, but lack the knowledge to exploit this information to its full advantage. Condition monitoring information, integrated with previous failure histories of equipment through a mathematical model such as the Proportional Hazards Model (PHM), can be used to derive useful management information, including the estimation of asset deterioration rates, residual life and/or the risk of operation. The PHM is used in the field of reliability modeling and utilizes the flexible Weibull distribution as parametrization. Fitting the PHM involves maximizing the likelihood of the function, thereby estimating the model parameters. Maximization of the likelihood can be complex and often involves multiple parameter optimization in a large search space. Knowledge of these parameters allows for the exploitation of condition monitoring information to its full potential. This paper discusses the development of a metaheuristic algorithm to maximize the multivariate likelihood function of the PHM, based on the Artificial Bee Colony (ABC) optimization algorithm. The paper concludes with a reflection on the validity and success of solving this problem through ABC optimization.

Key words: Metaheuristics, Maintenance, Multivariate, Proportional Hazards Model, Artificial Bee Colony.

1 Introduction

Physical Asset Management (PAM) is becoming a greater concern for companies in industry today. The British Standards Institutes' specification for the optimized management of

*Corresponding author: Stellenbosch University, South Africa, email: whiehancarstens@gmail.com

†Stellenbosch University, South Africa, email: tanyav@sun.ac.za

‡Stellenbosch University, South Africa, email: pjvlok@sun.ac.za

physical assets and infrastructure is PAS 55. According to PAS 55, PAM is "systematic and coordinated activities and practices through which an organization optimally manages its physical assets, and their associated performance, risks and expenditures over their life cycle for the purpose of achieving its organizational strategic plan".

A key performance area of PAM is Asset Care Plans (ACP). ACP are maintenance strategies which are used for the improvement of asset utilization and performance. Condition Based Maintenance (CBM) is a proactive maintenance strategy and has two important aspects namely, diagnostics and prognostics. Diagnostics utilize recorded Condition Monitoring (CM) asset data to identify, detect and isolate a fault condition before failure occurs. Prognostics is used to predict when failure may occur and to estimate the remaining time left before failure occurs.

One prognostic approach utilizes statistical survival models to analyse previous failure histories (failure data), incorporating both event data and recorded CM data. This type of analysis is called statistical failure data analysis. Failure data analysis is aimed at turning vast amounts of data collected from industry into useful decision making information for the development of ACP.

Event data includes information such as asset time-to-failure and the type of failure that occurred. Time-to-failure is the duration that an asset was operational up to the occurrence of a failure. The measured duration may be done on any time-scale such as time, tonnage handled and kilometres travelled.

The type of failure included in event data serves as an indicator variable. Indicator variables distinguish between failures and suspensions. Failures can be categorized as a breakdown, when a predefined failure condition is reached or when maintenance action is taken to influence the asset survival time. A certain vibration level could be an example of a predefined failure condition. A suspension in turn, happens when an asset is taken out of service before failure occurs. Both failures and suspensions are referred to as events.

CM data are recorded measurements which can include vibration, tribology, temperature, current, etc. These measurements are referred to as covariates in the statistical failure data analysis environment. These covariates are recorded continuously or at predefined intervals. CM data and event data are often rigorously recorded (often at great expense) but seldomly analyzed to enhance decision making. This leads to the use of suboptimized maintenance strategies, not suited for complex assets that require refined and sophisticated maintenance strategies. This might lead to unplanned failures and downtime. The statistical survival model used in this paper is the Proportional Hazards Model (PHM).

The analysis involves fitting a PHM to failure data using a Weibull distribution as parametrization. The failure data used in this paper are generic data typically found in industry. This data consists of event data and CM data containing six types of covariates. To fit the Weibull PHM to the failure data, parameter estimation has to be done by optimizing a multivariate function. The purpose of the paper is to develop a metaheuristic algorithm that performs the required optimization needed to determine the Weibull PHM parameters.

The paper commences with a brief discussion on the PHM and the parameter estimation. It is followed by a discussion on the optimization problem and the Artificial Bee Colony (ABC) algorithm used in this paper. Then, an evaluation of the ABC algorithm's success is discussed.

The paper then ends with a conclusion.

2 The Proportional Hazards Model

The PHM models the Force Of Mortality (FOM) of an asset as the product of a baseline FOM (dependent on the time-scale only) and a functional term (dependent on the time-scale and the CM data). FOM is the rate of mortality of which e.g. an asset fails at a certain age. The assumption is made that the underlying baseline FOM is continuous, thus the parametric PHM is used. The Weibull distribution is used for the parametrization of the baseline FOM, due to its flexibility. For the fully parametric Weibull PHM, the baseline FOM cannot be determined independently; the distribution- and regression-parameters have to be estimated simultaneously. Vlok [1] showed that the FOM is given by:

$$h(x,\overrightarrow{z(x)}) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}\cdot\exp\left(\overrightarrow{\gamma}\cdot\overrightarrow{z(x)}\right) \tag{1}$$

which represents the fully parametric Weibull PHM. Equation (1) represents the underlying failure behaviour of the asset under study. To develop the PHM, parameter estimations have to be done. These parameters include a scale parameter, β, shape parameter, η, and a coefficient vector $\overrightarrow{\gamma}$. This vector $\overrightarrow{\gamma}$ is a row vector of six coefficients associated with the model's covariates. There are thus eight model parameters that need to be estimated. Estimation of these model parameters are done by maximizing the log-likelihood. Vlok [1] also ahowed that the log-likelihood function is given by:

$$l(\beta,\eta,\overrightarrow{\gamma}) = r\ln(\beta/\eta) + \sum_{i}\ln[(X_i/\eta)^{\beta-1}] + \sum_{i}\overrightarrow{\gamma}\cdot\overrightarrow{z_i(X_i)} - \sum_{j}\int_0^{X_j}\exp(\overrightarrow{\gamma}\cdot\overrightarrow{z_j(x)})d((x/\eta)^{\beta}) \tag{2}$$

where i indicates all failures and j indicates all events (all events include all failures and suspensions). The optimal combination of values for β, η and $\overrightarrow{\gamma}$ will maximize the log-likelihood function. The only constraint is that the β and η parameters have to be positive. Knowing the β, η and $\overrightarrow{\gamma}$ parameter values for a specific asset, one can derive useful decision making information, including the estimation of asset deterioration rates, residual life and/or the risk of operation for ACP development.

From experience, optimal parameter values are known to be within a certain value range. This value range will be referred to as the parameter value brackets. A parameter value bracket serves as a search space for each parameter. The maximum and minimum parameter bracket values are given in table 1. Large search spaces were chosen for each parameter to ensure that the optimal parameters were within the value brackets and also to test the algorithm for robustness.

Table 1: *Parameter search spaces.*

Parameter	Minimum value	Maximum value
β	0.01	5
η	2000	8000
$\overline{\gamma}$	-5	5

3 The Optimization Model

A number of different methods exist to solve non-linear problems. These include quadratic programming, fractional programming, non-linear programming, stochastic programming, calculus of variations and metaheuristics. The objective of this paper was to apply a metaheuristic solution method. It is a computational method used to find a near optimal solution for a given problem, by iteratively manipulating a previous solution. New solutions are compared to previous solutions using an acceptance criterion to determine the quality of each solution. The higher quality solution is used for the following iteration. The process ends when the solutions converge to a single solution or when a maximum number of iterations are reached.

Metaheuristic models make few or no assumptions about the problem at hand and can be used to search for an optimal solution in a large search space. A major sub-field of metaheuristics is population-based optimization techniques. These algorithms often draw inspiration from nature. Karaboga and Akay [2] stated that there are two important classes of population-based optimization algorithms which are evolutionary algorithms and swarm intelligence-based algorithms. Popular evolutionary algorithms include genetic algorithms, genetic programming, evolution strategy and evolutionary programming.

Wenping and Akay [3] indicated that Swarm intelligence (SI) is an innovative method used to solve complex optimization problems. SI is described by Bonabeau [4] as “any attempt to design algorithms or distributed problem-solving devices inspired by the collective behaviour of social insect colonies and other animal societies”. In the last two decades many SI algorithms have been proposed such as the Ant Colony Optimization, particle swarm algorithm, bacterial foraging optimization and ABC algorithm.

It was found that the ABC algorithm performed better or similar to other SI algorithms shown by Karaboga and Akay [2]. The ABC algorithm was chosen for the problem at hand for its good performance relative to other metaheuristic algorithms, its efficient multivariate problem solving ability and ease of use.

3.1 Artificial Bee Colony Optimization

The ABC algorithm is based on the foraging behaviour of honey bee colonies, proposed by Karaboga [5]. The ABC model classifies bees into three groups: employed-, onlooker- and scout bees.

The algorithm starts by finding 1000 stochastic solutions. A solution is the log-likelihood obtained with equation (2) using eight model parameter values (β,η and $\overline{\gamma}$). To find a stochastic solution, eight parameter values are generated within their respective search spaces. Generating stochastic parameter values within their search spaces is done using the equation given by:

$$p_{new_{rand}} = p_{min} + \phi(p_{max} - p_{min}) \tag{3}$$

where $p_{new_{rand}}$ indicates the new parameter value, p_{min} the minimum parameter value, ϕ a random value between [-1,1] and p_{max} the maximum parameter value, proposed by Karaboga and Akay [2].

From the 1000 stochastic solutions, the ten with the highest quality are selected for further processing. The quality of a solution is an indication of the numerical size of a solution. If a solution is larger than another, it has higher quality. The ten solutions selected from the 1000 stochastic solutions are the 10 highest quality solutions in the set of 1000. Generating a 1000 solutions and only selecting the ten with the highest quality, ensures that a thorough search of the entire search space is done exhibiting the explorative nature of the algorithm. It also serves as a tool to generate high quality initial solutions. After the stochastic set of solutions have been generated, the three search processes (represented by three types of bees) are initialized.

Ten employed bees each take one of the ten stochastic solutions and memorize the solution value and its' eight corresponding parameter values. Each of these parameter values are modified using the ABC algorithm search process. The initial solution parameter values are each modified individually by manipulating a parameter value with the following equation:

$$p_{new_{search}} = (p - r_b) + (r_b)(\phi) \tag{4}$$

where $p_{new_{search}}$ indicates the modified parameter value, p the current parameter value being modified, r_b the search radius of the particular bee and ϕ a random value between [-1,1], proposed by Karaboga and Akay [2]. This process is called a *bee dance*. Each bee does a *bee dance* with each parameter value to obtain eight modified parameter values.

These modified values are then fed into equation (2) resulting in a modified parameter solution value, which is a log-likelihood solution of equation (2). If the quality of the modified parameter solution value is higher than the current parameter solution value, the modified solution value and eight corresponding modified parameter values are memorized replacing the current solution value and its eight current parameter values. When the quality of the current parameter solution is higher than the modified parameter solution quality, the modified solution value and eight modified parameter values are deleted.

The ten stochastic solutions are thus each modified by an employed bee and the highest quality solutions and its' corresponding parameter values are memorized. These ten memorized solutions and parameter values are then sent to ten onlooker bees. Each onlooker bee takes one solution and performs the same *bee dance* with its parameter values as the employed bee, but with a smaller search radius. The reduced search radius ensures further exploitation of a local maximum. As with the employed bees, the higher quality solutions are memorized and the others deleted resulting, again, in ten solutions and their corresponding parameter values.

The author added a fourth bee to the algorithm labelled the improvement bee. As with the employed- and onlooker bee, the ten solutions found by the onlooker bees are shared with ten improvement bees. Each improvement bee takes one solution and also performs a *bee dance* with its parameter values. The search radius of the improvement bee is equal to that of the employed bee. Further discussion of the improvement bee search radius will be done later.

Scout bees then generate new stochastic parameter values with equation (3) to find ten new stochastic solutions. Generating ten stochastic solutions ensures the algorithm has an explorative nature in searching for a global maximum in every iteration. These ten stochastic solutions are then compared to the ten solutions found by improvement bees. The higher quality solutions are memorized and the other deleted resulting in ten solutions and their corresponding parameter values. These ten solutions and their corresponding parameter values are then the final ten solutions for the current iteration. These solutions and parameters are then the new initial solutions and parameters used by the employed bees in the following iteration.

This concludes the search processes of the bees. The search processes of the bees rely on the size of the search radius to find a higher quality solution with equation (4). Selecting a large search radius might result in finding the global maximum without exploiting the local maximum. A small search radius might exploit a local maximum, but will however not ensure that the global maximum is found. Consequently, selecting a search radius can be quite complex when an algorithm has to exhibit an explorative and exploitative behaviour. To solve this problem, the improvement bee search radius is set to change to ensure the algorithm exhibits explorative and exploitative behaviour. This means that the algorithm has a function that reduces the improvement bee search radius if a certain condition is met. This condition is met whenever five consecutive iterations have equal highest quality solutions.

Finally, certain tests are done to test whether an optimal solution has been found. The first test evaluates whether the improvement bee search radius is less than 10^{-20}. When the improvement bee search radius is less than 10^{-20}, it is assumed that the search radius is extremely small and improved solutions will only be found by the stochastic solutions generated by the scout bees. Finding an improved solution using this method would consume too much time as it is almost completely explorative and weak in terms of exploitation. The second test evaluates whether the total number of iterations have been reached. The maximum number of iterations of 5000 ensures that the time spent finding a solution is acceptable. If any of these conditions are met, the algorithm is stopped.

To summarize, the algorithm starts by generating 1000 stochastic solutions. Four search processes are then performed resulting in 10 solutions and corresponding parameter values. The improvement bee search radius condition is tested. Then the two stop criteria are tested. All the mentioned processes, excluding the stochastic solution generation, are repeated until one of the stopping criterions are met.

4 Evaluation of Artificial Bee Colony Optimization Success

An eight parameter log-likelihood function had to be maximized to develop ACP. Results obtained with the algorithm were found to be satisfactory compared to other algorithms such as the Nelder-Mead simplex direct search algorithm. Figure **??** illustrates how the solutions of an algorithm run converged to a solution. For this given algorithm run, a solution of -60.2119 was obtained after 1843 iterations. The average time an iteration took was 0.27 seconds while the average number of iterations to meet one of the stopping criterions was 1830 iterations. With these values an average algorithm run took approximately 8 minutes and 17.8 seconds.

To test the variance of the algorithm solutions, it was run 50 times. The 50 solutions obtained are a sample set taken out of an infinitely large population of solutions. To analyze the solutions, the 95% Confidence Interval (CI) of the sample set was determined using the equation given by:

$$\left(\overline{x} - t_{n-1,0.025}\frac{s}{\sqrt{n}}; \overline{x} + t_{n-1,0.025}\frac{s}{\sqrt{n}}\right) \tag{5}$$

where $\overline{x}$ is the sample mean, n the sample size, $t_{n,0.025}$ is found in a t-distribution table and

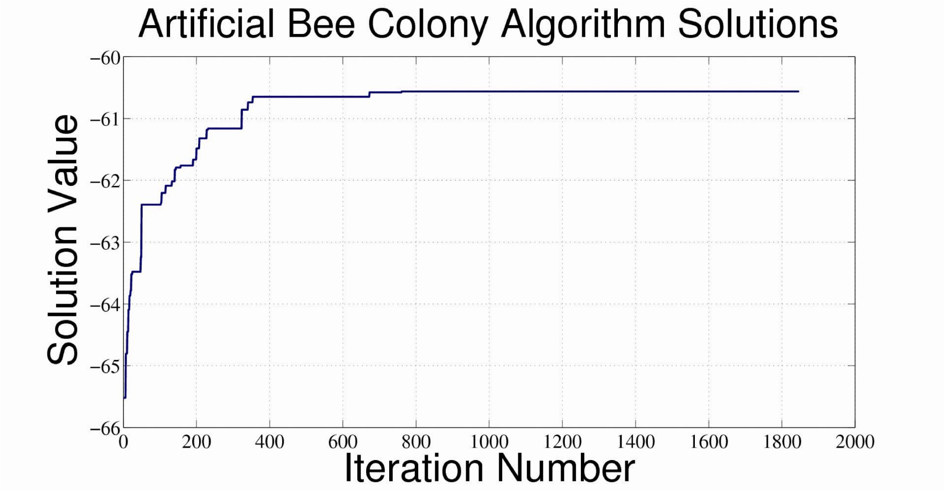


Figure 1: *ABC algorithm solution plot.*

s is the standard deviation as shown by Heiman, 2010. The sample mean was calculated to be -60.2419. $t_{49,0.025} = 2.0102$ was found by using a t-distribution table and interpolating to find the correct value. With these parameters, the 95% CI was calculated as:

$$(-60.2764; -60.2074)$$

This meant that 95% of the time, the solutions obtained with the algorithm solving this problem, should be within these values. Fortunately the variation was fairly small. However, it was necessary to determine whether this variation resulted with acceptable parameter values.

Stochastic parameter values were determined within their search spaces, to obtain a data set of a 1000 random solutions. This was done to indicate the effect random parameter values had on the solutions. It was decided to calculate the standard deviation for both sets. The standard deviation of the 1000 random solutions was calculated to be 45.6942 (10^{22}). The standard deviation of the 50 solutions was calculated to be 0.112. This was an enormous difference. Next, a F-test was also done to compare the variances in the two data sets. The two hypotheses for the F-test are as follows:

- H_0: There is no difference between the variances.
- H_A: The variances differ significantly.

H_0 is not rejected if the calculated F value was less than the critical value. With a 95% confidence level, the corresponding critical value is 0.05 (5%). The F value was calculated to be 1.63×10^{49} which is far greater than 5%. Thus H_0 was rejected which meant that there was not enough evidence to indicate that there are no difference in the two variances. Thus it can be assumed that the two data sets differ significantly. Therefore it could be concluded that the 50 sample solutions variation was insignificant and that the solutions were near optimal compared to the random data set variation. Consequently, the corresponding solution parameter values obtained using the algorithm were satisfactory.

5 Conclusion

The objective of the paper was to develop an algorithm that estimates the parameter values which maximizes a log-likelihood function in order to fit a Weibull PHM to asset failure data. The algorithm proposed by Karaboga [5], was adapted and a fourth bee was added to the algorithm. Results indicated that the developed ABC algorithm has the capability of obtaining satisfactory solutions. Variance in the solution values proved to be insignificant, not affecting the parameter values.

The runtime of the algorithm is slow compared to other algorithms such as the Nelder-Mead algorithm, but the ABC is a fairly new algorithm and has room for improvement. This indicates that the algorithm has the potential for further refinement to improve its performance. For the purpose of this paper, the ABC algorithm has shown sufficient capability and can be used in statistical survival modeling industry.

Bibliography

[1] Vlok PJ, 2002, *Vibration Covariate Regression Analysis of Failure Time Data with the Proportional Hazards Model*, Department of Mechanical and Aeronautical Engineering, Faculty of Engineering, Universty of Pretoria

[2] Karaboga D & Akay B, 2009, *A comparative study of artificial bee colony algorithm*, Elsevier, Applied Mathematics and Computation, 214, 1, pp 108–132, issn 0096-3003

[3] Wenping Z & Akay B, 2009, *A Clustering Approach Using Cooperative Artificial Bee Colony Algorithm*, Hindawi Publishing Corporation, issn 1026-0226

[4] Bonabeau E & Yunlong Z & Hanning, C & Xin S, 2010, *Swarm Intelligence: From Natural to Artificial Systems (Santa Fe Institute Studies in the Sciences of Complexity Proceedings*, Oxford University Press, USA

[5] Karaboga D, 2005, *An idea based on honey bee swarm for numerical optimization*, Techn. Rep. TR06, Erciyes Univ. Press, Erciyes

[6] Heiman G.W, 2010, *Basic statistics for the behavioral sciences*, Wadsworth Pub Co

Proceedings of the 2011 ORSSA Annual Conference
pp. 59–71
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Introducing Predictive Patient-Admission Algorithms to African Healthcare Systems

RA Daffue[*] TE Lane-Visser[†]

Abstract

Prediction algorithms, forecasting models and similar operations research (OR) tools have been practically applied in numerous service delivery sectors, with great success. Although many healthcare systems have benefitted from the practice of operations research internationally, OR's imprint on African healthcare systems is very faint. Lack of information management systems, data management processes and communications infrastructure has inhibited development in this area. Recent developments in Information and Communication Technology (ICT) and healthcare informatics have, however, rendered these barriers to entry amenable to resolution.

Health systems across Africa are fraught with resource management issues, for which acceptable, realistic solutions are required to improve service delivery. The development of Predictive Patient Admission Algorithms (PPAA), inspired by the Heritage Health Prize competition (http://www.heritagehealthprize.com), is a step towards improved management and resource utilisation in healthcare. A PPAA typically needs to accurately predict whether a patient will be admitted to hospital in the coming year and specify the estimated duration of admission, given the patient's medical history. The overall objective of a PPAA is to eliminate unnecessary hospital admissions, which, for example, resulted in healthcare expenditures of more than $30-billion in the United States in 2010.

Implementation of PPAAs can be beneficial for both healthcare providers (HCPs) (now enabled to schedule resources more efficiently and effectively to become more profitable organisations) and society in general (HCPs are enabled to intercede with patients early on, providing the required medical attention as a preventive measure, ultimately resulting in healthier societies.

The focus of this paper is to explore the feasibility of implementing PPAAs in African healthcare systems. The effect of organisational cultures, the state of information and communication technology (ICT), the role and requirements of management structures, as well as the associated potential benefits of the implementation and operation of PPAAs will be addressed in this paper.

Key words: Developing countries, health services, preventive care, prediction and forecasting

[*] Corresponding author: Stellenbosch University, South Africa, email: daff@sun.ac.za
[†] Stellenbosch University, South Africa, email: tanyav@sun.ac.za

1 Introduction

1.1 Background

African Healthcare Systems are fraught with resource allocation problems [23], ineffective management structures [1] and constrained innovation capabilities [6]. These problems are often aggravated and overshadowed by major healthcare pandemic issues, such as the management and treatment of HIV, tuberculosis, malaria and various other infectious diseases. Akta et al. [3] congruently found that the effective utilisation of limited resources in healthcare providers is a critical problem for healthcare management in developing countries, therefore *these issues should not be considered minor when national health reforms and programs are discussed* [1].

The World Health Organisation (WHO) believes that incompetent management may be the main bottleneck in the implementation and operation of new technologies and OR models in African Healthcare Systems [22]. Some of the issues identified include poor adherence to policy recommendations, poor record-keeping and reporting of information, poor information dissemination, ethical issues, interaction or competition with other interventions for other diseases, as well as marketing and advocacy of policies and regulations. These problems articulate the critical role that management plays, not only in the successful introduction of new healthcare technologies, but in the application of OR models that functionally utilise these technologies.

1.2 Predictive Patient-Admission Algorithms

Predictive Patient-Admission Algorithms (PPAAs) are concept based OR models, forming part of a strategy to reform the healthcare system of the United States. The algorithms *predict* and *identify* which individuals (from a given patient population) will be admitted to hospital in the near term, whilst specifying the patient's estimated duration of admission or *length of stay* (LOS). If this is achieved, healthcare providers (HCPs) can intercede with these individuals decreasing their probability/risk of admission and congruently schedule the HCPS resources according to the expected demographic of patient-admissions [12].

The algorithms are based on computationally-intensive ensemble methods of prediction and data mining to enable the management of big, complex datasets, as is the case with patient-data banks. The Random Forest™ model, designed by Leo Breiman and Adele Cutler in 2006, is an example of such ensemble methods of forecasting. These models enable thediscovery of useful predictive information from large data sets (number of entries greater than 1 million) in unstructured problems (number of variables greater than 1000) [8]

Figure 1 depicts a framework for the implementation of PPAAs in existing healthcare systems.

In Figure 1, the process is initiated by the patients entering the system (1). The patients can either be new patients (no health records exist) or current patients at the HCP (health records exist). The following patient-admissions possibilities exist; new patient with an unknown cause/illness, current patient with an unscheduled appointment with an unknown cause, current patient with an unscheduled appointment with a known illness, current patient with a scheduled appointment for a known illness or current patient with a scheduled

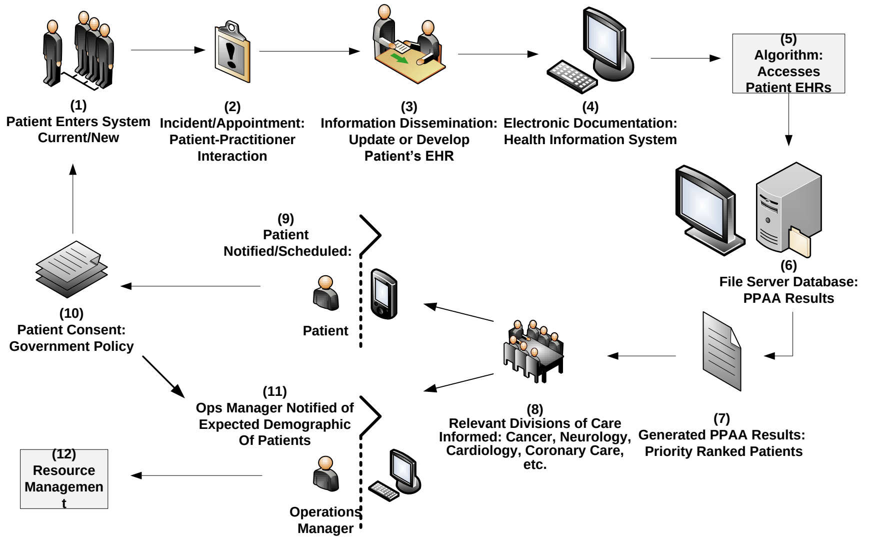


Figure 1: *Predictive Patient Admission Algorithms: Process Specification*

appointment with an unknown illness (2). These interactions result in new patient data that need to be disseminated to the patient's Electronic Health Record (EHR) (3). The EHRs are managed on the HCP's Health Information System (HIS) (4) to ensure that the records are easily attainable and in a suitable format as required by the PPAA. The PPAA executes the required mathematical permutations to identify potential hospital admissions from the given patient population (5) and the results are locally stored on the File Server Database (6) of the HCP and updated on a preference scheduled basis. The predicted patient-admission results are distributed to the individual divisions of care (8), which subsequently determine their capacity to provide preventive care to the identified patients from their division. The priority ranked patients are then notified of their risk of admission and an appointment is scheduled (9), providing that the patient gives his or her consent (10). The Operations Manager, responsible for resource management (RM) at the HCP, is notified of the expected demographic of patients the hospital will be able and unable to intercede with in the following year (11). This information is utilised by the Operations Manager to improve the RM capabilities of the HCP (12).

The feasibility of applying PPAAs in developed countries is beyond doubt. With the availability of Electronic Health Records (EHRs), Information and Communication Technology (ICT) infrastructure, Healthcare Informatics and competent management, overcoming the potential barriers to implementation of PPAAs in these countries, is possible. In the African context, however, these barriers are more prominent and greatly influence the feasibility of implementing PPAAs.

This paper serves as an indication of the extent to which healthcare systems in Africa, ***institutionally*** and ***operationally***, have to be adapted to support the use of computationally intensive OR solutions such as PPAAs. The benefits of and barriers to the implementation of Predictive Patient-Admission Algorithms in Africa are established as well as the differences between the application of PPAAs in first and third world health systems, which aid in the contextualisation of applying PPAAs in Africa, are established. Furthermore, a framework for successfully implementing PPAAs in Africa is discussed, illustrating the key role-players and their responsibilities required to seamlessly introduce PPAAs into typical African healthcare settings.

2 Potential Benefits of implementation

Healthcare systems are complex and depend on various economic, structural, and organisational factors and the interdependencies between these factors. When analysing the efficiency and effectiveness of a healthcare system, two factors are considered key performance indicators (KPIs): quality of service and patient waiting time [3].

PPAAs narrowly relate to both these measures; the implementation of PPAAs aims to assist healthcare providers with forecasted information regarding the expected behaviour of patients from the Healthcare Provider's patient database. The results generated by the PPAAs specify the expected cause of admission, the estimated duration of admission and the patient's medical history, enabling HCPs to make informed decisions with regards to healthcare resource management. For example, a local Healthcare Provider in Bloemfontein, implementing a PPAA, predicts that 150 patients are estimated acute cardiovascular patients, which, if admitted to hospital, would spend on average 10 days in hospital in the following year. The patients' identities are known, identifiable through Patient-Identification Numbers (PINs), and the medical background of the patient is readily available on the HCP's patient database. However, due to capacity constraints, the HCP is not able to provide the necessary preventive care to each identified patient; hence the patients are prioritised in the most ethical manner possible (all patients equal, worst-off first, maximising benefits or social responsibility). At this moment the HCP has the ability to intercede with the priority patients, ultimately reducing these patients' risk of admission and, in addition, effectively schedule cardiovascular medication, treatments, cardiothoracic surgeons, nurses, hospital beds, laboratory tests, etc. in accordance with the expected number of "unavoidable" cardiac admissions in the following year. If this is achieved, improvements in quality of service (due to the preventive care strategy) and reductions in service waiting time (due to more efficient healthcare resource management) are examples of the likely benefits.

The preventive care strategy focuses on establishing better patient-practitioner relationships, through the provision of patient-tailored care (higher quality of service) on a timely basis. Initially, however, yearly forecasts might not have this desired effect, but as the system progresses, half-year or even monthly forecasts will enable HCPs to *segment patients into priority and time based categories, that will determine each patient's scheduled frequency of interaction with the HCPs, preventing patients from becoming "lost" in the system. Furthermore, Merkin (2011) states that duplicated healthcare services, service starvation and service excess are common problems in healthcare systems worldwide* and *formed part of the initial rationale to develop Predictive Patient-Admission Algorithms. Leaner supply chains with less occurrences of service misalignment are expected to result from the implementation of PPAAs.*

Christensen et al. [5] questioned the various strategies for implementing new inventions in healthcare and found technological innovations to be more effective when the strategy was rather to "Invest less money in high-end, complex technologies and more in technologies that simplify complex problems." *OR models typically have the ability to solve complex problems without the necessary use of high-end technologies, which, in essence, resembles the beneficial functionality of PPAAs: these algorithms aid in the simplification of the complex problems associated with healthcare resource management.*

3 Potential barriers to implementation

Computationally-intensive methods of prediction and data mining pose various challenges for the developing countries in Africa and, if not resolved, these entry barriers can restrict the implementation of such tools entirely.

Certain technological prerequisites are essential for the functioning of PPAAs in any healthcare organisation. The *implementation framework* in Figure 1 illustrated the generic processes in the application of PPAAs and subsequently enables the identification and documentation of the *operational and institutional requirements* for the successful operation of PPAAs in any healthcare setting. These requirements include human resources, information and communication technologies (ICT), electronic health records (EHRs), health information management systems, financial resources, regulatory legislation and suitable patient profiles.

3.1 Human resources

Human Resources (HR) have to be allocated to the operation of the PPAAs, as patient data continuously changes, affecting the results generated by the PPAAs. Hence, the establishment of a permanent preventive care team (PCT), responsible for the daily management and operation of the PPAA, is advised. The PCT will need to undergo training with regards to the application and management of the PPAAs, but since the operability of PPAAs is still in the development phase, the extent of personnel training is largely unknown. In United States the ratio of management and support workers in healthcare per 1000 Population is 24.76 respectively, which is substantially better compared to that of South Africa (0.51), Nigeria (2.507) and Egypt (0.07) [11].

Managerial issues have been inadmissible in various efforts to transform the public health systems in Africa, especially with the introduction of Health Information Systems and the related ICTs, as the case with the Nigerian National Health Management Information System (NHMIS) in 2003 demonstrated. The Nigerian government initiated programs to transform their knowledge and information management systems and introduced e-documentation as part of their strategy to implement a new National Health Management Information System. Akande and Monehin [2] studied the effect of management and government policy in the effectiveness of the NHMIS. A survey was conducted among 37 active ***private*** clinics in Nigeria to determine the awareness and level of involvement of these clinics towards the NHMIS. 73% of the respondents were medical directors (MDs) and 67.6% of these MDs were *aware* of the NHMIS. The survey also found that a miniscule number of the private clinician respondents indicated that they were actively submitting patient information to the NHMIS. Therefore, without the support of private healthcare providers, which accounts for 34% of the Nigerian population's healthcare [16], a significant amount of patient information at private healthcare providers is lost, seriously compromising the effectiveness of the NHMIS, articulating the need for well-led management and regulating government policies in these and related project executions [2].

Therefore, the lack of competent human resources seriously affects the efficiency and effectiveness of OR models in Africa. The absence of quality healthcare managers, doctors, nurses and support workers will remain a major barrier to the implementation of OR tools, until efforts are initiated to educate and train these individuals to a globally acceptable standard.

3.2 Information and Communication Technology (ICT)

The role of ICT in the application of PPAAs, relates to the communication of patient information to the relevant healthcare decision-makers (Relevant Divisions of Care and Government) and priority ranked patients. Furthermore, the use of ICT will play an intricate part in the supply of real-time data to the preventive care clinician or individual responsible for the distribution and communication of PPAA results.

ICT based healthcare projects, like the Africa Health Infoway, *have been tried and tested by the WHO and aim to supply district managers and health workers with real time data to monitor and evaluate programs and healthcare resources, whilst enabling the general public to make informed decisions with regards to local health programs* [19]*. Although the Africa Health Infoway is still in the developmental phase, it is ICT based projects like these, funded by governments and academia, that could increase the opportunity for OR models and PPAAs to be successfully implemented in African healthcare systems.*

Various African countries have initiated efforts to implement ICT (Figure 2) [15]. *Although these technologies are seemingly available to Africa, very few ICT applications have made it passed the pilot implementation phase. Therefore the use of ICT, adapted for efficient use in healthcare systems, remains a challenge in Africa.*

PPAAs demand the use of readily available patient data. Electronic Health Records (EHRs) are longitudinal electronic records of patient health information generated by one or more patient-practitioner encounters in any care delivery setting [17]. Patient EHRs should constitute patient demographics, causes of admission, progress notes, and vital stats, past medical history, immunisations, laboratory data and claims data. In the process depicted in Figure 1, EHRs are essential in the provision of patient data to PPAAs, as EHRs provide the most efficient platform for the retrieval of- and access to patient information.

In Africa, patient data is rarely documented. If a patient record exists, it would typically be in the form of a paper-chart or file. These traditional methods of data documentation will constrain the use of PPAAs as computationally intensive operations require electronically stored data, to enable the extensive patient data analyses required for the generation of the PPAA's results. *A few East-African countries have implemented EHRs and professed they are enabling* success factors in the provision of higher quality care to a broader base of patients [18].

Apart from the limited use of EHRs in Africa, *poor record-keeping* & *reporting of medical information* as well as *inadequate information dissemination* increasingly inhibits the application of OR models, such as PPAAs, in African Healthcare Systems [22]. Therefore, the availability of patient data in the form of EHRs is currently less than desirable.

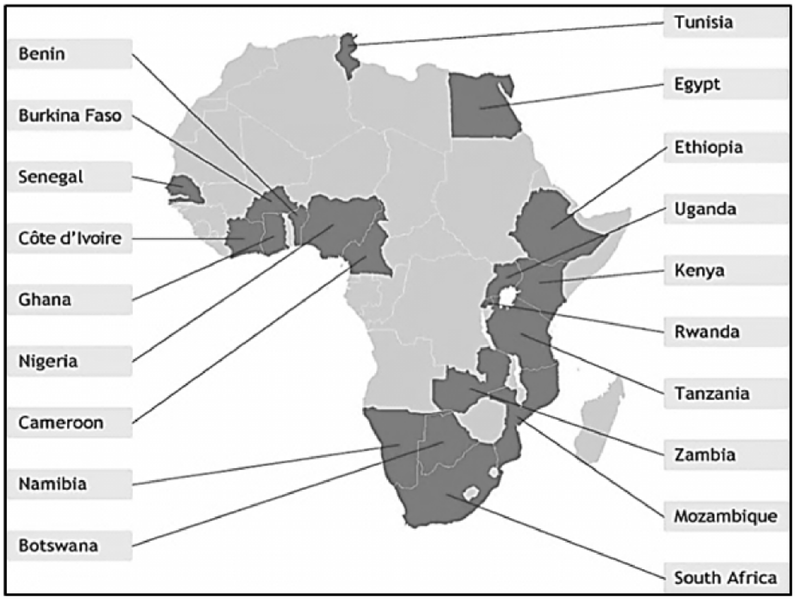


Figure 2: *African Countries Implementing ICT*

Health Information Management Systems integrate data collection, processing, reporting, and use of the information necessary for improving health service effectiveness and efficiency through better management at all levels of health services [21] PPAAs will require a revision of the current system of health information management, in order to effectively manage, protect and update relevant patient records.

Developed countries' health systems have evolved to such an extent that comprehensive patient data analyses are executable as a result of the implementation of Electronic Health Records (EHRs). Patient population datasets of more than 1-million individuals are electronically documented and managed with the availability of reliable Health Information Management Systems and knowledge management practices.

As mentioned in the Nigerian Health Information Management System, the primary inhibiting factor (incompetent management) resulted in various secondary problems, such as the lack patient data documentation and dissemination. Therefore, the use of these technologies is currently inhibited by various factors that mainly relate to incompetent management structures.

3.3 Financial resources

The Heritage Provider Network currently has full ownership of the developed algorithms [13] and no procurement costs have yet been established. Costs will however be incurred to align the Healthcare Provider's Health Information Management System and Electronic Health Records with the data requirements of the PPAA. Essentially, the magnitudes of these costs depend on the size of the respective healthcare systems. The costs associated with training and employing preventive care clinicians are also considered variable costs and need to be established by the HCP.

The implementation of ICT infrastructure, healthcare informatics and health information systems are not seen as costs solely incurred or associated with the implementation of PPAAs, as these technologies form part of various other healthcare reform strategies. However, funds for ICT, EHRs and the related healthcare technologies should be controlled locally. Furthermore, Academic Partnerships could be established to leverage research funds and government support should be effected to subsidise the technological infrastructure for PPAAs.

Although African countries spend substantially lower fractions of their gross domestic product (GDP) on healthcare compared to developed countries (South Africa and the United States spent 8.6% and 15.4% of their GDP on healthcare respectively in 2004 [14], studies have shown that the majority of African countries *do* possess the financial capability to transform their public health systems with quality Health Information Systems and the associated Information and Communication Technologies [1].

3.4 Regulatory legislation

The necessity of regulatory legislation and competent management in the governance of healthcare data management and information dissemination is essential to ensure optimal efficiency of the systematic implementation and operation of PPAAs.

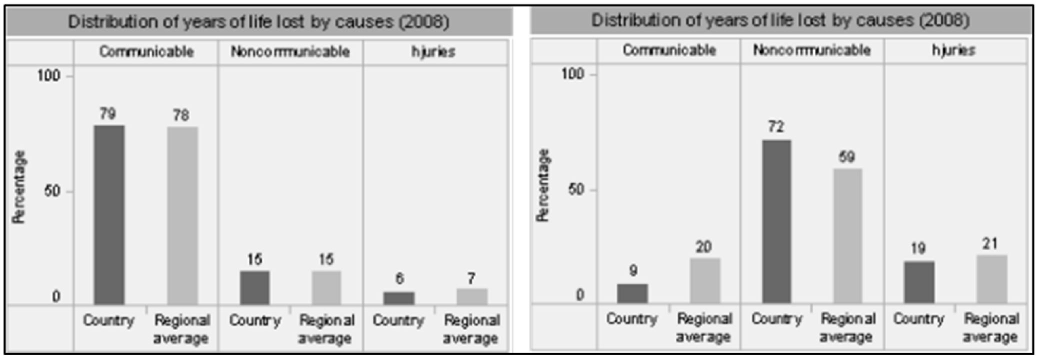
Various limitations on the use of patient information exist in most countries, emphasising the critical role government policies play in the regulation of PPAAs and the associated data management practices. In South Africa, for example, the National Health Act 61 of 2003: Chapter 2(16) specifies that healthcare providers may examine user health records for the purpose of:

1. Treatment ***with the sole authorisation of the user***; and
2. Study, teaching or research only ***with the authorisation of the user, head of the health establishment concerned and the relevant health research ethics committees***

The South African National Health Act specifies that national and provincial governments have concurrent legislative authority, enabling provinces to legislate on provincial health issues in terms of the Constitution [4]. Therefore, the process of applying PPAAs in the Western Cape, as opposed to Mpumalanga, might differ with regards to the individual provincial health legislation, complicating the implementation of PPAAs. However, with the introduction of policies to officiate and promote information management practices in healthcare, various African governments, including South Africa, have committed to the transformation of their public health systems.

3.5 Suitability of patient profiles

Non-communicable diseases (NCDs) such as cholesterol, diabetes and cancer are on average responsible for 78% of mortalities in the United States each year, compared to 15% in Southern Africa [11]. Furthermore, 72% of mortalities in Southern Africa are attributable to communicable diseases such as HIV, Malaria and Tuberculosis (Figure 3) [11]. As NCD patients illustrate patient-behaviour trends that are predicted with greater accuracy as opposed to patients with communicable diseases (CDs) [20], the *potential impact* of PPAAs in Africa are compromised by the average patient profile that exists on the continent.



(a) South Africa

(b) United States

Figure 3: *State of mortality rates in a) South Africa and b) America*

4 Platform for Implementing Predictive Patient-Admission Algortihms

PPAAs should form part of healthcare systems in its totality, with key role-players at different levels of the healthcare system hierarchy. A generic platform for the implementation of PPAAs will aid healthcare providers in their effort to introduce PPAAs into their organisations. Figure 4 depicts a hypothetical platform for PPAA implementation. The figure identifies the essential role-players in the implementation and operation of PPAAs and the hierarchical relationship between these entities in any healthcare system.

4.1 Government Policy and Regulations

Eysenbach [7] stated that the ideal integration of healthcare technologies and dependencies should incorporate government policy (to curb perverse incentives). Additionally, Friede et al. [9] emphasised the critical roles that government and academic public health leaders play in public healthcare and the necessity of government's commitment to public health technologies and informatics, for successful implementation in public health systems.

In essence, governments need to regulate HCPs to the extent that data collection is uniform and compulsory at all HCPs. The policies should also incorporate incentives for adherence to the regulations. If this is achieved, the application of PPAAs can form part of national healthcare reforms, which will improve the efficiency of the implementation of PPAAs.

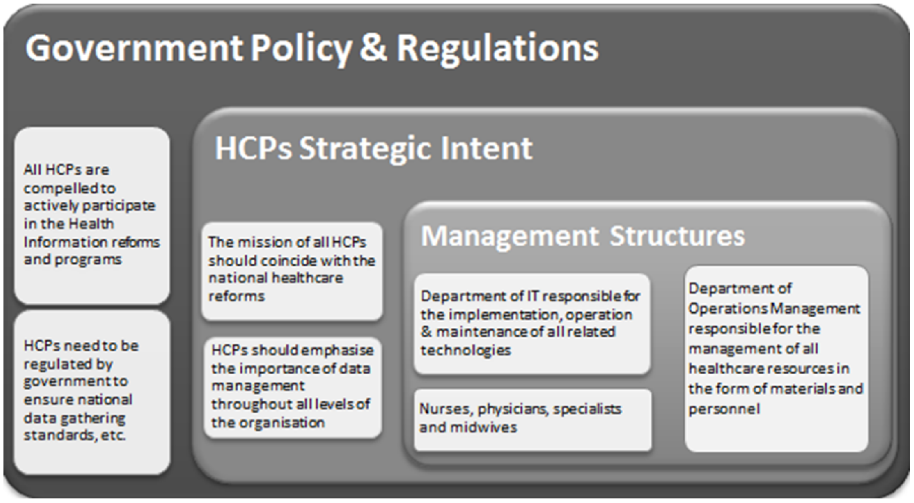

Government Policy & Regulations
All HCPs are compelled to actively participate in the Health Information reforms and programs
HCPs need to be regulated by government to ensure national data gathering standards, etc.
HCPs Strategic Intent
The mission of all HCPs should coincide with the national healthcare reforms
HCPs should emphasise the importance of data management throughout all levels of the organisation
Management Structures
Department of IT responsible for the implementation, operation & maintenance of all related technologies
Nurses, physicians, specialists and midwives
Department of Operations Management responsible for the management of all healthcare resources in the form of materials and personnel

Figure 4: *Hypothetical platform for PPAA implementations*

4.2 Strategic Intent of Healthcare Providers

The mission and vision of an HCP should be aligned with government policy and regulations, to establish and foster an environment which commits to the transformation from paper-based to electronically driven organisations; aspiring to make better informed decisions with the aid of quality data management processes and analyses. Organisational change is predicated on managerial leadership and support, and these elements are critical for successful implementation of quality improvement in healthcare [10].

Furthermore, the strategic intent of HCPs might differ in their vision and mission, resulting in a unique alignment between the strategic intent of the PPAA and the individual HCPs respectively. The onus lies on key decision-makers in the organisation to drive the re-alignment of the HCP's strategic intent.

4.3 Management Structures

The role of management in healthcare has become increasingly critical in the past few decades, as the inter-dependencies of relevant departments in a healthcare provider have become more complex in the information age [10]. As PPAAs affect various divisions and management structures within HCPs well-integrated management structures is crucial for the optimal efficiency of PPAAs, requiring the coordination of various departments.

The management structures in Figure 4 identify the key role-players that directly affect the data required and results generated by PPAAs. Doctors, physicians, nurses and midwives interact with patients and are typically responsible for gathering patient data. The data is then disseminated across the organisation with the use of ICT, HIMS and EHRs; technologies managed by the Information Technology (IT) department. Lastly, the PPAA results need to be communicated to the Department of Operations Management to effect the potential Resource Management benefits. These relationships illustrate the necessity of a horizontally integrated organisation, aligned with the strategic intent of the HCP and regulating legislation.

5 Evaluating the Barriers to Implementation

Based on the discussion in the preceding chapters, the various barriers to entry for PPAAs in South African healthcare systems have been quantified and the resulting scores are presented in Table 1. The barriers were weighted according their necessity in the application of PPAAs. Scores were concurrently allocated and each barrier was ranked. The values found in column two (R.S.A SCORE) had to be normalised, to correctly calculate a weighted score.

Table 1: *Evaluating the barriers to implementation*

Barrier	Necessity (out of 10)	R.S.A Score (10 Score Out of 10)	Weighted Score	Rank
Availability of Competent Management	8	07-Aug	56-64	1
Availability of EHRs and HISs	10	05-Jun	50-60	2
Suitability of Patient Profiles	9	05-Jun	45-54	3
Governments Commitment	6	06-Jul	36-42	4
Availability of ICT Infrastructure	7	04-May	28-35	5
Availability of Financial Resources	7	02-Mar	14-21	6

6 Conclusions

Various studies have proven that the main inhibiting barriers to the implementation of OR models in developing countries, first and foremost, relate to inadequate management structures and corporate governance. These issues need to be addressed if PPAAs are to be embraced by the African health sector.

The main differences in the application of PPAAs in a developing country as opposed to a developed country, relate to three factors: the average patient-profile, the extent to which healthcare technologies have been implemented and the availability of competent management staff. Currently, these disparities greatly affect the suitability of African healthcare systems in the adoption of OR models such as the PPAA. The barriers to implementation raise legitimate questions regarding the extent to which PPAAs can be introduced and ultimately benefit healthcare systems in Africa. The major barriers currently, relate to the availability of competent management and world-class EHRs and HIMS. Effectively addressing these barriers will positively affect the feasibility of applying PPAAs in Africa.

In summary, the identification and prioritisation of the main inhibiting barriers serves as an indicator for potential investors and developers of the PPAA and similar OR tools. African countries need to assess their performance with regards to these barriers and institute the necessary change required, to potentially effect the numerous benefits offered by PPAAs, consequently changing healthcare as we know it in Africa.

Bibliography

[1] African Union Conference of Ministers (2007). *African Health Strategy: 2007 - 2015.* Addis

Ababa: African Union.

[2] Akande, T. M. and Monehin, J. O. (2004). Health management information system in private clinics ilorin, nigeria. *Nigerian Medical Practitioner*, 46:130–146.

[3] Akta, E., lengin, F., and ahin, . . (2007). A decision support system to improve the efficiency of resource allocation in healthcare management. *Socio-Economic Planning Sciences*, 41:130–146.

[4] Carstens, P. and Pearmain, D. (2007). *Foundational Principles of South African Medical Law*. Durban: LexisNexis.

[5] Christensen, C. M., Bohmer, R., and Kenagy, J. (2000). Will disruptive innovations cure health care? *Harvard Business Review (October)*, pages 102–113.

[6] Dahlman, C. (2006). Innovation in africa - some basic concepts. In *Innovation in the African Context - A Forum for Policymakers*, pages 1–24. Dublin: Georgetown University.

[7] Eysenbach, G. (2000). Consumer health informatics. *BMJ*, 320:1713–1716.

[8] Fildes, R., Nikolopoulos, K., Crone, S. F., and Syntetos, A. A. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society*, 59:1150–1172.

[9] Friede, A., Blum, H. L., and McDonald, M. (1995). Public health informatics: How information-age technology can strengthen public health. *Annual Review Public Health*, 16:239–252.

[10] Glickman, S. W., Baggett, K. A., Krubert, C. G., and Peterson, E. D. (2007). Promoting quality: the health-care organisation from a management perspective. *International Journal for Quality in Health Care*, 19:341–348.

[11] Global Health Observatory (2011). Global health observatory data depository - health workforce - density per 1000. Retrieved August 28, 2011, from Global Health Observatory Data Depository: http://apps.who.int/ghodata/?vid=92100.

[12] Heritage Provider Network (2011a). Heritage provider network - health prize. Retrieved March 4, 2011, from Heritage Provider Network : http://www.heritagehealthprize.com/.

[13] Heritage Provider Network (2011b). Heritage provider network - health prize - leaderboard. Retrieved June 10, 2011, from Heritage Provider Network - Health Prize: http://www.heritagehealthprize.com/c/hhp/Leaderboard.

[14] NationMaster (2011). Nationmaster.com - statistics - total expenditure on health as gdp by country. Retrieved June 27, 2011, from NationMaster.com: http://nationmaster.com/graph/hea_tot_exp_on_hea_as_of_gdp-health-total-expenditure-gdp.

[15] Research ICT Africa (2011). researchictafrica.net - about. Retrieved June 20, 2011, from researchICTafrica.net: http://www.researchictafrica.net/about.php.

[16] Soyibo, A., Olaniyan, O., and Lawanson, A. O. (2009). *National Health Accounts of Nigeria: 2003-2005, Volume 1.* National Health Accounts of Nigeria: 2003-2005.

[17] The MITRE Corporation (2006). *Electronic Health Records Overview.* McLean, Virginia: NIH National Center for Research Resources.

[18] Tierney, W. M. (2010). Experience implementing electronic health records in three east african countries. *Studies in Health Technology and Informatics*, pages 371–375.

[19] UNIDO (2011). Unido nano - projects - africa health infoway. Retrieved June 26, 2011, from Unido Nano - International Centre of Nanotechnology: http://unidonano.org/inner.php?page=projectPageInfo&projectCode=4.

[20] Walker, J. D., Teare, G. F., Hogan, D. B., Lewis, S., and Maxwell, C. J. (2009). Indentifying potentially avoidable hospital admissions from canadian long-term care facilities. *Medical Care*, 47:250–254.

[21] World Health Organisation (2004). *Developing Health Management Information Systems - A practical guide for developing countries.* Manilla, Phillippines: World Health Organisation.

[22] World Health Organisation (2008). *Framework for Operations and Implementation Research in Health and Disease Control Programs.* Global Fund.

[23] Zere, E. (2006). Technical efficiency of district hospitals: Evidence from namibia using data envelopment analysis. *Cost Effectiveness and Resource Allocation*, 4(5):1–9.

Proceedings of the 2011 ORSSA Annual Conference
pp. 72–80
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Finding the optimal South African freight energy management strategy with Archived Multiobjective Simulated Annealing (AMOSA)

TE Lane-Visser* MJWA Vanderschuren

Abstract

The development of an optimised freight energy management strategy is a highly complex affair. The range of and variance between proposed freight energy management measures are vast. On one end of the spectrum there is the maintenance of adequate tyre pressure to reduce fuel consumption in trucks and on the other there is modal restructuring of the entire freight sector, for example. A freight energy management strategy consists of a particular combination of these measures and can vary in terms of the inclusion and exclusion of certain measures and in terms of the level of implementation planned for each included measure. There is, thus, an infinite amount of potential strategies that can be mathematically formulated. Furthermore, the management of energy use in freight transportation has many stakeholders with widely differing objectives and vested interests. Some of these objectives are positively correlated, some negatively correlated and some completely uncorrelated to each other. Different stakeholder objectives are grouped together and combined to form four distinct problem objectives: minimising overall energy use by the freight sector, minimising the negative environmental and social impacts incurred due to strategy implementation respectively and maximising the resulting positive impact on the economy. Finding a single optimal solution (energy management strategy) that simultaneously completely satisfies all of the stakeholder objectives is not deemed possible. The aim of this paper is to introduce Archived Multiobjective Simulated Annealing (AMOSA) as a solution approach capable of dealing with the full complexity of the problem, including the requirement for multiple objective optimisation. The proposed method for translating the model output into practically useful freight transport energy management strategies is explained.

Key words: Metaheuristics, multiple objective programming, simulated annealing, decision support systems, energy sector, freight transportation.

*Corresponding author: Stellenbosch University, South Africa, email: tanyav@sun.ac.za

1 Introduction

Transportation is an essential component of modern civilisation. Without freight transportation, the economy would grind to a halt. Unfortunately, the privilege of having an extensive freight transportation network comes at high a price. The transportation sector is the fastest growing source of greenhouse gas (GHG) emissions in South Africa, and the second largest source of GHG emissions in the world, accounting for 13% of the total global emissions [11]. This can mainly be ascribed to high levels of dirty energy use within the sector. Further to this, the sector consumes 27% of the total amount of energy, 78% of all liquid fuels and 1.6% of electricity in the country [4]. The sector is 98% dependent on crude oil based liquid fuels (approximately 70% of crude oil used in South Africa is imported) [4, 5]. The energy demand from non-renewable sources for freight transportation accounts for about 40% of total transport energy demand [9].

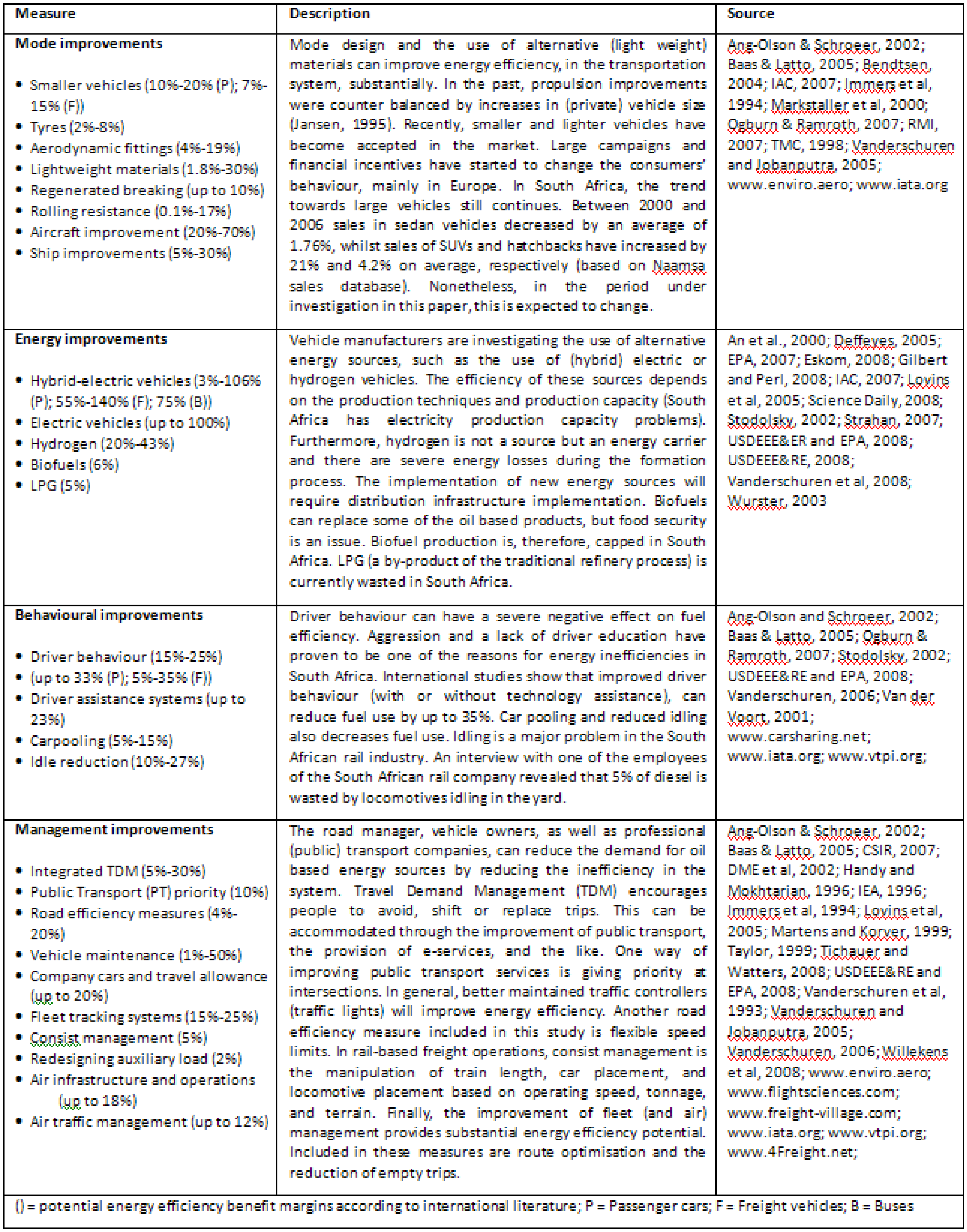
With resource availability under threat and environmental concerns becoming all the more urgent, globally, it is unavoidable that transportation systems cannot continue to operate in this fashion for much longer. Transport managers will have to find ways of delivering the same levels of mobility and connectivity, whilst requiring far less non-renewable resources in the near future. Fortunately, a range of potential transport energy demand mitigation alternatives are already being researched and proposed. These alternatives can be generally classified into three groups, namely avoidance measures, shifting measures and improvement measures. Avoidance measures include all measures that aim to remove excessive and wasteful transportation demand from the system. Shift measures, in turn, propose changes in the composition and distribution of demand in the freight system. Improvement measures are the most prevalent and encompass all measures that improve the energy demand of the current system components in operation today. Table 1 provides a summary of the most widely recognised transport energy management alternatives to date.

It is important to note that there is significant variation in the impact and complexity levels, as well as the application areas and expected lead times of measures. Despite there being quite a large body of literature on such measures, very few documents discuss the combination of measures into balanced (freight) energy management strategies. This paper forms part of a larger project that aims to develop a decision support model informing on the composition of the best freight energy strategies. The specific scope of this paper is limited to discussing the solution approach envisioned for use in the project and how this solution algorithm could be applied to the problem at hand. An investigation of the definition and classification of a good strategy is discussed in section 2. This is followed by an exposition of the nature and structure of the problem and a translation of the problem into modelling terms (section 3). Section 4 introduces the Archived Multiobjective Simulated Annealing (AMOSA) algorithm and establishes why this is the proposed solution approach. The paper concludes in section 5.

2 Definition of a good freight energy management strategy

Freight transport is intertwined with the South African economy. It is of the utmost importance that the sector is managed properly. Any future development strategies should err on

Table 1: *Transport efficiency measures and their potential impact*



Measure	Description	Source
Mode improvements • Smaller vehicles (10%-20% (P); 7%-15% (F)) • Tyres (2%-8%) • Aerodynamic fittings (4%-19%) • Lightweight materials (1.8%-30%) • Regenerated breaking (up to 10%) • Rolling resistance (0.1%-17%) • Aircraft improvement (20%-70%) • Ship improvements (5%-30%)	Mode design and the use of alternative (light weight) materials can improve energy efficiency, in the transportation system, substantially. In the past, propulsion improvements were counter balanced by increases in (private) vehicle size (Jansen, 1995). Recently, smaller and lighter vehicles have become accepted in the market. Large campaigns and financial incentives have started to change the consumers' behaviour, mainly in Europe. In South Africa, the trend towards large vehicles still continues. Between 2000 and 2006 sales in sedan vehicles decreased by an average of 1.76%, whilst sales of SUVs and hatchbacks have increased by 21% and 4.2% on average, respectively (based on Naamsa sales database). Nonetheless, in the period under investigation in this paper, this is expected to change.	Ang-Olson & Schroeer, 2002; Baas & Latto, 2005; Bendtsen, 2004; IAC, 2007; Immers et al, 1994; Markstaller et al, 2000; Ogburn & Ramroth, 2007; RMI, 2007; TMC, 1998; Vanderschuren and Jobanputra, 2005; www.enviro.aero; www.iata.org
Energy improvements • Hybrid-electric vehicles (3%-106% (P); 55%-140% (F); 75% (B)) • Electric vehicles (up to 100%) • Hydrogen (20%-43%) • Biofuels (6%) • LPG (5%)	Vehicle manufacturers are investigating the use of alternative energy sources, such as the use of (hybrid) electric or hydrogen vehicles. The efficiency of these sources depends on the production techniques and production capacity (South Africa has electricity production capacity problems). Furthermore, hydrogen is not a source but an energy carrier and there are severe energy losses during the formation process. The implementation of new energy sources will require distribution infrastructure implementation. Biofuels can replace some of the oil based products, but food security is an issue. Biofuel production is, therefore, capped in South Africa. LPG (a by-product of the traditional refinery process) is currently wasted in South Africa.	An et al., 2000; Deffeyes, 2005; EPA, 2007; Eskom, 2008; Gilbert and Perl, 2008; IAC, 2007; Lovins et al, 2005; Science Daily, 2008; Stodolsky, 2002; Strahan, 2007; USDEEE&ER and EPA, 2008; USDEEE&RE, 2008; Vanderschuren et al, 2008; Wurster, 2003
Behavioural improvements • Driver behaviour (15%-25%) • (up to 33% (P); 5%-35% (F)) • Driver assistance systems (up to 23%) • Carpooling (5%-15%) • Idle reduction (10%-27%)	Driver behaviour can have a severe negative effect on fuel efficiency. Aggression and a lack of driver education have proven to be one of the reasons for energy inefficiencies in South Africa. International studies show that improved driver behaviour (with or without technology assistance), can reduce fuel use by up to 35%. Car pooling and reduced idling also decreases fuel use. Idling is a major problem in the South African rail industry. An interview with one of the employees of the South African rail company revealed that 5% of diesel is wasted by locomotives idling in the yard.	Ang-Olson and Schroeer, 2002; Baas & Latto, 2005; Ogburn & Ramroth, 2007; Stodolsky, 2002; USDEEE&RE and EPA, 2008; Vanderschuren, 2006; Van der Voort, 2001; www.carsharing.net; www.iata.org; www.vtpi.org;
Management improvements • Integrated TDM (5%-30%) • Public Transport (PT) priority (10%) • Road efficiency measures (4%-20%) • Vehicle maintenance (1%-50%) • Company cars and travel allowance (up to 20%) • Fleet tracking systems (15%-25%) • Consist management (5%) • Redesigning auxiliary load (2%) • Air infrastructure and operations (up to 18%) • Air traffic management (up to 12%)	The road manager, vehicle owners, as well as professional (public) transport companies, can reduce the demand for oil based energy sources by reducing the inefficiency in the system. Travel Demand Management (TDM) encourages people to avoid, shift or replace trips. This can be accommodated through the improvement of public transport, the provision of e-services, and the like. One way of improving public transport services is giving priority at intersections. In general, better maintained traffic controllers (traffic lights) will improve energy efficiency. Another road efficiency measure included in this study is flexible speed limits. In rail-based freight operations, consist management is the manipulation of train length, car placement, and locomotive placement based on operating speed, tonnage, and terrain. Finally, the improvement of fleet (and air) management provides substantial energy efficiency potential. Included in these measures are route optimisation and the reduction of empty trips.	Ang-Olson & Schroeer, 2002; Baas & Latto, 2005; CSIR, 2007; DME et al, 2002; Handy and Mokhtarian, 1996; IEA, 1996; Immers et al, 1994; Lovins et al, 2005; Martens and Korver, 1999; Taylor, 1999; Tichauer and Watters, 2008; USDEEE&RE and EPA, 2008; Vanderschuren et al, 1993; Vanderschuren and Jobanputra, 2005; Vanderschuren, 2006; Willekens et al, 2008; www.enviro.aero; www.flightsciences.com; www.freight-village.com; www.iata.org; www.vtpi.org; www.4Freight.net;
() = potential energy efficiency benefit margins according to international literature; P = Passenger cars; F = Freight vehicles; B = Buses		

Source: Vanderschuren et al. 2010

the side of caution and try to cover as many bases as possible. Some of the key challenges facing the sector that need to be considered when evaluating the potential success of a proposed strategy as described in [8] will now be discussed.

Currently the freight sector is operating at extremely high risk due to its large exposure to energy supply shocks (in terms of pricing and availability). It is imperative that measures implemented improve the sustainability of the current system and not foster the high risk status quo even further.

South Africa is not an exceptionally affluent country and spending should be done judiciously. It is vital to consider the externalities associated with a particular measure, as externalities are often not attributed properly. An example of this is the hidden externality costs of road freight, such as road infrastructure and environmental damage, accidents and congestion. These externalities are seldom reflected in the cost of freight transport. The intended and unintended impacts of the implementation of a measure (for example, increased vehicular loads or load consolidation) on the maintenance of the freight network needs to be taken into account. The historic lack of investment in new infrastructure has hampered development and as a result, many improvement measures will require significant infrastructure investment. This is very expensive and has a long lead time. Improvement measures aimed at alleviating pressure on bottlenecks and capacity constraints should enjoy high priority.

Measures dependent on a high-tech freight system or modern equipment and facilities will not be viable in the short to medium term, at least [8]. The prevailing skills shortage can also limit the level of sophistication of viable improvement measures, for example: intelligent transport systems might be too advanced to be readily adopted. Furthermore, measures that require large volumes of accurate data (in order to successfully plan implementation) are presently a risk in South Africa, due to the lack and inaccessibility of good quality data in the country [8].

The viability of measures is related to the ownership of the infrastructure utilised. Present day South Africa has inherited a legacy favouring road transport above all other modes. This trend is still going strong, as the State of Logistics reports produced by the [2, 3] suggest. Fair competition within and between modes must be ensured and where possible, increased. Measures that do not ensure an equitable distribution of infrastructure cost recovery (including the capital, management, operation and maintenance thereof) will likely bolster the artificial modal shift and distorted tariff structures created by historic cross-subsidisation [8].

On top of this, South Africa is spatially challenged in terms of its freight transportation, with production and mining facilities in the hinterland and ports vast distances away, at considerably lower altitudes. Another disparity is that the movement of goods from Gauteng constitutes double the movement towards Gauteng on the two main corridors [7]. Areas in the hinterland that have no access to rail are highly exposed, with a danger that both people and assets could become stranded. Interventions that rely on fully loaded return trips are not viable, as are strategies that rely on a strong rural transport network.

Lastly, South Africa is a country severely afflicted by HIV/AIDS. The illness is especially prevalent under truck drivers, resulting in extremely high driver turnover. This only exacerbates the skills shortage in the country [8]. Other social issues affecting freight transport include crime (hijacking) and trade union activity. Striking or protesting truck drivers can have devastating economic repercussions [13]. The social climate in South Africa must be taken into consideration during the analysis of measure viability.

It is evident that there are many elements to consider when developing freight energy management strategies. There are many different stakeholders, each with their own needs and priorities. A good strategy will, thus, provide solutions that do not only cater for subsets of stakeholders, but that rather tries to simultaneously satisfy the requirements of as many stakeholders as possible. To achieve this, a strategy formulation approach must incorporate multiple objectives and enable stakeholders to determine the trade-offs between objectives.

3 Translation of the problem into modelling constructs

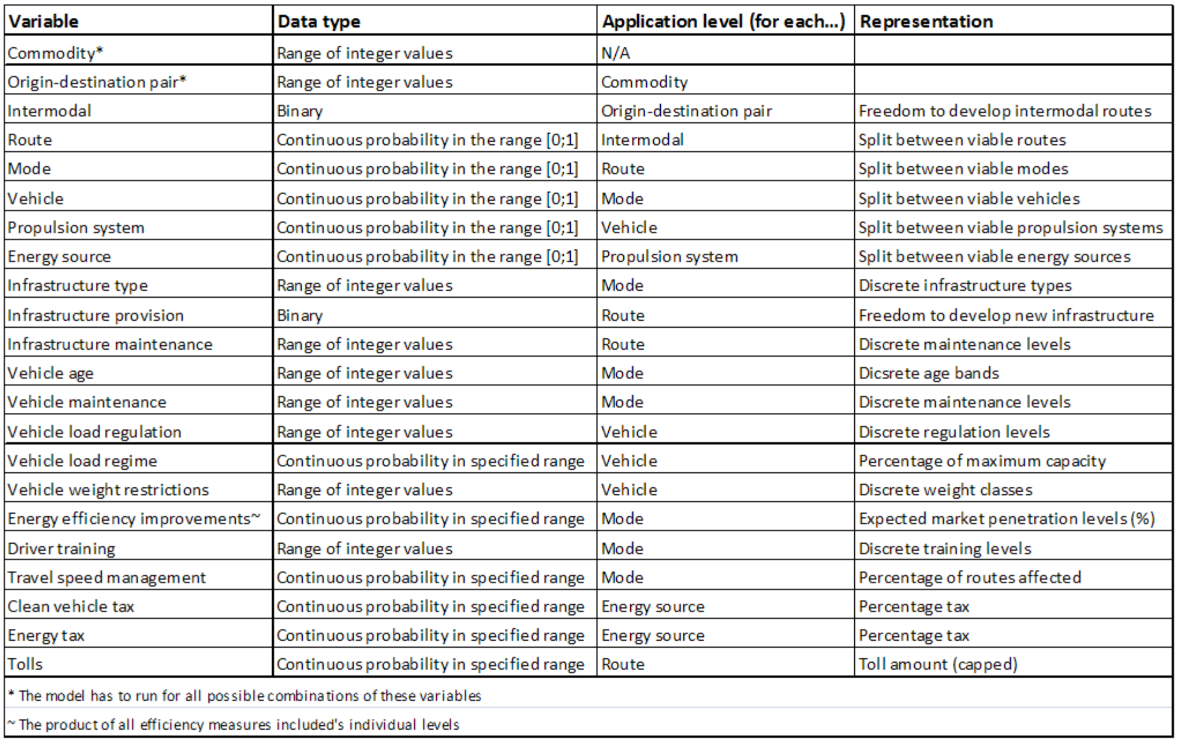
The first step in the formulation of a solution algorithm is to understand and define what a solution looks like. Based on this solution definition, an evaluation scheme (energy function) can be developed to assess the quality of the solution. A third step in the process is to define how the algorithm will progress from incumbent solution to incumbent solution through the search space when searching for the optimal solution.

3.1 Development of candidate solutions

A multi-level solution structure is proposed, based on the hierarchy of components in the freight transportation system. The reasoning behind the solution structure is as follows: for every commodity there are certain known transport demands between every origin-destination pair. For each of these pairs and the specific commodity, certain modes and routes are viable. A binary variable is included at this stage to indicate whether the model is free to propose intermodal solutions, or whether it must be confined to singular modes per route. A choice needs to be made between the different types of vehicles that can be included per mode (this can be different for every route, but is chosen from a discrete set of options). It should be noted that there is no discrete set of route options specified, a sub-model routing algorithm is used to stochastically vary the routing allocation. Different vehicle types can be associated with more than one type of propulsion system and a selection between these options needs to be made per selected vehicle. Finally, certain propulsion systems can operate on different energy sources. In short, a total transport demand allocation is made between all the energy sources, for all the propulsion systems associated with all the vehicle types selected for each modal routing option, for each origin-destination pair per commodity. These allocations all represent structural measures that can affect total energy demand.

There are, however, several other external, overarching elements that can affect total energy consumption, regardless of the specific demand allocation. Recent insights gleaned from on-going research pertaining to freight transportation energy management [10, 9, 8, 12, 7] has allowed the formation of a consolidated list of 22 freight transport energy management measures. These 22 measures are representative of the entire spectrum of energy management interventions proposed to date (internal as well as external). A solution is, thus, a set of unique vectors comprising of the 22 potential energy management measures included in this study. The values assigned to each vector component reflects the allocation decisions for internal measures, as well as the implementation levels of the external measures on the lowest level of aggregation (i.e. per energy source, per propulsion system, per vehicle, per route, per mode, per origin-destination pair, per commodity). The collective set of vectors represents all the commodities and their underlying allocations. A list of the variables used, as well as information on how they are utilised in the model is provided in Table 2.

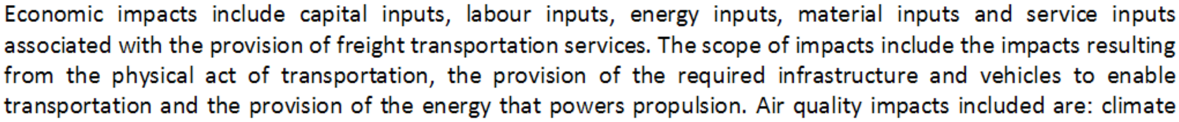
Table 2: *Summary table of the model variables constituting model solutions*



Variable	Data type	Application level (for each...)	Representation
Commodity*	Range of integer values	N/A	
Origin-destination pair*	Range of integer values	Commodity	
Intermodal	Binary	Origin-destination pair	Freedom to develop intermodal routes
Route	Continuous probability in the range [0;1]	Intermodal	Split between viable routes
Mode	Continuous probability in the range [0;1]	Route	Split between viable modes
Vehicle	Continuous probability in the range [0;1]	Mode	Split between viable vehicles
Propulsion system	Continuous probability in the range [0;1]	Vehicle	Split between viable propulsion systems
Energy source	Continuous probability in the range [0;1]	Propulsion system	Split between viable energy sources
Infrastructure type	Range of integer values	Mode	Discrete infrastructure types
Infrastructure provision	Binary	Route	Freedom to develop new infrastructure
Infrastructure maintenance	Range of integer values	Route	Discrete maintenance levels
Vehicle age	Range of integer values	Mode	Dicsrete age bands
Vehicle maintenance	Range of integer values	Mode	Discrete maintenance levels
Vehicle load regulation	Range of integer values	Vehicle	Discrete regulation levels
Vehicle load regime	Continuous probability in specified range	Vehicle	Percentage of maximum capacity
Vehicle weight restrictions	Range of integer values	Vehicle	Discrete weight classes
Energy efficiency improvements~	Continuous probability in specified range	Mode	Expected market penetration levels (%)
Driver training	Range of integer values	Mode	Discrete training levels
Travel speed management	Continuous probability in specified range	Mode	Percentage of routes affected
Clean vehicle tax	Continuous probability in specified range	Energy source	Percentage tax
Energy tax	Continuous probability in specified range	Energy source	Percentage tax
Tolls	Continuous probability in specified range	Route	Toll amount (capped)

* The model has to run for all possible combinations of these variables

~ The product of all efficiency measures included's individual levels



Economic impacts include capital inputs, labour inputs, energy inputs, material inputs and service inputs associated with the provision of freight transportation services. The scope of impacts include the impacts resulting from the physical act of transportation, the provision of the required infrastructure and vehicles to enable transportation and the provision of the energy that powers propulsion. Air quality impacts included are: climate

3.2 Evaluation of solution quality and development of the energy function

The four key objectives, representative of the stakeholder concerns in this project include freight transport's impacts on the environment, the economy, South African society and, finally, the impact of strategies on overall energy consumption. The scope and extent of these impacts and the interactions between them is too intricate to briefly explain in this article, but the key issues pertinent to each objective evaluation are highlighted.

Economic impacts include capital inputs, labour inputs, energy inputs, material inputs and service inputs associated with the provision of freight transportation services. The scope of impacts include the impacts resulting from the physical act of transportation, the provision of the required infrastructure and vehicles to enable transportation and the provision of the energy that powers propulsion. Air quality impacts included are: climate impacts, earth impacts and water impacts. The social impacts are impacts on mental well-being, impacts on material well-being and other socio-economic impacts. These impacts are discussed in depth in [10, 9]. Every one of the 22 measures included in the model affects these impacts in its own way and all of these relationships are modelled to form the energy function of the problem.

It is important to mention that there is only one constraint imposed on the model – all freight transport demand (specified in terms of tonnages to be moved between origin-destination pairs per commodity) must be met. The model cannot simply shut down the freight sector in order to save energy, but is free to apply any combination of measures in the transport system.

3.3 Solution perturbations

Due to the scale of the model and its multilevel structure, the solution search space is incredibly large. Table illustrates the solution range for each of the included variables. Perturbations of each variable have to be made bounded by the restrictions implied by the data structure of each variable. There are a vast number of combinations and permutations for perturbing solutions, ranging between perturbations to single variables and perturbations across the various levels in the solution structure. Adding further complexity to the model, measures can interact with one another and are not necessarily additive; hence a solution needs to include all the variables to be considered. Determining the best perturbation heuristic is part of the on-going research.

4 Archived Multiobjective Simulated Annealing (AMOSA)

Multiobjective optimisation (MOO) develops a set of solutions, called the Pareto-optimal (PO) set, which are considered to be equally important, as all of them constitute global optimum solutions. Over the past decade, a number of multiobjective evolutionary algorithms (MOEAs) have been suggested. The main reason for the popularity of evolutionary algorithms (EAs) for solving multiobjective optimisation is their population-based nature and ability to find multiple optima simultaneously [1]. Simulated annealing (SA) is another popular search algorithm. It utilises the principles of statistical mechanics regarding the behaviour of a large number of atoms at low temperature, for finding minimal cost solutions to large optimisation problems, by minimising the associated energy of the solution. Geman and Geman [6] suggested that SA, if annealed sufficiently slowly, converges to the global optimum.

In addition to the earlier aggregating approaches of multiobjective SA (MOSA), there have been a few techniques that incorporate the concept of Pareto-dominance [1]. These include multiobjective simulated annealing (MOSA), surrogate assisted simulated annealing (SASA), set-based multiobjective simulated annealing (SAMOSA), and volume-based multiobjective simulated annealing (VOLMOSA). Bandayopadhyay et al. [1] proposed a new MOSA - archived multiobjective simulated annealing (AMOSA), which incorporates a novel concept of amount of dominance in order to determine the acceptance of a new solution. The PO solutions are stored in an archive. In contrast to most other MOO algorithms, AMOSA selects dominated solutions with a probability that is dependent on the amount of domination measured in terms of the hypervolume between the two solutions in the objective space. Results demonstrate that the performance of AMOSA is comparable to, often better than that of

MOSA in terms of purity, convergence, and minimal spacing (as defined by Bandayopadhyay et al. [1].

The problem of finding optimal freight transport energy management strategies is complex, nonlinear, constrained, has multiple objectives and an extremely large search space. The multilevel structure of the transportation system leans itself towards programming based solution algorithms, such as metaheuristics. Based on a literature review, AMOSA is seen to be a good metaheuristic technique for the formation of a solution Pareto front. The development of a Pareto front of solutions for the problem at hand is highly desirable. It is proposed that cluster analysis is performed on the front to determine if there are underlying structural measure combinations that speak to the various stakeholder objectives, respectively. Understanding such structural differences will greatly advance decision making capabilities in this field. For further information on the AMOSA algorithm, it is highly recommended to read the article by Bandayopadhyay et al. [1].

5 Conclusions

Looking towards the future and the risks of energy price spikes and environmental issues becoming of paramount global concern, the key questions in freight transport energy management – how do we transform the freight transport sector into a greener sector, without impeding the benefits and services provided by the current system? Developing balanced and apt freight transport management strategies is the first step in this direction. The various stakeholders of the freight system, its far reaching impacts on sustainability and the scope and complexity of the sector renders this a difficult task. Sophisticated multiple objective optimisation algorithms are tools that can aid in the formulation of such strategies. In this paper a translation of the problem into modelling constructs is proposed, as well as the application of an archive multiple objective simulated annealing algorithm (AMOSA) as a suitable solution approach. Future work will include the development of a solution perturbation heuristic and the actual implementation of the proposed model in a South African case study.

Bibliography

[1] Bandayopadhyay, S., Saha, S., Maulik, U., and Deb, K. (2007). A simulated annealing-based multiobjective optimization algorithm: Amosa. In *IEEE Transactions on Evolutionary Computation.*

[2] CSIR Built Environment (2008). The fith annual state of logistics survey for South Africa. Available from http://www.csir.co.za/sol/.

[3] CSIR Built Environment (2009). The sixth annual state of logistics survey for South Africa. Available from http://www.csir.co.za/sol/.

[4] Department of Minirals and Energy (DME) (2006). RSA energy balance 2006. Online. Available from http://www.energy.gov.za/Energy_statistics/Energy_statistics.html (Cited 1 July, 2011).

[5] Department of Minirals and Energy (DME) (2007). Energy security master plan–liquid fuels. Government Report, Government Publications, Pretoria.

[6] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

[7] Hendler, P., Lane, T., Ratcliffe, S., Vanderschuren, M., and Wakeford, J. (2008). Energy and transport status quo: Demand and vulnerabilities. Section 2 of report submitted to the National Department of Transport.

[8] Lane, T. (2009). Assessing sustainability and energy efficiency improvement measures in freight transportation. Paper presented at the 28th Southern African Transport Conference.

[9] Lane, T. and Vanderschuren, M. (2010). Exploring the sustainability impacts of the contemporary South African freight transport sector. Paper presented at the 29th Southern African Transport Conference.

[10] Lane-Visser, T. and Vanderschuren, M. (2011). Evaluating the environmental impact of the South African freight systems energy tendencies. Paper presented at the 30th Southern African Transport Conference.

[11] Tran:SIT (2007). Energy, climate change and transport. Tran:SIT Update, Vol. 3.

[12] Vanderschuren, M., Lane, T., and Wakeford, J. (2010). Can the South African transport system surmount reduced crude oil availability? *Energy Policy*, 38:6093–6100.

[13] Williams, F. and De Waal, J. (2010). Transnet strike: exporters still paying. Fin24.com, June.

Proceedings of the 2011 ORSSA Annual Conference
pp. 81–92
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Local Government Elections – Some Personal Perspectives

HW Ittmann*

Abstract

The Constitution of South Africa requires that both national and local government elections be held every five years. These elections are not held simultaneously; currently, the local government elections are held two years after the national elections. In 2011, the 4th general local government elections took place. These were possibly the most hotly-contested elections since 1994, when South Africa became a democracy. As is the case in many national elections, a forecasting model has been developed to predict the ultimate outcome of the elections, based on early voting results. Such a forecasting model was developed by the Council for Scientific and Industrial Research (CSIR) for the 1999 national elections and has since been used to predict the results for all the national and local government elections to date. Each election poses its own challenges that need to be addressed in the forecasting model. Predictions are computed at national level (the race for votes), provincial level and also in the eight metros in the country. Local elections are different from national ones in that there is more than one ballot paper - individuals vote for a candidate in a ward as well as a party; these votes are used for candidates elected through proportional representation. In cases where municipalities fall within districts, there is a third vote for candidates to be elected for district councils. As predictions are computed, these need to be shared through the national public broadcast system, including radio and television stations and the broadcasters website. Subsequently, other media houses also request specific explanations or analysis. This, in itself, poses challenges. This paper will briefly outline the forecasting model used in the elections, what the expectations were, discuss the predictions as these unfolded over time as well as the challenges, the experiences and the interactions in dealing with the media.

1 Introduction

US President Abraham Lincoln (1809-1865) defined democracy as: “Government of the people, by the people, for the people” [9]. In most democracies - at least in representative democracies - regular elections are held to elect new governments. This is one of the most visible and true cornerstones of a democracy. Since the newly-established democracy in 1994,

*CSIR Built Environment, South Africa, email: hittmann@csir.co.za

South Africa has held regular elections every five years at both the national and local levels. To date, these elections have not coincided and are held two years apart. The last national elections took place in 2009, while the most recent local elections were held in May 2011.

A team of operations researchers from the CSIR has been involved on contract to the SABC, the national public service and commercial broadcasting corporation, in predicting the election results through a forecasting model developed in the '90s. This model has been used since 1999 to predict results for all the national and local elections held in the country. The predictions need to be communicated to the media in different forms namely digitally, by radio and by television. Operations researchers are trained primarily to understand and address difficult decision-making problems and then to construct and develop models to solve these. Communication skills in interacting and explaining the results to clients are very important. In this specific instance, one of the most critical aspects that contributes significantly to the success of the prediction model is the ability to interact with the different types of media to convey the predicted results on a continuous basis as these become available. The author has been privileged to be the main spokesperson of the operations research (OR) team of the CSIR since the 1999 national elections and shares his personal perspectives and insights gained from these experiences.

The paper briefly describes different types of elections and election-forecasting methodologies. A short outline of the forecasting model developed for the South African elections is presented. The experiences during the run-up to the elections, on election day itself and the predicted results compared to the final results are discussed for the recent local elections, with reference to some of the previous election results as well. Finally, a fairly extensive section is devoted to the interaction with the different types of media and the experiences gained and the lessons learnt are presented. It is hoped that operations researchers can gain new insight from this and realise the importance of appropriate communication to ensure the message reaches the target audience, in this instance, the public at large.

2 Election forecasting

Elections create enormous interest worldwide. The media report on election campaigns months beforehand, while there is continuous speculations about the possible outcomes. In this regard, the final outcomes of any elections are not only of importance to the politicians involved, but have a huge influence on how countries will be managed and ruled, on the future economy of a country, etc. The policies of the party elected into power will play the strongest role in government policy and decision-making until the next elections.

Various forms of election forecasts exist and are being used in most elections [3, 6, 7, 8, 10, 12, 14]. Before an election, market surveys - or opinion polls - are conducted to establish how eligible citizens will vote. This is one way of getting a sense of what the ultimate outcome of the elections may be. Representative sampling methods are used for these surveys and, although efforts are made to ensure accuracy, the element of the lie factor must always be taken into account. People are not always totally honest about who they would vote for. Another type of forecast is conducted through exit polls. Here voters are asked - again through sampling methods - for which candidate or party they had voted as they leave the voting station. This method has its own drawbacks and challenges but it is used in countries such as the UK. A third way of forecasting elections is to use the early results as these are received to forecast the final outcomes. Various methods and models have been developed to enable operations researchers to predict the results in this manner. The last approach is the one used by the CSIR team.

The forecasting methods are all very dependent on the political representative model and electoral system being used in a country. The key components of political representation, on almost any account, will exhibit the following four components [13]:

- some party that is represented (the representative, an organisation, movement, state agency, etc.);
- some party that is being represented (the constituents, the clients, etc.);
- something that is being represented (opinions, perspectives, interests, discourses, etc.); and
- a setting within which the activity of representation is taking place (the political context).

In 1994, South Africa changed its electoral system to that of proportional representation whereas previously it was based on constituencies [1] In a system of proportional representation, citizens vote for a party; from the number of votes received by the party, the number of seats is allocated to the party. Upfront, the party provides a list of candidates and seats are allocated from the top of this list - in this way, the individuals who will represent the party are determined. This holds true for national elections, while for local elections one finds individual representation as well as proportional representation, and the number of voting papers required differs for municipal and districts councils [15]. In both cases there are wards where the citizens vote for a candidate (of a specific party, typically) and also cast votes for a party, which are used for proportional representation. Some areas of the country have district councils - in those areas, citizens cast a third vote, which is used to determine the proportional representation in district councils - again, the individual needs to vote for a political party.

3 Forecasting models

As indicated earlier, various approaches are followed to predict the outcome of an election in a country and this is dependent on the election system. In the UK, a constituency system is used and because of the homogeneous voter make-up of a constituency, exit polls are used as well as previous election results to determine the share of vote in making predictions [3]. In the USA, the election of the next president is possibly of greatest interest during the elections held there every four years. Lichtman and DeCell [8] have correctly predicted the popular vote outcome of every US presidential election since 1984. These predictions are based on 13 questions, each with a “yes” or “no” answer [11]. The “yes” answers favour the incumbent party candidate. If five or fewer answers are “no”, the incumbent party retains the presidency, while if six or more are “no”, the challenger wins. This method is based on a statistical pattern recognition algorithm that incorporates “a test of competing theories of politics”, which closely resembles kernel discriminant function analysis. The prediction results validate some of the theories and contradict others. Although the next US presidential election is still 18 months away, Lichtman predicts President Obama to win [12]. Election-night forecasting approaches for other countries have been developed, inter alia for New Zealand [10], India [7] and Sweden [14].

The model developed for the South African elections is described in detail in [6] as well as [5]. The method used is firstly to formulate a cluster model, which aims to divide the population/electorate into groups, or clusters, with similar voting behaviours. The clusters are determined before the elections and are then used during the elections to extrapolate the initial partial results to the whole cluster and thereby for the whole electorate. Initially, the clusters were determined by using demographic data, but after the 1999 elections these were based on those election results. In following this approach, the one major assumption that drives this approach is that voters in the various voting districts belonging to the same cluster would vote according to the same patterns in future elections.

The formal method used in determining the clusters is called the fuzzy clustering approach, or C-means method [2]. As explained by Greben et al. [6], the fuzzy clustering process was carried out for the 1999 election results, using 20 clusters. As a result, there was a compromise between a large number of clusters, about 40, to allow for sufficient discrimination between different cluster centroids and a smaller number, say five, allowing predictions on a minimum of results. Having obtained these clusters, they are then used as the results come in from specific voting districts, of which the model knows the cluster membership, to calculate the predictions by extrapolating these for the entire electorate. The results are obtained from the Independent Electoral Commission (IEC) via the SABCs information system that is linked to the IEC system. The forecasting system also endeavours to predict voter turnout, which is another interesting statistic associated with elections.

4 Run-up to the elections

Operations researchers know that it is only possible to develop a model, or represent reality in mathematical terms, if the researcher has an in-depth knowledge and understanding of the real problem situation. Whether the problem to be solved is in the forestry area, health or airline industry, it is important to understand that specific domain. In this case, the domain is the political environment. Although a forecasting model has been developed to predict the election results, the spokesperson needs to have more than a superficial knowledge of the political scene and political developments. In addition, the political environment is different for each election and this also needs to be understood. When interacting with the media, e.g. during interviews, it is impossible to have pre-prepared written notes or answers that can be used during interactions.

Duckworth and Lewis [4], arguably crickets most modern and well-known partnership, were the formulators of the mathematically-based method that is currently used to ensure fairness to the problem of how to adjust targets if rain interferes in one-day, or limited overs, cricket matches. They very soon realised, in explaining their method, *"interviews dont work like that. Whereas you will certainly have the gist of responses to likely questions embedded in the mind its a mistake to try to use pat responses. For a start a response read from notes sounds very unnatural and the way the questions are varyingly phrased requires a different emphasis and hence different wording from anything that has been prepared in advance. In other words it is important to be spontaneous and also not overly complicated"*. Not only is it essential that one knows the forecasting method used very well but, more importantly, one needs to be able to explain this in easily understandable, digestible terms and language to the public!

What complicates matters even more is that elections take place every few years and one needs to keep updated with each election regarding the methodology as well as the current political situation in the country. For the latter, one needs to stay on top of developments in the political arena by tracking these in the media, for example by reading a variety of newspapers, listening to the radio and watching television. It is critical to know what the "hot spot" areas are with the highest interest in terms of the possible outcomes. These are situations where the outcomes could be very close such as, in the case of South Africa, prior to the 2011 elections, some local communities had severe protests due to poor service delivery, which could have affected the outcomes of the elections in those municipalities. All of this is included in preparation and research that need to be done consciously before the elections. Obviously, it is not appropriate to act or be seen to act as a political analyst. However, one needs to be prepared to answer political-type questions and one should have an opinion! The media do not understand, and dont care, what the difference is between an operations researcher, a results analyst or modeller, and a political analyst.

Background material on elections is also available on the website of the IEC. This includes information on the number of registered voters, the number of voting stations, number of wards and number of candidates. All of this is very valuable for any spokesperson. Before the recent local elections, the SABC news journalists had a briefing session, which proved to be very informative and interesting to the author as the forecasting teams spokesperson. Journalists in all nine provinces gave their views on what could happen in the different provinces.

The CSIR team needs to do quite a bit of preparatory work before election day. All the models/programs need to be loaded on the computer and these needs to be tested - at least whether the data sent from the SABC system are read correctly. Communication between the various systems also needs to be tested. Any last minute changes or additions, if and as required, need to be made. In all cases the number of parties needs to be entered into the model, as well as the demographic data, etc. Because of the number of parties involved, the model does not predict the results for all of these but only for the 10 to 12 largest. Examples of new challenges encountered for the prediction model were, for example, in 2009, COPE was a new entrant into the elections, while in 2011, independent candidates - people not on the ANC lists but actually representing the ANC caused some real concern whether the model would be able to handle them correctly. Significant tasks usually need to be performed in this regard.

Lastly, one cannot underestimate the experience gained during previous elections, also in media interaction, and this proved again to be invaluable in all aspects.

5 Election day and the results event

Election day is a normal day for an operations researcher or results analyst, for whom it is important to vote and get a feel for the vibe of the elections. The work of the researchers and results analysts starts only after the elections when the first results start coming in. It is with trepidation and excitement that one goes to the IEC results centre at the Pretoria show grounds. To experience this is possibly one of the most exciting involvements the author has had in any project over many years. The experience may be short-lived it lasts for a day or two, but it is extremely intense. During that time the adrenaline is pumping, one interacts and mingles with the media and also with well-known politicians and personalities while the election outcomes unfold. Initially there is also the unspoken but constant concern, in all the analysts minds, about whether the model will perform as one would expect or not! Although the model did work in the past, one is still concerned as it has not been tested fully with the real data before a specific election.

The SABC news team was going to report on the results of the elections around four specific topics, namely:

1. The race for votes this is an indication of the percentage voter support for each political party at national level, provincial level as well as for metros;
2. The race for wards the number of wards that each party would get in the various municipalities;
3. The race for seats the proportional representation of each party in the municipality and/or district council. These are determined from the support a party receives as well as the number of wards. The IEC has a number of formulas which are applied to determine this; and
4. Seat allocation the allocation of seats for the municipalities and district councils when all the results are in.

The forecasting model contributed mainly to predicting the race for votes.

6 The predictions and final results a comparison

The forecasting model described here was used, as indicated, during the previous five South African elections, namely the 1999 National Elections, the 2000 Municipal Elections, the 2004 National Elections, the 2006 Municipal Elections and the 2009 National Elections. The model used in these five elections proved to be very robust and achieved a high degree of accuracy. During the 2000 municipal elections, the ANCs final result was predicted to within 1% after only 10% of the votes were counted - this was at 02:45 in the early morning hours after voting on the previous day. At that time, the actual results still showed a 20% deviation from the final results. For the DA, a 1% accuracy was achieved after 20% of the votes were counted and released by the IEC.

In the 2004 elections, the prediction for the ANCs final percentage result (69%) was within 1% when only 5% of the votes had come in. In the municipal elections of 2006, the model predicted the overall ANC result (65.8%) with a relative error of 1% (66.5%) when 10% of the votes had come in, at 02:23 in the morning following the elections. At that point-intime, the actual results (62.6%) still deviated about 5% from the final results. For the smaller parties, the relative errors were larger, however, the relative error for the DA (5%) was dramatically smaller than that of the actual results at that time (38% over-estimate). Similar levels of accuracy were achieved during the 2006 and 2009 elections. The national elections of 2009 were very interesting, since the main discussion point at the results centre, and in the entire country for that matter, was whether the ANC would obtain a two-thirds majority. Early the morning after elections, around 07:00, the model predicted that the ANC would get 65.6%, whereas the party ultimately obtained 65.9% of the votes casted! However there was a point when the actual numbers on the results scoreboard at the IEC results centre showed that the ANC had secured just over 70% of the votes. Not all votes were counted yet at that stage, and since the prediction model knew which voting station results were still outstanding and took this into consideration, it predicted correctly - that the ANC would not get the two-thirds majority! The CSIR team was adamant that the model was correct and the spokesperson had to stick to his guns and ensure the nation that the team was confident of the outcome this prediction was, of course, not popular, especially with the ruling party and its supporters!

For the 2011 elections, there was huge interest in the contest between the ANC and the DA - the two main parties at national level, at provincial level (mainly in the Western Cape), and in the case of a number of the metros. With these elections it became very obvious, however, that the country is moving towards a two-party state. The smaller parties did not really feature and most of them received very limited support. The model predicted in the race for votes at the national level, with 1.8% of voting districts counted, that the ANC would get 62.8% and the DA 27%, respectively. This was the prediction at 23:30, late on the night of the elections. Ultimately, the ANC got 62.93% and the DA 24.08%. Early the morning after the elections, at 07:52, with 30% of the voting districts counted, the prediction given on radio was 62.8% and 24.27%, respectively, for the ANC and the DA.

A number of predicted results are shown in Table 1 for only the ANC and the DA. These are for the main areas of interest and the percentage voting district results that were out, formally released by the IEC, is also shown.

As can be seen from Table 1, the predictions were fairly accurate even when very few of the results had been available. It is noticeable that, especially in the metros, the early-predicted results were not that close to the final results. The main reason for this is the fact that the clustering is done at national level but for metros the voting trends in clusters are used at a much lower level where the representation is not that well distributed. This is very noticeable in the City of Cape Town metro where the predictions only started to converge to the final results after 50% of the voting district results were received. Nevertheless, from the results shown in Table 1 it is very clear that the predictions were excellent with very few results known at that stage.

7 Communication challenges

The author was the sole spokesperson of the CSIR team during the initial years, with a colleague joining him in that role since the national elections in 2009. Over the past two decades, the author has had experience of media interviews during the course of his other work, including OR, not relating to election forecasting. With election forecasting being a very specific topic of interest to all in South Africa during elections, the author had a two-hour “training session” with a television presenter before the elections in 1999. Some general communication skills were discussed and shared with him, while he picked up subsequent skills through experience during the five elections since 1999. A good foundation and understanding of quantitative methods are essential. This section, however, describes personal perspectives of what is required in conveying the results of a mathematical model to the general public,

Table 1: *The race for voter predictions*

Party	% Voting district results out	Prediction in %	Final result in %
National			
ANC	1.8	62.80	62.93
DA	1.8	27.00	24.08
ANC	30	62.80	62.93
DA	30	24.27	24.08
Provincial	**Western Cape**		
ANC	16.5	37.65	34.10
DA	16.5	53.15	58.10
Metros			
(i) City of Tshwane			
ANC	0	56.39	56.46
DA	0	37.06	38.74
(ii) City of Cape Town			
ANC	2.3	36.64	33.17
DA	2.3	53.94	61.15
(iii) City of Johannesburg			
ANC	0	56.77	59.29
DA	0	33.96	34.35
(iv) Nelson Mandela Bay			
ANC	9.8	54.49	52.13
DA	9.8	39.12	40.24

generally regarded as lay people in this sphere. Some anecdotes are used for illustrative purposes. This is by no means a comprehensive view on requirements to communicate with the media.

Good communication skills are essential when interacting with the media. Lewis from Duckworth and Lewis [4] states that after his education, “useful seeds had therefore been sown in me; an interest in applying quantitative ideas practically and a training in communication skills”. Other important factors include a good command of the language(s); the ability to think and speak in a logical manner and to convey a difficult concept in an uncomplicated, concise (to-the-point) and understandable way. It is very important to have self-confidence and also confidence in what one is commenting on, with the ability to “think on ones feet” being crucial. One must also not be afraid to present an informed opinion with conviction. One should not get flustered during a radio or television interview or debate. A simple way to handle this is to focus on the questions asked and respond as best as one can or to comment and participate during a discussion - this forces one to never even think about all the people out there listening to or watching you. One should try to not even think that one is on radio or television, it is totally immaterial one could even pretend to be speaking to a family member to be more at ease. What is important, though, is to be sensible in what one says and to convey the results, or facts, as honestly and correctly as possible. For some people it comes more naturally than for others, but practice and experience will help anyone improve. Fortunately in the case of a quantitative model and its results, one is talking about facts and real numbers, which do make things easier.

In presenting the forecast results, this should be done and discussed with absolute conviction and confidence. It should be crystal clear that the team absolutely believes in the model and the results presented by the spokesperson. If that does not come across strongly, how should the audience (listeners and viewers) believe, accept or take these predictions seriously? One should have faith in the models results, and never question these. This was the case in 2009 with the general perception of the ANC going to obtain a two-thirds majority, whereas the model consistently predicted that this was not going to happen.

Early-on during any elections, the inevitable question in media interaction is: “now explain the model and how it works”. Clearly one cannot say it is a fuzzy clustering method or a C-means method that is used in modelling the voting behaviour! The author typically starts by saying a model is a representation of reality and, in this case, the voting population needs to be represented in a mathematical way. Based on certain grouping criteria, the voting population is combined into a number of groups (maybe use the term “clusters”), and the main assumption is then that those who populate a specific group vote in the same way.

Duckworth and Lewis [4] decided very early on in developing and testing their method that they would never admit publically if something was wrong with the method. Their argument was that people would immediately lose faith in the method. The same could be said of the prediction model. Two examples are cited: In 2009, due to data communication problems, the voter turnout percentage was very clearly wrong. That specific prediction was just never mentioned or used. In 2011, while being interviewed on radio, the author was passed an unexpected note stating “what do you make of this final result in the Northern Cape province”. The final results of the main parties in that province were printed on this piece of paper as well. The models prediction numbers were totally different from the final results reported in that province. Sweating and blushing, with a sudden increased heart beat, the author just talked around this without the radio presenter even noticing anything! Therefore, if something goes wrong, it should not be obvious to anyone both on radio and television.

It later turned out that there was a mistake in the printed output from the prediction model: the headings for the predicted results of the Northern Cape and North Western provinces were swapped around! This leads to another very critical point - testing of the system is crucial. That specific SABC report format was added for the 2011 elections, and since the team became involved only two weeks before the elections, this report was not properly scrutinised for correctness beforehand.

A tough but good lesson learnt through experience is that one should always answer or respond to any question. It is even possible to not answer a specific question directly, while providing a response! If a presenter wanted to get an answer to a specific question and it was not answered, (s)he will push for an answer again - by then one has already conveyed ones argument to the audience. Once during a television show, the presenter introduced the two guests as “political analysts” and then proceeded to ask the author a question. The author started his response with: “I am not a political analyst”, and before he could continue, the presenter promptly passed the question to the actual political analyst and ignored the author for the rest of the show.

Although one analyses modelled results, it is inevitable that political questions will be asked and one is expected to provide a comment. If one can use predicted results or facts in answering the question, it is ideal. However, this is not always possible and then it is critical that, like in every problem-solving situation that OR people face, there is proper knowledge and understanding of the domain. One should have a view and an opinion. This one formulates and accumulates by reading and listening to the experts! Luckily, in these situations one can never be wrong; however, quantitative people deal with facts! It is thus also important to stay abreast of current terminology or phrases that are being used, as these can even change from one election to another. Obviously, one still needs to be careful what one says and what one is prepared to comment on - one should never make a fool of oneself. In one of the previous national elections, where the New National Party (the old National Party) lost horribly and got only 1.7% of the vote, the author did say on an early-morning Afrikaans radio show that this “was clearly the end of the road for the National Party” as a non-political analyst, it was a very dangerous thing to say.

Appearing on television has its own frustrations too. The presenters are directed by the producers and as a guest on the show one is not privy to the direction to be taken. The implication is that one is in the dark most of the time about how the show will proceed. It happens very often that one had not finished an explanation, but that there was never a chance to return to that point. One also has to be prepared to wait long periods on a show where one basically sits and not say a word. One gets onto a show and the presenter says: “Welcome, you will be with us for the next two hours” - nobody tells one this in advance, and on top of that, one speaks or discusses issues for maybe only 5 or 10 minutes at most, during the two hours!

One should also expect the unexpected and be prepared to respond to questions in a confident, logical way. How does one respond to a question like: “What is the actual purpose of this prediction model and why should one spend money on doing forecasting?” This is the “so what" question that the public is interested in "what is in it for me?" The author did not have enough time to respond, but nevertheless said: "To get an indication of what the ultimate result would be as soon as possible". Not bad, but there could have been a whole range of responses.

8 Conclusion

This paper endeavours to give a very unusual perspective of an aspect that is required in OR during the development of any mathematical model, namely communicating with the public at large as the client. However, this case is unique with the kind of exposure obtained during an election. In this instance, the models or methodology used is very critical since it requires very accurate prediction results, but these then need to be shared with a wide audience. The prediction model, briefly described here, has been very successfully used since 1999 to forecast election results in South Africa. In addition, the election team was able to provide inputs in communicating the predictions to the wider public. From an OR point of view, this is a unique occasion with its own unique experiences, but nevertheless with the same steps required in developing any model. The exception is that the communication of the results rendered by the model through the different media is so crucial.

Acknowledgements

All the OR modelling team members, namely Jan Greben, Chris Elphinstone, Peter Schmitz, Jenny Holloway, Rosalie de Villiers, Zaid Kimme and Maanda Rasuba are thanked for their dedication and commitment in developing the model, the running and use of the model and the support provided throughout all the elections. The forecasting model not only performed well, but produced excellent results! The SABC is thanked for its support in this initiative and also for giving the CSIR the opportunity to be involved in this exciting event.

Bibliography

[1] Alvarez-Rivera, M. (2004). Election resources on the internet: The republic of South Africa electoral system. http://electionresources.org/za/system/.

[2] Bezdek, J., Trivedi, M., Ehrlich, R., and Full, W. (1981). Fuzzy clustering: a new approach for geostatistical analysis. *International Journal of Systems, Measurement and Decision*, 1–2:13–24.

[3] Brown, L. and Chappell, H. (1999). Forecasting presidential elections using history and polls. *International Journal of Forecasting*, 15:127–135.

[4] Duckworth, F. and Lewis, T. (2011). *Duckworth Lewis The method and the men behind it.* Sportsbooks Limited, Cheltenham, UK.

[5] Greben, J., Elphinstone, C., and Holloway, J. (2006). A model for election night forecasting applied to the 2004 South African elections. *Orion*, 22:89–103.

[6] Greben, J., Elphinstone, C., Holloway, J., De Villiers, R., Ittmann, H., and Schmitz, P. (2005). Prediction of the 2004 national elections in South Africa. *South African Journal of Science*, 101:157–161.

[7] Karandikar, R., Payne, C., and Yadav, Y. (2002). Predicting the 1998 india parliamentary election. *Electoral Studies*, 21:69–89.

[8] Lichtman, A. and DeCell, K. (1990). *The Thirteen Keys to the Presidency.* Madison Books, Lanham, Md.

[9] Lincoln, A. (1865). *The Papers of Abraham Lincoln.* Manuscript Division, Library of Congress, Washington, DC.

[10] Morton, R. (1988). Election night forecasting in New Zealand. *Electoral Studies*, 7:269–277.

[11] Samuelson, D. (2008). Road to the white house. OR/MS Today. pp 26–28.

[12] Samuelson, D. (2011). Elections 2012: The 13 keys to the white house. OR/MS Today. p 26–28.

[13] Stanford Encyclopedia of Philosophy (SEP) (2006). First published Mon Jan 2, 2006; substantive revision Fri Jun 24, 2011; http://plato.stanford.edu/entries/political-representation/.

[14] Thedeen, T. (1990). Election prognosis and estimates of voter streams in Sweden. *New Zealand Statistician*, 25:54–58.

[15] website, I. (2011) Cahange

Proceedings of the 2011 ORSSA Annual Conference
pp. 93–102
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

A Nurse Rostering Algorithm for a District Hospital in South Africa

MJ Treurnicht[*] TE Lane-Visser[†] L van Dyk[‡] SS Friedrich[§]

Abstract

An acute shortage of healthcare professionals is the rule rather than the exception in South Africa. Effective scheduling of nurses is critical to ensure good quality of care, while limiting staff related healthcare costs and abiding by labour laws. South African district hospital nurses are presently scheduled through the manual production of duty rosters on a monthly basis. In this paper, two related nurse rostering problems (NRP) are formulated for a district level public hospital in Stellenbosch (South Africa). The first problem addresses the scheduling of the months that nurses are on night shift duty. The other problem addresses the scheduling of the days that nurses are working night or day shifts within a month, respectively.

A hierarchy of four levels exists among the nursing staff at Stellenbosch Hospital. Distributed over the four levels, the nursing staff totals ninety employees. Due to hospital policy, no casual nurses are employed. Fluctuations in demand are met by scheduling overtime shifts that are limited by current labour legislation. The hospital consists of seven wards, each with separate staff requirements. The NRP for Stellenbosch Hospital is solved using the genetic algorithm. The algorithm is adapted to specifically adhere to the requirements and constraints given by the formulated NRP. The algorithm outputs optimal feasible rosters for each problem and provides data required to evaluate the performance of the algorithm. Roster results are interpreted and verified using the initial nursing requirements of the hospital. The robustness of using such an algorithm for sustainable use is also discussed. The paper ultimately aims to promote the use of operations research in healthcare on a practical level in South Africa.

Key words: Genetic algorithms, hospitals, metaheuristics, scheduling

1 Introduction

One of the major challenges in most developing countries is the provision of quality healthcare for all. Similar to many other developing countries, South Africas public sector has an acute

[*] Corresponding author: Stellenbosch University, South Africa, email: miekie@sun.ac.za

[†] Stellenbosch University, South Africa, email: tanyav@sun.ac.za

[‡] Stellenbosch University, South Africa, email: lvd@sun.ac.za

[§] Stellenbosch University, South Africa, email: 15297839@sun.ac.za

shortage of medical expertise. Proper scheduling of nursing staff is critical to effectively utilise the available scarce resources. Nurse scheduling could have a large impact on consistency of care provided, eliminating excess or waste resources and reducing health costs as well as maintaining high staff morale [2, 5].

The nurse rostering problem (NRP) and the nurse scheduling problem (NSP) have attracted much research attention due to the time consuming and often complex nature of nurse scheduling. A large variety of NRP models and solution methods have been researched. These models typically involve the development of a periodic duty roster, subject to some constraints that are mostly hospital-specific. The objectives of the rosters could also vary between maximising the shift preferences of nurses to minimizing costs [2].

The NRP could be simplified by limiting the number constraints and complexity of the objective function. Many research articles limit the complexity of the problem to emphasize a solution method [5]. Hence, in academic papers complex optimisation methods and heuristic models are researched using simplified NRP. According to Kellogg and Walczak [5] the simplification of a problem results in a situation that the study is not likely to be implemented. For successful implementation of the NRP in hospitals, a complex set of constraints are mostly required. The result is that many real-world applications of the NRP are over-constrained, and complex to solve (Ernst et al. 2004).

Meta-heuristics such as the Tabu Search method, Simulated Annealing and Genetic Algorithms have proven to be effective in finding near-optimum solutions to NRPs [2, 4]. These solution methods use random orchestrated search strategies to explore a solution space, looking for a global optimum while avoiding local optima [3]. Despite the variety of commercial rostering software available and the many publications on the topic, the majority of hospitals still rely on manual scheduling. Kellogg and Walczak [5]) attribute this to a mismatch between practical applications and the type of research that is published by academia.

The aim of the paper is to investigate whether a genetic algorithm can be used to solve a real-world NRP at a public hospital in South Africa. In this paper we investigate the use of a genetic algorithm to produce nurse duty rosters for Stellenbosch Hospital, a district hospital in the Western Cape, South Africa. Two similar NRPs are formulated to represent the current rostering specifications of the hospital. Genetic algorithms are specifically programmed to solve these NRPs. The solutions are discussed with recommendations for future work.

2 Formulating the NRP's for Stellenbosch Hospital

The majority of public hospitals in South Africa produce their nurse duty rosters manually. Nursing managers spend a substantial amount of time developing these rosters. Nurses are generally allowed to make requests that complicate the process even further. Nurse scheduling has always been a rather complex task. The primary reason for this is that hospitals are operational, 24 hours a day, 7 days a week. Furthermore, nursing staff needs to be scheduled in such a way that services with acceptable quality of care are always available, while simultaneously limiting the number of nurses employed and healthcare costs [2].

The nurse rostering problems are formulated specifically to model the existing practice at Stellenbosch Hospital. To deliver a 24 hour service at the hospital, shifts are defined as being day or night shifts, each shift lasting 12 hours. It is compulsory that all hospital nurses work night shifts. To promote staff morale and to abide by labour laws, nurses work night shifts in blocks that should include at least three months. These night shift blocks are scheduled annually, allowing nurses to adapt to a night shift lifestyle.

Nursing managers produce two types of rosters. The first roster is produced annually, scheduling night or day shift blocks as well as ward allocation. This roster outputs in which months each nurse has to work night shift as well as ward allocations for each month. Monthly rosters are separated as day or night shift rosters, scheduling the days of the month in which each nurse is on duty. The monthly rosters use the output from the annual roster as input. Constraints are shared between the monthly and annual rosters, therefore if constraints should change, all the rosters should be modified.

2.1 Annual Duty Rosters

The purpose of this roster is to determine the months in which nurses have to work night shifts by considering nurse preferences as well as allocating nurses to wards. The problem is subject to some constraints that considers ward staff levels and nurse specialities. Nurses are asked to complete a preference matrix, indicating which months they would prefer to work night shifts. Another matrix contains the ward experience of each nurse. The objective function maximises the match between ward preference of nurses and the years of ward experience for all the nurses. For the purposes of this paper, the weighting of the nurse preferences and experience terms in the objective function are equal. Nevertheless, the importance of ward experience with respect to nurse experience would be different for other hospitals and could be changed according to the specifications of nursing managers. The problem is formulated as follows:

Defining variables:

$i = 1, 2, 3, \ldots, 90$	Nurse index
$j = 1, 2, 3, \ldots, 12$	Month index
$k = 1, 2, 3, \ldots, 7$	Ward index
$l = 1, 2, 3, 4$	Speciality index

Parameters:

$B_{i,j}$ $\triangleq$ Preference of nurse i to work during month j

$$S_{i,l} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ is qualified to work as speciality } l \\ 0, & \text{Else,} \end{cases}$$

$$P_{T_i,l} \triangleq \begin{cases} 1, & \text{A nurse following pattern } T_i \text{ works night shift during month } j \\ 0, & \text{A nurse following pattern } T_i \text{ works day shift during month } j, \end{cases}$$

$W_{i,k}$ $\triangleq$ Experience (in years) of nurse i in ward k

$C_{k,l}$ $\triangleq$ Contraints for total number of nurses working night shift with specialty l in ward k.

Decision variables:

$$T_i \triangleq \begin{cases} 1, & \text{Nurse } i \text{ is scheduled to work night shift pattern 1} \\ 2, & \text{Nurse } i \text{ is scheduled to work night shift pattern 2} \\ 3, & \text{Nurse } i \text{ is scheduled to work night shift pattern 3} \\ \dots \\ 48, & \text{Nurse } i \text{ is scheduled to work night shift pattern 48} \end{cases}$$

$$V_{i,j,k,l} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ with speciality } l\text{, works night shift during month } j \text{ in ward } k \\ 0, & \text{Else,} \end{cases}$$

$$X_{i,j,k,l} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ with speciality } l\text{, works day shift during month } j \text{ in ward } k \\ 0, & \text{Else,} \end{cases}$$

Objective Function:
Maximise the total preference and experience of all nurses over the period of a year:

$$\max z = \sum_{j}^{90} \sum_{j}^{12} B_{i,j} P_{T_i,j} + \sum_{i}^{90} \sum_{j}^{12} \sum_{k}^{7} \sum_{l}^{4} W_{i,k} V_{i,j,k,l} \tag{1}$$

Subject to:
Each nurse should be assigned to one ward per month:

$$\sum_{k=1}^{7} \sum_{l=1}^{4} \left(V_{i,j,k,l} + X_{i,j,k,l} \right) = 1 \quad i = 1,2,3,\dots,90;\ j = 1,2,3,\dots,12 \tag{2}$$

Each nurse is only assigned to their qualified speciality per month:

$$\sum_{k=1}^{7} \left(V_{i,j,k,l} + X_{i,j,k,l} \right) = S_{i,l} \quad i = 1,2,3,\dots,90;\ j = 1,2,3,\dots,12;\ l = 1,2,3,4 \tag{3}$$

Number of night shift nurses according to speciality in wards per month

$$\sum_{i=1}^{90} = V_{i,j,k,l} \leq C_{k,l} \qquad j = 1,2,3,\ldots,12;\ k = 1,2,3,\ldots,7;\ l = 1,2,3,4 \tag{4}$$

All nurses are scheduled according to shift patterns (expressed in terms of night shifts months)

$$\sum_{k=1}^{7}\sum_{l=1}^{4} V_{i,j,k,l} = P_{T_i,j} \qquad i = 1,2,3\ldots,90;\ j = 1,2,3,\ldots,12. \tag{5}$$

2.2 Monthly Duty Rosters

The monthly rostering problem determines the shift patterns for the months in which nurses are on duty. The monthly rostering problem is similar to the night shift rostering problem. Since wards are assigned for each month on an annual basis, the assignment of wards is not included in the monthly rosters. The monthly rostering problem inputs the results from the annual roster for specific constraints, such as the speciality of nurses scheduled for day or night shift and their assigned wards. The objectives of the monthly rostering solutions are to maximise the sum of the preference for the entire nursing staff. The problem is subject to a set of constraints that are specific to each of the seven wards and four specialities. For example, in the maternity ward, exactly two sisters have to be on duty for each day of the month.

Defining variables:

$i = 1,2,3,\ldots,90$	Nurse index
$d = 1,2,3,\ldots,12$	Day index
$k = 1,2,3,\ldots,7$	Ward index
$l = 1,2,3,4$	Speciality index

Parameters:

$D_{i,d} \triangleq$ Preference of nurse i to work during day d

$$S_{i,l} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ is qualified to work as speciality } l \\ 0, & \text{Else,} \end{cases}$$

$$N_{i,k} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ is scheduled to work in ward } k \text{ for this month} \\ 0, & \text{Else,} \end{cases}$$

$$Q_{R_i,k} \triangleq \begin{cases} 1, & \text{A nurse following pattern } R_i \text{ works night shift during day } d \\ 0, & \text{A nurse following pattern } R_i \text{ works day shift during day } d, \end{cases}$$

$K_{k,l} \triangleq$ Contraints for total number of nurses with specialty l in ward k.

Decision variables:

$$R_i \triangleq \begin{cases} 1, & \text{Nurse } i \text{ is scheduled to work pattern 1} \\ 0, & \text{Nurse } i \text{ is scheduled to work pattern 0} \end{cases}$$

$$V_{i,d,k,l} \triangleq \begin{cases} 1, & \text{Nurse } i \text{ with speciality } l\text{, works during day } l \text{ in ward } k \\ 0, & \text{Else,} \end{cases}$$

Objective Function:
Maximise the total preference of all nurses over the period of a month

$$\max z = \sum_{i}^{90} \sum_{d}^{31} D_{i,d} Q_{R_i,d} \tag{6}$$

Subject to:
Each nurse is assigned according to the wards scheduled for each day in the monthly roster

$$\sum_{l=1}^{4} V_{i,d,k,l} = N_{i,k} Q_{R_i,d} \qquad i = 1,2,3,\ldots,90;\ d = 1,2,3,\ldots,31;\ k = 1,2,3,\ldots,7 \tag{7}$$

Each nurse is only assigned to their qualified speciality for each day

$$\sum_{k=1}^{7} V_{i,d,k,l} = S_{i,l} Q_{R_i,d} \qquad i = 1,2,3,\ldots,90;\ d = 1,2,3,\ldots,31; l = 1,2,3,4 \tag{8}$$

Number of night shift nurses according to speciality in wards per day

$$\sum_{i=1}^{90} V_{i,d,k,l} \leq K_{k,l} \qquad i = 1,2,3,\ldots,90;\ k = 1,2,3,\ldots,7; l = 1,2,3,4 \tag{9}$$

All nurses are scheduled according to specific shift patterns

$$\sum_{k=1}^{6} \sum_{l=1}^{4} V_{i,d,k,l} = Q_{R_i,d} \qquad i = 1,2,3,\ldots,90;\ j = 1,2,3,\ldots,31 \tag{10}$$

2.3 Limitations to the NRPs

The formulated NRPs do not address fluctuating demand, leave, overtime shifts or hiring temporary nurses. Unit managers assess the demand for nurses continuously and assign overtime shifts accordingly. Nurses are requested to apply for leave a month in advance. It is against Stellenbosch Hospitals policy to hire additional temporary nurses. Overtime is scheduled during the month to compensate for demand fluctuations and nurses leave requirements. These aspects could be added to the problem, however, at this stage of developing a rostering solution, it is desired that the unit managers still have manual control over detailed shift changes and overtime shifts.

3 Solving the NRP using Genetic Algorithms

Two genetic algorithms were programmed for the respective annual and monthly NRPs. Microsoft Excel was chosen as a computing platform together with Visual Basic as programming language. The reasoning behind using Microsoft Excel as user interface is that administrative staff and unit managers at the hospitals are already trained to use Microsoft Office. Unit managers are currently using Microsoft Excel to manually produce duty rosters. They would therefore easily adapt to using an algorithm to produce a duty roster that can be modified manually using the well known spreadsheet functionality of MS Excel.

Optimization problems require both variety and progression. Natural phenomena, therefore, provide valuable principles for algorithms. Genetic algorithms mimic the biological theory of evolution where plants and animal species breed to form new offspring with unique features. With the birth of a new offspring a new generation is formed. The concept of evolution is that new generations possess different characteristics than the previous generation with the capability of improved performance in the new environment. The survival of the fittest principle is applied and thereby as time progresses the new generations become stronger through abandoning weaker individuals. Mutations that occur randomly reduce the possibility of inbreeding and therefore the chances that the population is trapped at a local optimum [4].

Unfortunately, solving highly constrained problems with genetic algorithms is rather complex. A reason for this is that the generation of new offspring through crossover does not allow for constraint consistency. Hence, the birth of a new offspring through crossover does not necessarily allow for a new feasible solution. The new offspring has to meet with the requirements of the constraints before it can be considered a possible solution. If infeasible solutions dominate the solution space, it is unlikely that the genetic algorithm will succeed in finding a good feasible solution [1].

Infeasible offspring can be avoided by using penalty functions or repairing infeasible solutions. Penalty functions steer the search away from infeasible solutions, thereby avoiding the problem. Unfortunately, there are no guarantees that a solution can be found through avoiding this complication with penalty functions. Repairing infeasible solutions are also not the perfect solution. Through repairing, the characteristics of the offspring are often changed in such a way that a weaker individual is formed, resulting in a time consuming process of generating many weak offspring that are abandoned and does not contribute in creating a stronger population [1].

3.1 Initiation

The initial population consists of a number of feasible trial solutions. Each individual solution in the population is created randomly and repaired to meet the set of constraints. An individual solution consists of genes that determine the strength of the individual. In the NRP a gene could be the shift type chosen for a corresponding nurse. A basic feasible solution would therefore consist of a number of genes, indicating the chosen shift types for a number of nurses. The fitness (value of the objective function) is calculated for each feasible member of the initial population.

3.2 Iterations

The iteration phase of the algorithm randomly selects parents to compete in a tournament. Two parents with relative high fitness are chosen and paired with each other in a multiple crossover process. Crossover points are chosen at random, and therefore if two parents were to pair more than once, different offspring will be created. Offspring inherit a combination of both parents genes, a different combination of genes from the same parents would therefore result in a different child. Mutation occurs on a random basis. The operator of the algorithm can specify a mutation probability. This probability determines the mutation rate. If the mutation possibility is met, the new offspring is generated with new genes that do not necessarily belong to any specific parent.

New offspring is subject to the set of constraints. If a new offspring is created that does not fit criteria and are therefore infeasible, the solution (offspring) is repaired in the same way that the individuals of the original population were repaired. The fitness of each offspring is calculated. If the offsprings fitness is better than the weakest member of the population, any random old member of the population is replaced with the offspring. This process of creating new offspring is continued until the termination criterion is met.

3.3 Termination rule

The iteration phase is a loop that continues until a termination condition is encountered. There are a number of different termination rules that can be used such as a fixed CPU time, a fixed number of iterations or a fixed number of consecutive iterations without improvement. If the algorithm is executed for an adequate time period, the entire population will have the same fitness. If no new offspring are created during a number of iterations, it can be assumed that the algorithm has found a near optimum solution. Therefore, after reaching the termination criterion the solution space would have converged to the final solution.

4 Results

At first the genetic algorithms were programmed to randomly search for a feasible solution, without repairing or using penalty functions. As literature suggested, the algorithm was unable to find any feasible solutions in an acceptable time period. The reason for this is most likely the diversity in shift patterns that could be assigned. If no structure is provided, the algorithm will randomly search for a combination of shift patterns among the nurses that satisfies the constraints. Unfortunately, the possibility of finding a combination of 48 patterns among 90 nurses that fits the constraints is very small.

By repairing individuals to create a population of feasible solutions, infeasible solutions were removed from the solution space. It should be noted that it is possible that the repairing method could influence the impact of the crossover method. This is due to the fact that an offspring is repaired after crossover. It is therefore likely that the offspring after repairing might not be similar to either parent. The good characteristics might not survive the repair process. It is therefore likely that through repairing the offspring, the good characteristics might get lost and the algorithm might stop at a local optimum.

The genetic algorithms were successful in finding feasible rosters for both annual and monthly rosters. Using genetic algorithms to solve two related NRPs we were able to accomplish the following:

- An annual roster that allocates each nurse to night shift blocks and wards for each month of the year
- A monthly roster that specifies the days that each nurse are on duty

The genetic algorithms output for the annual is in the form of shift pattern numbers for all the nurses, which are translated to a roster showing the months in which the nurse are on night or day shift duty. The monthly rosters output is similar to the annual roster, with the exception that wards are not allocated on a monthly basis. The monthly roster will indicate the days of the month that the nurses are on duty. If the annual roster indicates that the nurses are on night shift duty, the days that the nurses are on duty according to the monthly roster will apply to night shifts. The same principle applies to day shifts. Thus the two rosters are dependent and should not be used independently.

5 Conclusions and Recommendation

The final feasible rosters are good solutions, but are most likely not the optimum solutions. Nevertheless, the aim of this study was to provide a good alternative to producing manual rosters at Stellenbosch Hospital. The algorithms were able to find a variety of different rosters that would fit the constraints of the hospital and choose the best rosters for nursing staff preference and ward experience.

The next step is to test and validate the method for Stellenbosch Hospital. If necessary, constraints such as leave, overtime and demand fluctuations could be added. Further alterations will be done at the request of the hospital staff. If the hospital staff would like to replace manual scheduling with the algorithm, implementation should include adequate training and support.

Results from the algorithms should be validated during implementation to ensure that all the nurses are treated fairly and hospital demands are met. If possible results in terms of nurse preferences should not be made public to the general nursing staff. It is inevitable that some nurses will have higher preference ratings than others. It should therefore be avoided that nurses compare their preferences with each other. This will reduce the jealousy effect of introducing the preference method.

The genetic algorithm was programmed to be problem specific. This enabled the algorithm to search within a structure as opposed to absolute random searches. Therefore if this algorithm is considered for another hospital similar to Stellenbosch Hospitals, alterations are necessary.

Acknowledgements

The authors wish to thank the staff of Stellenbosch Hospital for their enthusiasm and support in this project.

Bibliography

[1] Aickelin, U. and Dowsland, K. (2004). An indirect genetic algorithm for a nurse scheduling problem. *Computers & Operations Research*, 31:761–778.

[2] Cheang, B., Li, H., Lim, A., and Rodrigue, B. (2003). Nurse rostering problems - a bibliographic survey. *European Journal of Operational Research*, 151:447–460.

[3] Glover, F. and Kochenberger, G., editors (2003). *Handbook of Metaheuristics.* Kluwer Academic Publishers, Boston.

[4] Hillier, F. and Lieberman, G. (2005). *Introduction to Operations Research.* McGraw-Hill, Singapore.

[5] Kellogg, D. and Walczak, S. (2007). Nurse scheduling: From academia to implementation or not? *Interfaces*, 37:355–369.

Proceedings of the 2011 ORSSA Annual Conference
pp. 103–110
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Operations Research in Telemedicine

AJ van Zyl*   L van Dyk†

Abstract

Telemedicine is the delivery of healthcare, where distance is an issue. This paper explores the use of operations research techniques to contribute to the successful implementation of telemedicine systems. A study of literature concerning the application of operations research together with telemedicine, showed research gaps, but also identified a few examples with respect to linear programming, simulation modelling, Markov decision-making processes, Bayesian networks, queuing theory and meta-heuristics. There are many areas in telemedicine where operations research can be used to ensure the success and sustainability of telemedicine as a viable alternative to providing specialist healthcare that is economically feasible.

Key words: telemedicine, operations research, health systems, healthcare

1 Introduction

Telemedicine is the delivery of healthcare, where distance is an issue. Broens (2007) found in a study of 50 international telemedicine projects that most of these projects failed. This observation is confirmed by the low success rate in the implementation of telemedicine projects in the public health sector of South Africa. Health systems engineering is an academic discipline where researchers and practitioners treat the healthcare industry as complex systems, and further identify and apply engineering applications in healthcare systems. Health Systems Engineers make us of - amongst others - operations research techniques. Operations research allows us to solve complex decision-making problems, a common phenomenon when developing and implementing a telemedicine system. This paper explores the use of operations research techniques to contribute to the successful implementation of telemedicine systems.

Methodology: The first part of this paper is devoted to the study of literature pertaining to the use of operations research techniques within the context of health systems engineering, with a specific focus on telemedicine. This is followed by a qualitative and quantitative evaluation of a selection of techniques. A few examples of operations research techniques used within the context of healthcare is also given to support findings from the literature study and technique evaluation.

*Corresponding author: Stellenbosch University, South Africa, email: 14834340@sun.ac.za

†different Stellenbosch University, South Africa, email: lvd@sun.ac.za

2 Literature Review

Two main components play a vital role when developing and implementing a telemedicine system, namely the telemedicine system itself and the project management team who is responsible for the successful implementation of the system. Both of these components can be supported by operations research individually and also simultaneously. Operations research provides the project management team with decision support and the telemedcine system with operating support.

A multiple keyword search is used to identify the availability of literature containing operations research in telemedicine. eHealth, mHealth, and telehealth, are just some of the words that are associated with telemedicine.SciVerse Scopus was used as the search engine to gather articles containing the key words and phrases.

SciVerse Scopus is a bibiliographic database containing abstract and citations for scholarly journal articles. The results returned from Scopus is of a scholarly nature excluding advertisements and irrelevant articles. Scopus searches both the internet and its local (scholarly) database of article. A summary of the results obtained from the Scopus search is discussed in the following section.

2.1 Scopus Search Result - Scholarly Database

Table 1: *Summary of Search Results - Scopus Scholarly Database*

Keyword	Keyword	Found	Date
Operations Research	-	30 371	1947 - 2011
Operations Research	health	1 374	1952 - 2011
Operations Research	health care	883	1964 - 2011
Operations Research	decision support	882	1974 - 2011
Operations Research	project management	778	1968 - 2011
Operations Research	health systems	56	1966 - 2011
Operations Research	telemedicine	6	1994 - 2007
Operations Research	electronic health	1	2007

Table 1 shows the keywords entered into the search space and the number of articles that contain those keywords. This table is a summary of the local Scopus database. This database only contains scholarly articles and does not include any results obtained from the internet. Operations research was used as the fixed variable in all of the searches. From the results it is clear that there is a multitude of articles containing operations research literature. However, as soon as the search is tweaked to be more specific, focusing on the medical field, the number of scholarly level articles decreases dramatically.

2.2 Scopus Search Result - Internet

Table 2 shows the summary of the results obtained with the exact same keywords as mentioned in the previous section. The difference is that the results include data obtained from the internet. The only filter applied in this instance is that the results returned must be articles. This is done to avoid corrupting the data with advertisements or irrelevant website data. Also, note all the results returned from the internet contains dates, hence it is omitted from the table.

Table 2: *Summary of Search Results - Scopus Internet*

Keyword	Keyword	Found
Operations Research	-	766 469
Operations Research	health	256 870
Operations Research	health care	55 381
Operations Research	decision support	74 279
Operations Research	project management	63 200
Operations Research	health systems	9 054
Operations Research	telemedicine	2 199
Operations Research	electronic health	896

It is interesting to note that a much larger portion of the articles pertain to the health environment. This implies that operations research is indeed being used extensively in the health sector, however the documentation and implementation of the operations research techniques used in the healthcare sector are not being documented academically and is mostly personal research. Another point of interest is the fact that the same ratio applies in the case of telemedicine related searches. Few articles of any nature exist which clearly makes use of operations research in telemedicine.

2.3 Search Conclusion

The Scopus search engine indicated that few works about the use of operations research in the telemedicine field have been published. This can be contributed to the fact that telemedicine is a relatively new branch in the medical field and also that the use of operations research techniques to support decision making in the telehealth industry is a modern trend that is only now starting to gain popularity in the profession.

From the search results it can also be seen that early attempts have been made to incorporate operations research into healthcare, but the focus was more on clinical support rather than health support.

This is an indication that more research is required to identify appropriate uses of operations research which will benefit the medical field and in this manner improve the success of health and telemedicine projects. The next section describes some of the operations research techniques and possible application of these techniques to support healtcare, especially telemedicine.

A summary of the search results obtained from Scopus is shown in figure 1:

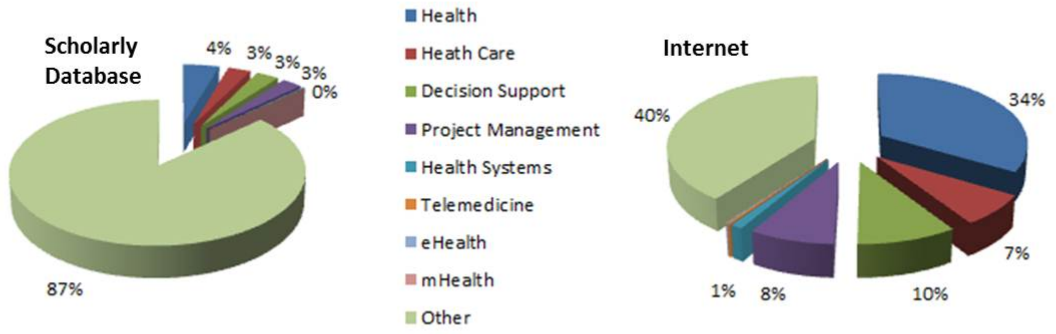


Figure 1: *Summary of Scopus Search Results*

3 Operations Research Applications in Telemedicine

This section describes some operations research techniques used in the telemedicine field. A brief discussion and a relevant example is also given for each operations research technique.

3.1 Linear Programming

Linear programming (LP) is a tool for solving optimization problems and is used in industries as diverse as banking, education, healthcare, etc. [2]. Linear programming has proved useful in modelling diverse types of problems in planning, scheduling, assignment, and design [7].

An example of the use of linear programming to aid the cause of telemedicine in South Africa is the development of a decision support framework for telemedicine by Treurnicht [8]. A telemedicine workstation was developed as a first step in addressing the need of specialized healthcare in rural communities in South Africa. This was followed up by the development of a decision support framework that will assist telemedicine decision makers with a scientifically based needs assessment [8].

Integer programming was implemented to find the best alternative given the utilization and cost of the telemedicine devices, by maximizing the benefit or utilization while reducing the costs involved. The outcome of the project was that devices could be recommended for the workstation given patient data and cost of devices.

3.2 Simulation Modelling

The complexity of many organizations in today's society is such that the totality of the effects of the interacting elements includes sometimes cannot be captured adequately through mathematical model. This is particularly evident when many of the factors affecting the performance of the organization cannot be predicted with certainty. In cases like these, the analyst will often turn to simulation modelling as an alternative approach to the problem [3].

An example of simulation used in telemedicine is the project by Lach [5] that provides people living in poverty conditions in Mexico with medical care. A mobile unit is sent to location, fully equipped with telecommunications gear and satellite connection, where a patient can be diagnosed by a specialist located hundreds of miles away in a hospital. The program's processes were simulated on a computer model. This allowed for various scenarios to be tested and to identify any possible problems. The results were used to assist decision making procedures that lead to increased performance and efficiency of the program [5].

In this instance, simulation modelling was used to analyse a new process. Simulation modelling can also be useful when an already existing process with a lot of variables and constraints need to improved. The popularity of simulation as an optimization tool has lead to many software developers developing powerful simulation packages such as Arena and more recently Simio (a powerful simulation modelling tool with an extensive three dimensional library for visualization).

3.3 Markov Decision Processes

Diagnosing a patient more often than not requires sequential decision making in an uncertain environment. Most of these decisions are made relying on the skill of the practitioner and his ability to make the correct assumptions. Due to the fact that the decision made can be the difference between life and death, a support tool to aid the practitioner in his decision making process is crucial. The use of mathematical decision models can reduce the burden on the practitioner as well as increase the likelihood of a successful diagnoses. Markov decision processes (MDPs) are one such technique for aiding the practitioner with certain types of treatment decisions [4].

A serious problem in the medical field is that patients with dementia often cannot remember how to do simple, everyday task. Boger [6] proposed the use of a computerized guidance system that can support a person with dementia, reducing the person's reliance on a caregiver. A planning system that uses MDPs to assist a person with dementia perform the simple task of washing hands was developed. Results from the study gives a clear indication that MDPs can be used to great effect in this type of guidance problem [6].

3.4 Bayesian Networks

A Bayesian network is a graphical representation of a probabilistic model of conditional dependencies using variables of interest [4]. A Bayesian network could for example represent the probabilistic relationship between diseases and symptoms.

A model of a real-world application which aim is giving a daily diagnosis on the hydration state of kidney disease to people is used as an example. Based on a collaboration with the ALTIR and the C.H.U. of the Nancy Center for the Treatment of Kidney Disease Patients and the University Hospital of Nancy, the Diatelic project aims to monitor kidney disease patients at home who are waiting for a kidney transplant. These patients must have a substitution process in order to replace the kidney before the transplant. To improve the performance of the system and enhance the interaction of the telemedicine system with the specialist and patients, a Dynamic Bayesian Network is used. The goal of the model is to predict a problematic situation before it happens rather than discovering them as they occur [9].

The approach to model such a domain in terms of data fusion using the dynamic Bayesian network was successful and it is possible to detect pathological situations by directly using the results issued from the dynamic Bayesian network, which allow a synthetic view of the possibility that the patient is in an abnormal hydration state [9].

3.5 Queueing Theory

Queueing theory is a mathematical model approach to describe queues. This mathematical technique allows for the derivation and calculation of several performance measures. Queueing theory is applied in various fields, including telecommunications, traffic engineering, and hospitals. It is considered an operations research technique since the results are often used when making decisions about the resource needed to provide the service.

An example of the application of a queueing system in telemedicine is the study done by Tarakci [10] of the optimal strategy to provide traditional face-to-face consultation via experts and remote medical services via tele-specialists to a remote hospital. The whole system is modelled as a queuing problem and the optimal staffing policy for this hospital is provided by taking into account the various cost components, such as those for staffing, mistreatment, and waiting. An optimal investment in telemedicine technology that offers the best trade-off between the quality and accuracy of telemedicine services is also found. The outcome is that the queueing system determines the optimal tele-specialist policy of which patients to treat remotely via telemedicine and which patients to refer to the experts for a face-to-face consultation [10].

3.6 Meta-Heuristics

Meta-heuristics are ways to go about searching for the optimal solution in a structured and logical fashion. It is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regards to certain user defined constraints. A candidate solution and a method to test whether the solution is good or bad, better or worse is required. Normally, a meta-heuristic is applied in situations where there is very little heuristic information available and a brute force search is out of the question because the search space is too large. There are many different meta-heuristics such as particle swarm, ant colony, and genetic algorithm. Each technique has its strong- and weak points and the use of the correct technique is subject to research and experience.

Telemammography and telemedicine requires that the characteristics of an image, such as the image resolution, bit-depth and intensity response are standardized to ensure the integrity of the diagnosis. Qian [11] developed a method based on the genetic algorithm to allow for the standardization digital mammography images. The techniques that are used to standardize the image are based on geometric and intensity transformations that are discovered by using calibration images. The result of the study was that the genetic algorithm method improved the detection sensitivity rate (true positive%) from 60% to 87% while the false positive was successfully reduced from 3.5 per image to 1.9 per image [11].

4 Conclusion

Few articles on a scholarly level exist that illustrates the use of operations research techniques in telemedicine. This should be seen as an opportunity for operations research specialists to contribute to the field of telemedicine and the use of operations research in Telemedicine. The previous section showed that there are people using operations research techniques to facilitate the development and implementation of telemedicine systems. It also indicated that there are many areas in telemedicine where operations research can be used to ensure the success and sustainability of telemedicine as a viable alternative to providing specialist healthcare that is economically feasible.

The growing need for specialist healthcare in the developing countries due to economic growth has pressured governments and health institutions to increase its capacity and reach when providing healthcare. Telemedicine provides an economically feasible solution to one man's most basic rights; a right to proper healthcare. It is thus worthwhile to invest in telemedicine, not only financially, but also academically.

Bibliography

[1] Bashshur RL & Shannon GW, 2009, *History of Telemedicine*, Mary Ann Liebert Inc. Publishers, New York, USA.

[2] Winston WL, 2004, *Operation Research: Application and Algorithms*, Brooks/Cole - Thomson Learning, Belmont, USA.

[3] Miller DM & Schmidt JW, 1984, *Industrial Engineering and Operations Research*, John Wiley and Sons, USA.

[4] Brandeau ML Sainfort F & Pierskalla WP, 2004, *Operations Research and Health Care*, Kluwer Academic Publishers, Norwell, Massachusetts, USA.

[5] Lach JM & Vazquez RM, 2004, *Simulation Model of the Telemedicine Program*, Proceedings of the 36th Conference on Winter Simulation, Winter Simulation Conference, Washington, D.C.,pp 2012–2017.

[6] Boger J & Poupart J, 2006, *A Planning System Based on Markov Decision Processes to Guide People with Dementia through Activities of Daily Living*, IEEE Transactions on Information Technology in Biomedicine, 2nd Edition, pp 323–333.

[7] Wikipedia, 27 June 2011,*Linear Programming*, [Online], [Cited: 06 July 2011],Available: `http://en.wikipedia.org/wiki/Linear_programming`.

[8] Treurnicht M, 2009,*A Decision Support Framework for Telemedicine Implementation in the Developing World*, [Online], [Cited: 06 July 2011],Available: `http://www.iienet2.org/uploadedFiles/SHSNew/Tools_and_Resources/2010_Conference_Proceedings/TreurnichtPaperSHS2010.pdf`.

[9] Bellot D Boyer A & Charpillet, *Designing Smart Agent Based Telemedicine with Dynamic Bayesian Networks: an Application to Kidney Disease People*, [Online], [Cited:

06 July 2011],Available: http://www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.8848&rep.

[10] TARAKCI SHARAFALI & OZDEMIR, 2007, *Optimal Staffing Policy and Telemedicine*, Melbourne Business School, [Online], [Cited: 06 July 2011],Available: http://works.bepress.com/hakan_tarakci/5.

[11] QIAN W SANKAR R & SONG X, 2003, *Standardization for Image Characteristics in Telemammmography Using Genetic and Non-linear Algorithms*, Computers in Biology and Medicine, pp 183–196.

Proceedings of the 2011 ORSSA Annual Conference
pp. 111–123
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Paving the way for the use of prediction modelling in a hospital environment

I van Zyl* TE Lane-Visser† L van Dyk‡

Abstract

The high cost of hospitalisation is a challenge for many health insurance companies, governments and individuals alike. In 2006, studies concluded that well over $30 billion was spent on unnecessary hospitalisation in the United States of America. In all likelihood this could have been prevented through early patient diagnosis and treatment so, there is room for improvement in this regard. Prevention is always better than cure; especially where lives are at stake and successful decisions regarding hospitalisation prediction may make unnecessary hospitalisation a realistic possibility. The aim of this paper is to pave the way for the development of successful hospitalisation prediction models.

Key words: Prediction modelling, hospitalization

1 Introduction

The Heritage Provider Network, a health insurance provider and sponsor of the Heritage Health Prize (HHP) Competition, has recently come to realise the potential benefits of a hospitalisation prediction model [7]. The competition is aimed at producing an effective hospitalisation prediction algorithm, using health insurance claims data. The purpose of this competition is ultimately, to prevent the unnecessary hospitalisation of members in their network. If successful, this could lead to fewer critical medical cases, fewer claims and consequently lower expenses for all the stakeholders in the affected system. The competition serves as inspiration for this study.

This study is not the first to consider the use of operational research techniques to assist in hospital predictions. For example Miyata et al. [12] used multivariate logistic regression to predict in-house mortality in hospitalisation, using records obtained from a nation-wide

*Corresponding author: Stellenbosch University, South Africa, email: 15324745@sun.ac.za
†different Stellenbosch University, South Africa, email: tanyav@sun.ac.za
‡different Stellenbosch University, South Africa, email: lvd@sun.ac.za

administrative database in Japan. Ganster et al. [6] also used logistic regression to predict consequent health care costs associated with stressful work conditions and personal control. These studies used medical records as well as occupational classification data. Decision tree analysis was done by Lee et al. [11] to predict an outbreak of dengue haemorrhagic fever (DHF). These predictions would be able to help doctors decide whether to hospitalise or do outpatient monitoring.

The shortcomings of these studies, when compared to the HHP case study, are that the response outputs of these models are binary in character. Miyata et al. [12] predicted for mortality or non-mortality, Miyata et al. [12] used logistic regression, which mostly has binary output and lastly, Lee et al. [11] required a prediction output of either hospitalisation or outpatient monitoring, which is also binary in character.

The aim of this study is to pave the way for predictive patient admission algorithm (PPAA) developers by providing insights and identifying possible pitfalls in the development of such an algorithm.

The paper consists of two parts. In the first part, typically available hospitalisation data, which serves as input for the PPAA, are briefly described, together with methods to extract, transform and load (ETL) data within this context. Next, a list of contender techniques and technologies is assembled, based on the given data, the algorithms expected input requirements and the techniques ability to meet these needs.

The prediction modelling techniques reviewed include regression methods, neural networks and ensemble methods. The expected outputs of promising techniques are also discussed briefly. The paper ultimately provides a recommendation on the preferred technique and technology to use in the development of hospitalisation prediction models of this kind. Potential pitfalls, which may be encountered, are highlighted and discussed throughout.

2 Data and data handling

The problem at hand relies heavily on data, which makes it imperative to first understand the available input data. It is also important to process the data in such a way as to maintain its accuracy and integrity. Section 2 thus attempts to gain a good understanding of the data and to find ways to handle it, while preserving its integrity.

2.1 The data

The Heritage Health Prize (HHP) data was received from the Heritage Health Provider Network. It is realistic data although some distortion occurred when members identities were hidden by the competition organisers. The data consisted of the following elements:

- General information about members who are part of the Heritage Health Provider Network health insurance company.
- Information about the claims made by members every year.
- Information about the amount of days that members spent in hospital every year.
- Information about drug prescriptions claimed by members.

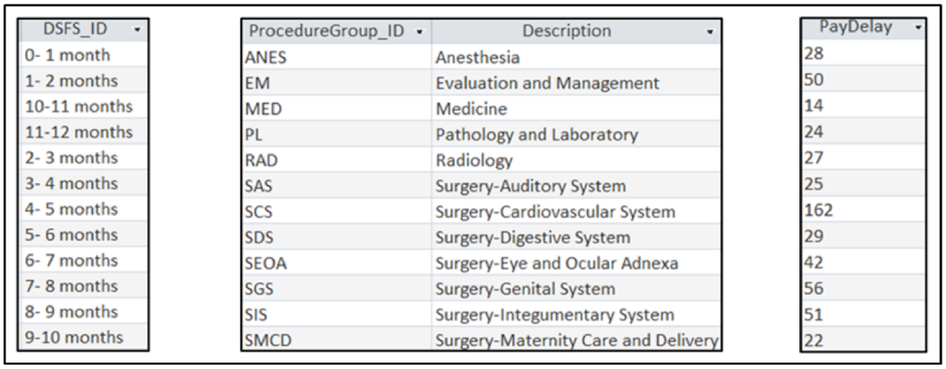
- Information about lab tests claimed by members.

- Metadata to describe codes used in certain data fields.

A data dictionary is presented in Table 1, as different types of variables can be expected in health insurance claims and hospitalisation data. Firstly, the most general type is continuous numeric variables. Many data modelling techniques can only use categorical variables, and in these cases, continuous numeric variables can be converted into discrete numbers (also called discretizing). For example, if the numeric continuous variables range is 1 to 100, these variables can be discretized by dividing them into bins (sub-ranges) of four: 0-25, 26-50, 51-75, 76-100 [13]. Another kind of variable is categorical variables, which can be either nominal or ordinal. Examples of these different kinds of variables can be seen in Figure 1. Column DSFS_ID is an example of an ordinal categorical variable, column Procedure Group_ID is an example of a nominal categorical variable and column PayDelay is an example of a continuous numerical variable. The PPAA should be able to accommodate all these variables.

Table 1: *Heritage Health Prize data dictionary*

Variable	**Description**
MemberID, ProviderID, Vendor	Member, provider and vendor pseudonym.
AgeAtFirstClaim	Age in years at the time the first claims date of service was computed.
Sex	Biological sex of member: M = Male; F=Female.
PCP	Primary care physician pseudonym.
Year	Year in which the claim was made: Y1; Y2; Y3.
Speciality	Generalized specialty.
PlaceSvc	Generalized place of service
PayDelay	Number of days delay between the date of service and date of payment
LengthOf Stay	Length of stay (discharge date admission date + 1)
DSFS	Days since first claim, computed from the first claim for that member per year
Primary Condition Group	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes.
Charlson Index	A measure of the effect diseases have on overall illness, grouped by significance, that generalizes additional diagnoses.
Procedure Group	Broad categories of procedures.
SupLOS	Indicates if the NULL value for the LengthOfStay variable is due to suppression done during the de-identification process.
DrugCount	Count of unique prescription drugs filled by DSFS. No count is provided if prescriptions were filled before DSFS zero.
LabCount	Count of unique laboratory and pathology tests by DSFS.
DaysInHospital_Y2, DaysInHospital_Y3	Days in hospital Y2, Y3
ClaimedTruncated	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise. If truncation is indicated (in years 2 and 3) it means that a certain member had more that 43 claims for a specified year. Truncation is used as part of the suppression done during the de-identification process.



DSFS_ID
0- 1 month
1- 2 months
10-11 months
11-12 months
2- 3 months
3- 4 months
4- 5 months
5- 6 months
6- 7 months
7- 8 months
8- 9 months
9-10 months

ProcedureGroup_ID	Description
ANES	Anesthesia
EM	Evaluation and Management
MED	Medicine
PL	Pathology and Laboratory
RAD	Radiology
SAS	Surgery-Auditory System
SCS	Surgery-Cardiovascular System
SDS	Surgery-Digestive System
SEOA	Surgery-Eye and Ocular Adnexa
SGS	Surgery-Genital System
SIS	Surgery-Integumentary System
SMCD	Surgery-Maternity Care and Delivery

PayDelay
28
50
14
24
27
25
162
29
42
56
51
22

Figure 1: *Examples of different types of variables found in health insurance claims and hospitalisation data.*

2.2 Issues with data interpretation

It is important to note that data sets are often riddled with ambiguities and uncertainties. Examples found in the HHP data set are listed below, and can be expected in similar data recorded in the hospital environment:

- Each member specifies a primary care physician. This could be one doctor or a group of doctors.
- A similar situation is found in MemberID, as a MemberID can represent either one person or a family. That is why, in some cases, it has been found that a male member might have a condition of pregnancy (the person who is pregnant is simply a dependent of the main member who happened to be male) [8].
- Where the length of hospital stay (LengthOfStay) column is blank, it is assumed that patients stayed less than a whole day [9].
- The amount of drugs consumed in a specified year (DrugCount) is described in the data dictionary as the "Count of unique prescription drugs filled by DSFS." This is more easily understood by means of an example: if two Paracetamol prescriptions and one Ponstan prescription are claimed in the same claim time frame (DSFS), then it will count as two unique types of drugs and will display as a 2 in the DrugCount column [1].

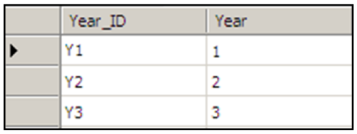
2.3 Data handling

Initial data handling can be described in terms of the extract, transform and load (ETL) procedures. ETL is a crucial part of data modelling and if it is done properly, will prevent a garbage-in-garbage-out situation for the life cycle of the project.

ETL is a three stage process that enables integration and analysis of the data from different sources and in a variety of formats. For typical hospitalisation and claims data, such as data used in the HHP, the following ETL steps were followed:

Extraction is the step where data is collected from different sources; the HHP data was divided into seven separate Comma Separated Values (.csv) files. These files, provided by the Heritage Health Provider Network, were all downloaded from the Heritage Health Prize website, combined in a SQL Server database and relationally linked to each other. Some tables were added to make the data numeric (this is explained under the Transformation section). From this, queries could be run with the preferred combination of fields. An alternative approach could be to extract .csv-files straight into programs like STATISTICA, SAS-Enterprise Miner or SPSS Clementine which also has the functionality to provide in-database access to data via low-level interfaces [13]. The last mentioned alternatives are not recommended if very complex relationships are present.

Transformation concerns the formatting, cleaning and conversion of data. When importing .csv files into Microsoft Access, care has to be taken to format each column appropriately, for example, columns with values like "10-19" have to be formatted as text, to prevent certain programs from converting it to "19-October". To cleanse the data into usable data for analysis, test fields can be converted to numeric values to make it compatible with programs like Matlab which can only accept numeric values. An example of this conversion can be seen in Table 2 where "Y1", which is text data, is substituted with "1", which is numeric data. This is not recommended when using programs with functionality such as STATISTICA, because this program has built-in functions that could manipulate text fields and save them for later use. Records containing blank entries in predictor variable cells should be deleted, as these cause problems when running queries. When the data set is small, one should be wary of deleting records in this way, but because the HHP case study has large amounts of data, deleting is an efficient way of preventing certain future problems. Predictor variables that have no variance were removed, as these variables could cause errors for certain analysis tools. It is also recommended that data only be stored once, for example, the field DaysInHospital is a sum of the length of hospital stay per claim (LengthOfStay) over the time frame of a year [9]. Consequently the derived item DaysInHospital will be preserved, but LengthOfStay will be removed.



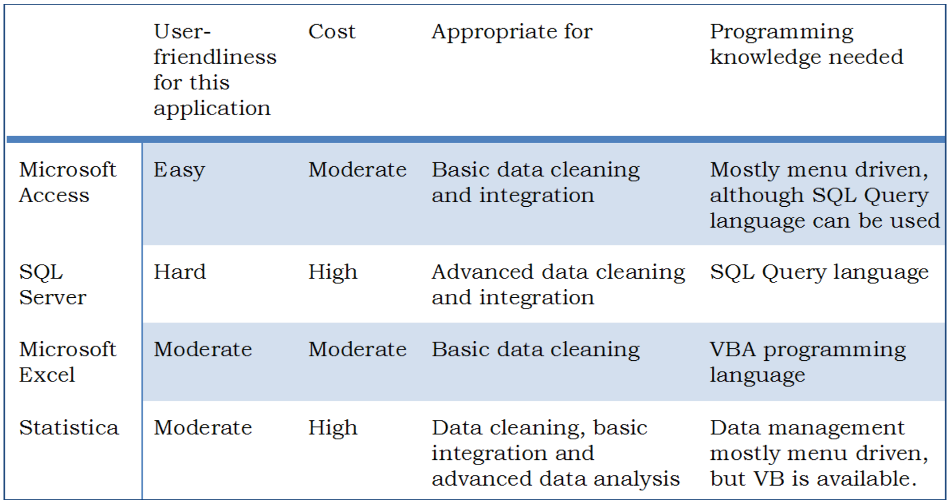
Year_ID	Year
Y1	1
Y2	2
Y3	3

Figure 2: *Example of converting text data into numerical data.*

Loading was done by firstly, importing a data sample into Excel (for preliminary analysis) to help understand the data better. This was followed by converting the bulk of the data to .csv (from Microsoft Access or SQL Server), to be copy-pasted into programs like Matlab, or STATISTICA for intensive statistical analysis.

Different technologies were tested for the ETL process and a summary of the findings for the tested technologies can be seen in Table 3. Based on the limitations of Microsoft Access and Microsoft Excel, in terms of the amount of records it can process, a conclusion can be drawn that these two programs are probably too basic for this application, and STATISTICA or SQL Server should rather be considered. The drawbacks of the last mentioned technologies are that SQL Query language has to be learnt for SQL Server and Visual Basic programming language for STATISTICA.

Table 2: *Technology ETL decision matrix*



	User-friendliness for this application	Cost	Appropriate for	Programming knowledge needed
Microsoft Access	Easy	Moderate	Basic data cleaning and integration	Mostly menu driven, although SQL Query language can be used
SQL Server	Hard	High	Advanced data cleaning and integration	SQL Query language
Microsoft Excel	Moderate	Moderate	Basic data cleaning	VBA programming language
Statistica	Moderate	High	Data cleaning, basic integration and advanced data analysis	Data management mostly menu driven, but VB is available.

3 Prediction modelling techniques

The task of the appropriate contender technique is to use the claims and member data for year $x - 1$ and the days in hospital count for year x to build a prediction model that will be used to predict for year $x + 1$.

There are certain characteristics that a prediction modelling contender technique needs to exhibit before being considered viable for the application in the HHP case study. These include:

- Multivariate modelling approach: This approach encompasses the analysis of more than one predictor variable. The input data in this study consists of an $n \times p$ (n rows by p columns) rectangular array of real numbers. Claims are summarised per member and the data set then consists of a record per member, containing characteristics of such a member. Each of the n members are thus characterised with respect to p variables. The values of the p variables may be either quantitative or a numeric code for a classification scheme [10]. All the contender techniques were chosen on the basis that they can handle multiple predictor variables.
- Linearity and non-linearity: Contender techniques should be able to handle linear as well as non-linear data, because variables are distributed linearly as well as non-linearly.
- Different variable types: As described in the previous section, the dependent variables consist of continuous variables, but the predictor variables can consist of continuous, binary, ordinal or nominal categorical variables. The contender technique should therefore, be able to handle such variations in variable types.
- Robust: This refers to the contender technique being able to model for different datasets, especially if they contain illogical data and the like (for example data sets with missing values). New data sets are made available by the competition and the technique must be able to model from these new data sets as well.
- Resistance to over fitting: Over-fitting tends to occur when more parameters than necessary are used to fit a function to a set of data [15] and causes a model to generalize

poorly to the new data. However, there are specific and different ways to avoid over-fitting with every technique used and these will be discussed further with each technique description.

- Comprehensiveness of results: This refers to the ease with which the response output of the technique can be logically understood and interpreted.
- Compatible with available technologies: It may happen that a certain technique will be able to perform prediction flawlessly in theory, but that in practice, the available technologies are limiting or too complex to use. This is an important aspect to consider in the choice of technique as well as the choice of technology. Considered technologies include: Excel and VBA, Matlab, Statistica, R, SAS and SPSS Clementine. Each tool is briefly discussed in Table 5, in terms of the: degree to which it is open source or menu driven, cost, software capabilities and the known advantages and disadvantages of each.

This study considers four multivariate prediction techniques: Multivariate Adaptive Regression Splines (MARS), Classification and regression trees (CART), Neural Networks and Ensemble Methods.

3.1 Regression modelling

Since regression is one of the simpler methods available, it is often used as the first analysis. However, basic linear regression will be insufficient as this is a complex data application and some relationships in the dataset could be linear and others non-linear. A technique used to bypass this problem is called Multivariate Adaptive Regression Splines (MARS). MARS is a nonparametric regression technique that makes no assumptions about the underlying functional relationship between the dependent and independent variables [14]. Instead, it adapts a solution to the local region of the data that has similar linear responses. MARS also has a useful characteristic in that it only picks up those predictor variables that make a sizable contribution to the prediction. MARS can also handle multiple dependent variables, although this is not required for this specific application. Outputs of this model will keep only those variables associated with the bases functions that were retained for the final model solution. If no counteract measures are taken, nonparametric models may exhibit a high degree of flexibility that, in many cases, result in over-fitting. A measure to counteract over-fitting in this kind of technique, is called pruning [14] and should be applied if this technique is used.

3.2 Classification and regression trees (CART)

The CART methodology is technically known as binary recursive partitioning [2]. It is binary because the process of modelling involves dividing a data set into exactly two nodes, by asking yes/no questions [3]. Typical questions for this application are, “Is the member male?”, “Is the member in the age group of 0-9?”, “Is the member suffering from cardiac problems?” and so on. Data is recursively partitioned by trees that divide data into more homogeneous sets, with respect to the response variable, than is the case in the initial data set. A tree keeps on growing until it is stopped by a criterion or if splitting is impossible.

CART is nonparametric, nonlinear and can analyse very complex interactions. Modelling variables are not selected in advance, but are picked by the algorithm. This model can use either categorical or continuous independent variables, or a combination of the two. It is also robust enough to handle missing or blank values and data sets with outliers will not negatively affect this model. CART is also said to be simple and easy to use and it can be incorporated into hybrid ensemble models with neural networks [13]. They often reveal simple relationships between only a few variables that could have easily gone unnoticed using other analytic techniques [14]. Timofeev and W. [16] also found CART results to be invariant to monotone transformations of its independent variables.

Some disadvantages of the decision tree models, such as CART, include:

- A small change in the value of an independent variable can sometimes lead to a large change in the predicted response.
- CART also does not capture linear structure effectively. Due to the discrete nature of the CART technique.
- A very large tree can be produced in an attempt to represent very simple linear relationships [3].
- Deciding when to stop splitting trees is a well known issue when applying CART to real life data, because real life data usually has lots of errors and random noise. An approach that can be used to address this issue is to first put a procedure in place that will stop the generation of new split nodes when improvement of the prediction is very small (Electronic Statistics Textbook, 2011).

CART trees are usually larger than is necessary and then pruned to find the optimal tree. Pruning is accomplished by testing the data set or using cross-validation or V-fold cross-validation methods. Cross-Validation can be done by comparing the tree computed from the training sample to another completely independent test sample. By doing this, one is able to see if most of the splits determined by the analysis of the training sample are essentially based on "random noise". If this is the case, the prediction for the testing sample will be poor [14].

V-fold cross-validation is accomplished by repeating the analysis many times with different randomly selected samples from the data, for every tree size (starting at the root of the tree, and comparing it to the prediction of observations from randomly drawn test samples). The best tree is the one with the best average accuracy for cross-validated or predicted values [14].

Most advanced statistical analytics software today have built in functions for CART, e.g. CART menu option in STATISTICA and treefit and treeprune functions in Matlab.

3.3 Neural Networks (NN)

Neural networks were originally based on the understanding of how the brain is structured and how it functions. This model type can do both time series prediction (univariate) and causal prediction (multivariate). The latter is required for the HHP case study.

Causal prediction (multivariate) refers to an assumption that the data generating process can be explained by the interaction of causal (cause-and-effect), independent variables [5].

Other features of neural network, according to Crone [4], are that they are non-parametric and can approximate any linear and nonlinear function to any desired degree of accuracy directly from the data. NN do not assume a particular noise process, although it is considered a flexible forecasting paradigm. Input variables are flexible: binary [1;0], nominal/ordinal [0,1,2] or metric [0.237, 7.76, ..]. This is required for the HHP case study as there are binary as well as ordinal variables. Output variables are also flexible: prediction of a single class member (binary), a multi class member (nominal) or a probability of class member (metric). NN can have any number of inputs and outputs.

NN are very powerful in terms of capability to model extremely complex functions, as is required in the HHP case study. NN learn by example and it is therefore expected that they are quite easy to use. The user invokes training algorithms to automatically learn the structure of the representative data. The level of knowledge needed to successfully apply neural networks is somewhat lower than would be the case using other, more traditional, nonlinear statistical methods. The user needs to have some knowledge of how to select and prepare data, how to select an appropriate NN, and how to interpret the results [14].

NN are not extremely robust as they do not tend to perform well with nominal variables that have a large number of possible values. This causes a problem if data is in an unusual range or if there is missing data. As mentioned earlier, missing data are not a problem in the HHP case study. NN are noise tolerant to a certain extent, but occasional outliers, far enough outside the range of normal values for a variable, may bias the training. It is best to remove outliers [14].

As with most nonparametric techniques, NN are also prone to over-fitting. Over-fitting (over-training for NN) can be prevented by validating progress against an independent test set. Validation can be done by monitoring selection error. Once the selection error starts to increase, it is an indication that the network is starting to over-fit the data, and training should be stopped. In such a situation, the network is too powerful for the problem at hand and it is recommended that the number of hidden layers should be decreased. On the other hand, if the network is not sufficiently powerful to model the underlying function, over-learning is not likely to occur, and neither training nor selection errors will drop to a satisfactory level [14].

Nowadays, most advanced statistical analytics software has built-in functions for NN, for example, the SANN menu option in Statistica and NeuroXL add-in, in Excel.

3.4 Ensemble Methods

Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. It is natural to ensemble "smooth" modelling techniques such as linear models, neural networks and MARS with decision trees in such a way that their strengths can be combined effectively [3]. The result of such a union is usually more accurate than the best of its components and it also improves the generalization of the model. Steps to building ensembles are firstly, to construct varied models, and secondly, to combine models estimates. Two popular and recommended methods for creating accurate ensembles are bagging and boosting.

Bagging, also known as bootstrap aggregation, is a method of using a variety of algorithms to model a single problem and then to use the prediction of each, as a vote. The majority ruling determines the final classification for a given case and the final model is a compromise of its component models.

Boosting, on the other hand, is a method of creating variety by weighting cases, according to which models were easier or harder to model correctly (harder cases get higher weights and vice versa). Boosting works well over a wide range of different modelling approaches [13].

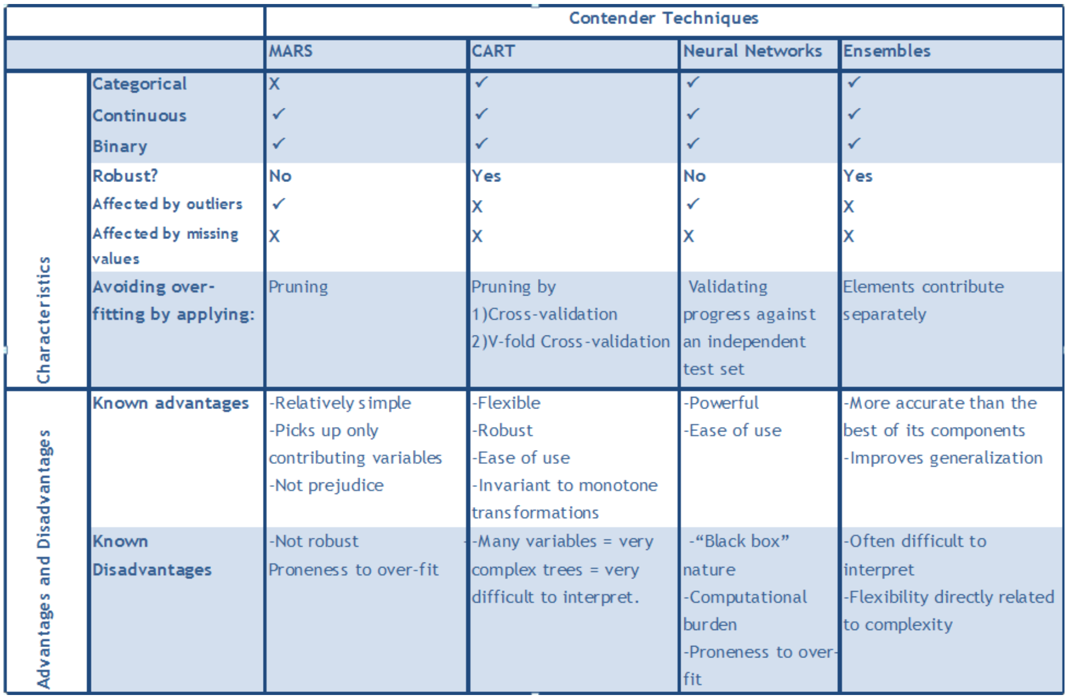
Criticisms of ensembles are that the more flexible an ensemble is built, the more complex it becomes to interpret its response. In addition to this criticism, is the expectancy that more complexity could also lead to over-fitting [13].

These days, most advanced statistical analytics software has built in functions for ensemble methods, for example, the ensemble menu option in STATISTICA and Treebagger functions in Matlab.

3.5 Comparing contender prediction modelling techniques

Table 4 shows a comparison of the four contender techniques in terms of characteristics, needed for the HHP case study application (as discussed in the commencement of Section 2). From Table 4 it can be reasoned that CART and ensemble methods are the preferred techniques to use because of their robust nature which is necessary for the HHP case study application.

Table 3: *Contender techniques decision matrix*

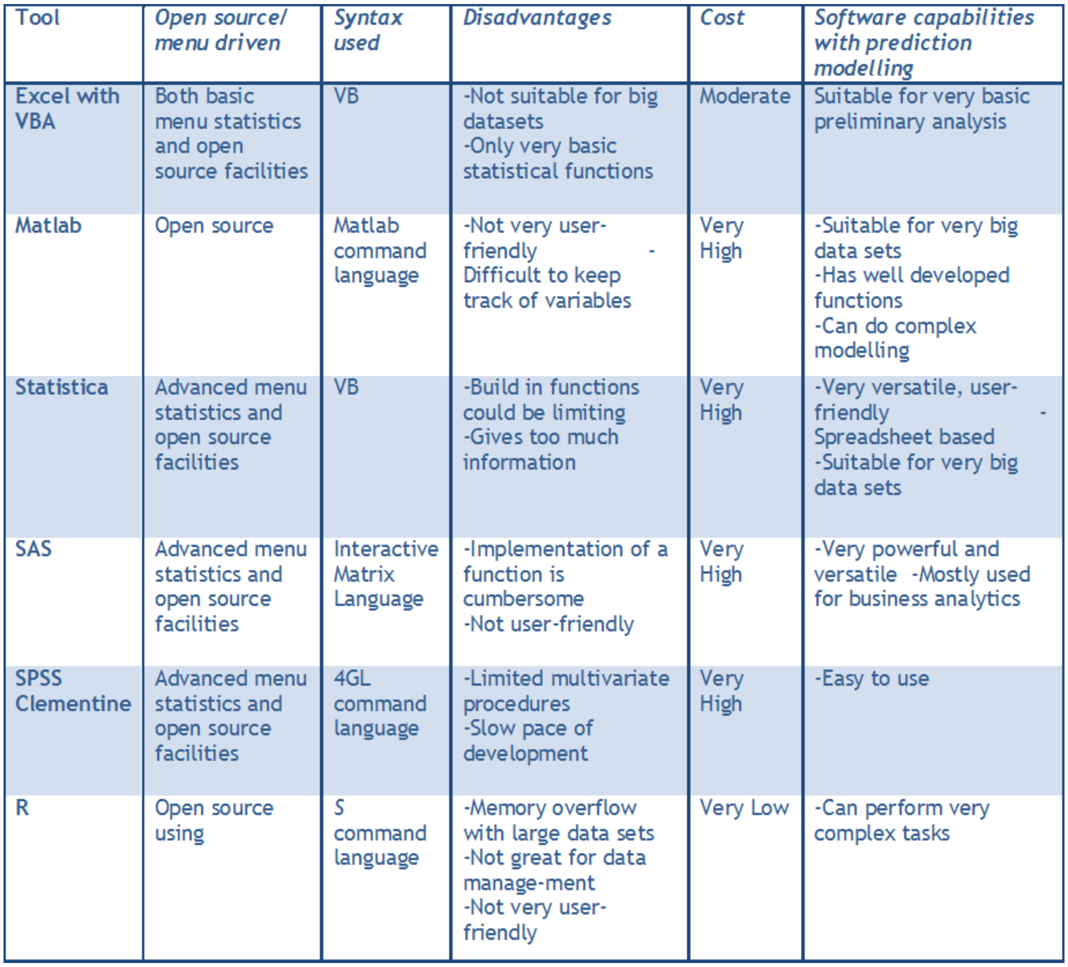


		Contender Techniques			
		MARS	CART	Neural Networks	Ensembles
Characteristics	Categorical	X	✓	✓	✓
	Continuous	✓	✓	✓	✓
	Binary	✓	✓	✓	✓
	Robust?	No	Yes	No	Yes
	Affected by outliers	✓	X	✓	X
	Affected by missing values	X	X	X	X
	Avoiding over-fitting by applying:	Pruning	Pruning by 1)Cross-validation 2)V-fold Cross-validation	Validating progress against an independent test set	Elements contribute separately
Advantages and Disadvantages	Known advantages	-Relatively simple -Picks up only contributing variables -Not prejudice	-Flexible -Robust -Ease of use -Invariant to monotone transformations	-Powerful -Ease of use	-More accurate than the best of its components -Improves generalization
	Known Disadvantages	-Not robust Proneness to over-fit	-Many variables = very complex trees = very difficult to interpret.	-"Black box" nature -Computational burden -Proneness to over-fit	-Often difficult to interpret -Flexibility directly related to complexity

4 Technologies considered

Different technologies are appropriate for the different stages of the process. A summary of the technologies used in the HHP case study is provided in Table 5.

Table 4: *Contender techniques decision matrix*



Tool	*Open source/ menu driven*	*Syntax used*	*Disadvantages*	*Cost*	*Software capabilities with prediction modelling*
Excel with VBA	Both basic menu statistics and open source facilities	VB	-Not suitable for big datasets -Only very basic statistical functions	Moderate	Suitable for very basic preliminary analysis
Matlab	Open source	Matlab command language	-Not very user-friendly - Difficult to keep track of variables	Very High	-Suitable for very big data sets -Has well developed functions -Can do complex modelling
Statistica	Advanced menu statistics and open source facilities	VB	-Build in functions could be limiting -Gives too much information	Very High	-Very versatile, user-friendly - Spreadsheet based -Suitable for very big data sets
SAS	Advanced menu statistics and open source facilities	Interactive Matrix Language	-Implementation of a function is cumbersome -Not user-friendly	Very High	-Very powerful and versatile -Mostly used for business analytics
SPSS Clementine	Advanced menu statistics and open source facilities	4GL command language	-Limited multivariate procedures -Slow pace of development	Very High	-Easy to use
R	Open source using	S command language	-Memory overflow with large data sets -Not great for data manage-ment -Not very user-friendly	Very Low	-Can perform very complex tasks

Each technology has its advantages and disadvantages and often the choice of technology depends heavily on the analysts ability and preference. Programs like SAS, SPSS and Excel seem to have limited imbedded functions of contender techniques and therefore, the use of these technologies for this application, will be quite labour intensive (hard coding will have to be done to fill in any gaps that limited functions leave). However, Statistica, Matlab and R have sufficient built-in functions for these complex modelling applications. Last mentioned programs also have appropriate visualisation resources and are powerful enough to handle the HHP case study data set size and complexity. Statistica, Matlab and R were thus identified as the preferred tools for this application.

5 The road ahead

This paper provided comparisons of contender techniques and technologies to use in the development of hospitalisation prediction models (such as those for the HHP), based on theoretical knowledge and study. Potential pitfalls were discussed, mostly concerning the implementation of the ETL process, and aspects to keep in mind, when using each technique and technology, were highlighted. Recommendations concerning the choice of technologies for ETL, are SQL Server or Statistica and for prediction model building, Statistica, R or Matlab. The next step is to pilot test these techniques and technologies, so as to verify the suitability, functionality and practicality of each, and to determine which one gives the most appropriate response. Based on currently available data, this ought to be the preferred strategy, when developing PPAAs.

Other prediction modelling techniques that can also be researched for future use in this application are probabilistic Bayesian methods as well as Structural Equations Methods (SEM).

Bibliography

[1] Arbuckle (2011). Drugcount? count of drugs or prescriptions? Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited July 7th, 2011].

[2] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

[3] Brookes, R. and Kolyshkina, I. (2002). Data mining approaches to modelling insurance risk. Paper presented at the IXth Accident Compensation Seminar.

[4] Crone, S. (2005). Forecasting with artificial neural networks. *Journal of Intelligent Systems*, 14:99–122.

[5] Galkin, I. and Lowell, U. (2011). Crash introduction to artificial neural networks. Unpublished course material. Available oneline: http://ulcar.uml.edu/ iag/CS/Intro-to-ANN.html [Cited August 8th , 2011].

[6] Ganster, D., Fox, M., and Dwyer, D. (2001). Explaining employees' health care costs: A prospective examination of stressful job demands, personal control, and physiological reactivity. *Journal of Applied Psychology*, 86:954.

[7] Heritage Provider Network Health Prize (2011). Available online: http://www.heritagehealthprize.com/c/hhp [Cited August 7th, 2011].

[8] Howard, J. (2011). Male pregnancy? Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited June 10th, 2011].

[9] Igor, I. (2011). Data problems: inpatient hospital stays w/o lengthofstay & outpatient los. Available online: http://www.heritagehealthprize.com/c/hhp/forums [Cited June 10th, 2011].

[10] Jobson, J. (1991). *Applied multivariate data analysis: regression and experimental design.* Springer, Faculty of Business University of Alberta Edmonton, Alberta, Canada.

[11] Lee, C., Parr, R., and Yang, W. (1988). Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37:785.

[12] Miyata, H., Hashimoto, H., Horiguchi, H., Matsuda, S., Motomura, N., and Takamoto, S. (2008). Performance of in-hospital mortality prediction models for acute hospitalization: hospital standardized mortality ratio in japan. *BMC health services research*, 8:229.

[13] Nisbet, R., Elder, J., Elder, J., and Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Academic Press, London.

[14] StatSoft Inc. (2011). Electronic statistics textbook. Available online: http://www.statsoft.com/textbook/ [Cited August 23th, 2011].

[15] Steig, E. (2009). On overfitting. Available online: http://www.realclimate.org/index.php/archives/2009/06/on-overfitting/ [Cited September 17th, 2011].

[16] Timofeev, R. and W., H. (2004). Classification and regression trees (cart) theory and applications. Unpublished MSc thesis, Humboldt University, Berlin.

Proceedings of the 2011 ORSSA Annual Conference
pp. 124–136
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

Solving A Multi-Objective Micro Milling Problem Using Metaheuristics

EC Essmann* TE Lane-Visser†

Abstract

The standard management challenge of balancing the conflicting objectives of time, cost and quality applies to the manufacturing domain in general, and, hence, to the field of micro milling. Assuming that quality levels are non-negotiable, the time/cost relationship becomes an important consideration. Managers typically want products manufactured at the lowest cost and in the shortest time. This is not always possible, however, because reducing the machining time typically reduces the tool life, which in effect, increases the cost. This is especially relevant in the case of micro milling, where tooling costs are relatively high. Understanding the relationship between machining time and machining costs, and how these objectives are influenced by the cutting parameters of the machine will be very useful from a management perspective. Cutting parameters include cutting speed, feed per tooth. The purpose of this paper is to determine the relationship between machining time and costs for bipolar plates, calculated in terms of the machine cutting parameters. As the solution space is characterised by all possible combinations of these parameters and the parameters are bounded continuous variables, the solution space for this problem is infinitely large. The problem is further complicated by the inclusion of two nonlinear objectives. Multi-Objective Simulated Annealing (MOSA) was chosen as the technique with which to solve this problem. The paper discusses the development of the MOSA algorithm, the shape of the resulting pareto front and the interpretation of the results in a micro milling manufacturing environment.

Key words: Prediction modelling, hospitalization

1 Introduction

The background to this paper has its roots in the field of hydrogen fuel cell technology, a dynamic and strategically important research field for the South African economy. In a global move toward environmental awareness, governments are focusing on the development of sustainable energy systems as a means to reduce their countrys carbon footprint and ensure

*Corresponding author: Stellenbosch University, South Africa, email: 14808706@sun.ac.za
†different Stellenbosch University, South Africa, email: tanyav@sun.ac.za

the security of energy supply. One sustainable energy system is the use of hydrogen as an energy carrier, which combined with fuel cell technology, shows much promise as a means of electricity production [3].

Hydrogen and other fuel cells use platinum as a catalyst to promote the chemical reactions at work. Platinum is a rare mineral with known reserves in only five countries. South Africa the leading supplier of these five countries with 80% of the worlds known platinum reserves. This places South Africa at the centre of the worlds intensifying research efforts towards a hydrogen economy [3].

A vital component in fuel cell stacks is the bipolar plate. Bipolar plates allow several fuel cells to be connected in series, thereby increasing the achievable voltage and greatly improving the potential usefulness of fuel cells.

The production cost of hydrogen fuel cells is driven significantly by the manufacture of these bipolar plates. This is because bipolar plates are complex in design and account for most of the mass and volume in a fuel cell stack. The need, therefore, exists to find materials and manufacturing techniques that result in the cost effective production of these components.

Several techniques exist that could be used for the manufacture of bipolar plates. Micro milling is one such technique that shows promise, especially for small to medium batch sizes. Micro milling is defined as the milling of components with two or more dimensions in the sub millimetre range. This technique is characterised by the ability to manufacture complex three dimensional, free form geometries of small to medium batch sizes cost effectively. This manufacturing technique, therefore, warrants further investigation

The standard management challenge of balancing the conflicting objectives of time, cost and quality applies to the manufacturing domain in general, and, hence, to the field of micro milling. Assuming that quality levels are non-negotiable, the time-cost relationship becomes an important consideration.

Understanding the relationship between machining time and machining costs, and how these objectives are influenced by the cutting parameters of the machine will be very useful from a management perspective. Cutting parameters include cutting speed (m/min), feed per tooth (m) and depth of cut (mm). The purpose of this paper is to facilitate understanding of the relationship between machining time and costs for bipolar plates. This is stated formally in terms of the following objectives:

- The primary objective is to determine the relationship between machining time and cost where these are calculated in terms of cutting parameters.
- The secondary objective is to demonstrate the influence that external factors have on the relationship. External factors include overhead rates, price of micro tools and machine capabilities.

2 Problem formulation

In formulating this problem, it should be noted that only the main feature of the bipolar plate design is considered. This is because this feature, known as the flow field channels, accounts for approximately 70% of the machining time and cost. The machining time and cost objectives can be formalised as follow:

$$T_{max} = \frac{N_{ch} \times L_{ch}}{F_R} + (T_{ref} + T_{tc}) \times z \tag{1}$$

$$C_{mach} = T_{mach} \times \frac{OR}{60} + TC \times z \tag{2}$$

where

T_{mach} $\triangleq$ Time (min) to machine channel feature.
L_{ch} $\triangleq$ average channel length (mm).
T_{ref} $\triangleq$ time (min) to reference the tool to the plate.
T_{tc} $\triangleq$ time (min) to change tools.
F_R $\triangleq$ feed rate (mm/min).
C_{mach} $\triangleq$ cost (R) to machine the channel features.
OR $\triangleq$ overhead rate (R/hr).
TC $\triangleq$ tool cost (R/tool).

The term z, in Equations (1) and (2) above, represents the number of tools used per plate, according to

$$Z = \frac{N_{ch} \times L_{ch}}{\frac{L_T}{d \times D}} \tag{3}$$

where

$$L_T = e^{5.842+5.275d-0.000348v^2-0.00506{f_t}^2-1.724d^2+0.00270vf_t} \tag{4}$$

and

L_T $\triangleq$ tool life (mm^3/tool).
d $\triangleq$ axial depth of cut (mm).
D $\triangleq$ tool diameter (mm).
v $\triangleq$ cutting speed (m/min).
f_t $\triangleq$ feed per tooth (μm)

For the purpose of this application, the parameters d and D are assigned a value of 0.7mm. This is dictated by the design of the bipolar plate channels. That is, the axial depth of cut, d, corresponds to the channel depth and the tool diameter, D, corresponds to the channel width. Equation (4) above deserves special mention.

Equation (4) is an empirically determined function, developed by Essmann and Van Schalkwyk [1], which relates tool life in terms of volume of material removed (mm3) to cutting parameters v, f_t and d. This tool life formula was developed for the particular situation at hand. That is, the correct material (polymer-graphite composite) and micro end mills (solid carbide) were used.

T_{ref} and T_{tc}, from Equation (1) are typical setup times applicable to micro milling, while N_{ch} and L_{ch} are features of the bipolar plate design. For the purpose of this paper, the following constants are used:

$$\begin{aligned} T_{tc} &= 34.3/60 \text{ min} \\ T_{ref} &= 46.4/60 \text{ min} \\ N_{ch} &= 60 \\ L_{ch} &= 309 \text{ mm} \end{aligned}$$

The problem formulated above is subject to the constraints placed on it by the machine. That is, the machine cannot exceed its maximum achievable feed rate (F_R(mm/min)) and rotational velocity (n(rev/min)) such that,

$$\frac{v \times 1000}{\pi \times D} = n \le n_{Max} \tag{5}$$

$$\frac{v \times f_t \times 2}{\pi \times D} = F_R \le F_{RMax} \tag{6}$$

where n_{Max} and F_{RMax} are the maximum achievable rotational velocity and feed rate respectively. These constraints are prescribed by the capability of the machine in use. The value of 2 in Equation (6) above represents the number of flutes/cutting edges on the micro end mill. This value is prescribed by the type of end mill used. For the purpose of this machining application, a two-flute end mill is recommended. Further, as stated previously, the relevant tool diameter, D, has the value of 0.7mm.

The solution space for this problem is infinite because it is characterised by all possible combinations of the cutting parameters, which are all bounded continuous variables. The problem is further complicated with the inclusion of two non-linear objectives. This suggests the use of a metaheuristic to solve this problem. Although metaheuristics are not the only feasible approach to solving this problem, they are amongst the most versatile.

According to Talbi [8], metaheuristics represent a family of approximate optimisation techniques. These techniques are approximate because they do not guarantee exact optimal solutions. However, metaheuristics have gained much popularity over the last two decades, according to Talbi [8], because they do provide acceptable solutions in a reasonable amount of time. Further, their use in a variety of applications such as engineering design and supply chain management, demonstrate their efficiency and effectiveness in solving complex problems [8].

Using metaheuristics to optimise cutting parameters in micro milling has been seen before in literature. One such example was done by Sreeram et al. [7]. They used a genetic algorithm to optimise tool life and, in effect, machining cost in terms of cutting parameters for a particular component. Their results indicated that the optimal cutting parameters were different from those recommended by the suppliers, especially the depth of cut parameter. Sreeram et al. [7] showed that the use of an optimisation algorithm could increase machining efficiency and reduce production costs by the correct selection of machining parameters. This is due to the effect of machining parameters on tool life and machining time, two significant cost drivers in micro milling.

The algorithm solution developed by SSreeram et al. [7] is specific to the component and type of material used. Therefore, it is not suitable to implement their solution directly to this situation. Since the problem formulated in this paper is unique, it was decided to build a new algorithm to solve it.

Modelling costs and cost drivers in micro milling problems is also common in literature. Further, several researches have recognised the contribution of tooling to manufacturing cost. This is due to the brittle and costly nature of micro end mills. As such, several attempts have been made to model tool life in terms of cutting parameters. This is done for the purpose of optimising manufacturing costs.

Prakash et al. [6], developed an empirical tool life model using a 1mm diameter tool to machine copper under dry cutting conditions. In this study, axial depth of cut, cutting speed and feed rate are considered. It was found that only axial depth of cut and cutting speed are relevant to the progression of flank wear; while feed rate had no influence.

Filiz et al. [2] on the other hand came to a different conclusion. They studied tool life on a 254μm diameter tool. Their results showed that tool life depends highly on feed. That is, faster feed rates led to reduced wear relative to slower feed rates.

3 Solution building

The solution developed for the purpose of this paper is derived from the simulated annealing algorithm. Simulated annealing gets its name from the process of cooling molten metal; also known as annealing. If the metal is cooled rapidly, the atoms are forced to settle into a random structure. If, however, the temperature is decreased slowly, allowing the atoms enough time to settle, a strong crystal lattice structure is formed [8]. This process is analogous to an optimisation problem in the sense that search pattern needs to shift slowly and carefully from a wide-ranging pattern to a localised pattern.

The algorithm built for the purpose of this paper is based on the generic algorithm for multi-objective simulated annealing presented by Nam and Hoon Park [5].

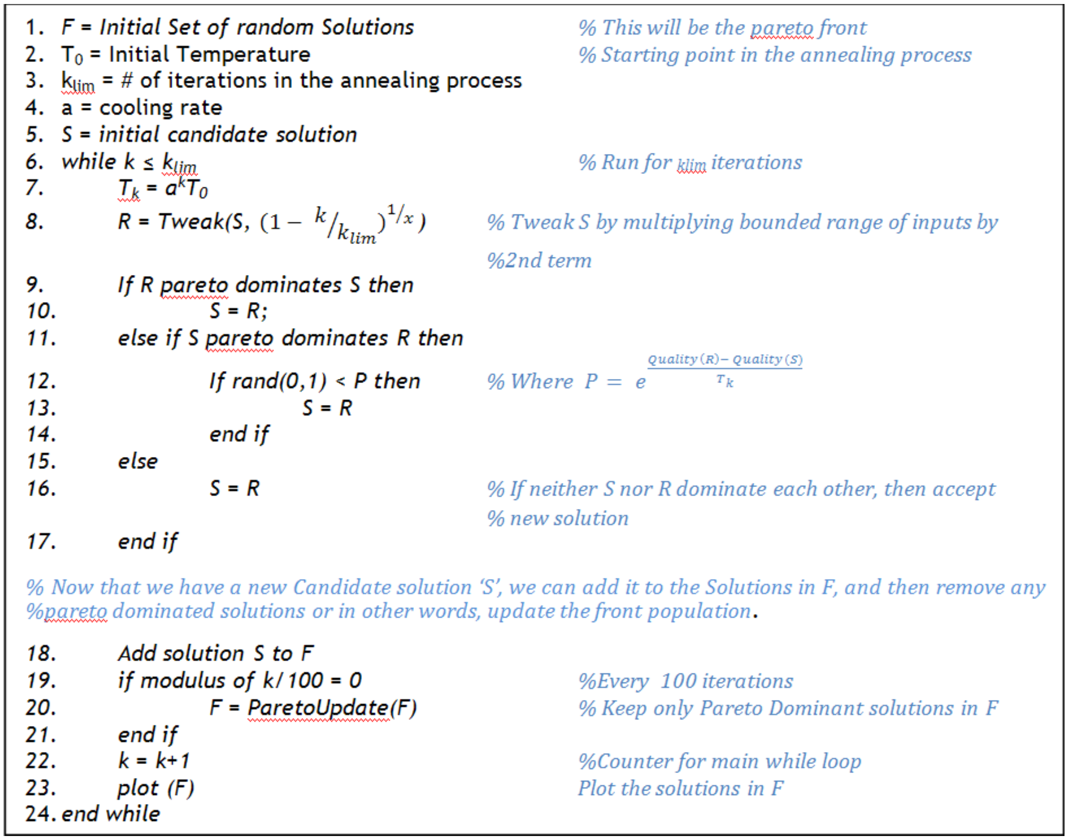
3.1 Simulated Annealing Pseudo Code

Consider Figure 1, which presents the pseudo code of the algorithm built for the purpose of this paper.

Lines 1–4 define a few variables, which are used in the algorithm. F, stores a set of solutions. Although F is initially empty, it will, as the algorithm progresses, be transformed into the pareto front (set of non-dominated solutions). T_0 stores the initial temperature of the annealing process. klim defines the number of iterations in the process. a is defined as the rate of cooling in the annealing process when it is considered in Equation (7) below (where k is the iteration counter of the algorithm). Further elaboration on this and its influence in the algorithm is presented in section 3.2.

$$T_k = a^k \times T_0 \tag{7}$$

Lines 5 populates S with a initial candidate solution chosen randomly from within the solution space. The candidate solution, S, is coded as a vector containing the input



```
F = Initial Set of random Solutions                  % This will be the pareto front
T_0 = Initial Temperature                            % Starting point in the annealing process
k_lim = # of iterations in the annealing process
a = cooling rate
S = initial candidate solution
while k ≤ k_lim                                      % Run for klim iterations
    $T_k = a^k T_0$
    R = Tweak(S, $(1 - {}^{k}/_{k_{lim}})^{1/x}$ )   % Tweak S by multiplying bounded range of inputs by
                                                     %2nd term
    If R pareto dominates S then
        S = R;
    else if S pareto dominates R then
        If rand(0,1) < P then                        % Where $P = e^{\frac{Quality(R) - Quality(S)}{T_k}}$
            S = R
        end if
    else
        S = R                                        % If neither S nor R dominate each other, then accept
                                                     % new solution
    end if
% Now that we have a new Candidate solution 'S', we can add it to the Solutions in F, and then remove any
%pareto dominated solutions or in other words, update the front population.
    Add solution S to F
    if modulus of k/100 = 0                          %Every  100 iterations
        F = ParetoUpdate(F)                          % Keep only Pareto Dominant solutions in F
    end if
    k = k+1                                          %Counter for main while loop
    plot (F)                                         Plot the solutions in F
end while
```

Figure 1: *Simulated Annealing Algorithm*

parameter (v, f_t and d) values as well as the objective values corresponding to those inputs. Similarly, F is coded as an array containing a set of S-type vectors.

Line 6 is the initiation of the main loop with k being the iteration counter.

Line 7 calculates the 'temperature' for that iteration. This is a function of the cooling rate a, the iteration number k and the initial temperature T_0 as indicated in Equation (7).

Line 8 determines the new candidate solution R by tweaking the current candidate solution S. The second input term in the Tweak function defines the maximum half-range noise to be added to each variable in the tweaking process according to the following formula (where t is the second term):

$$Maximum\,Half\,Range\,Noise = [Upperbound - Lowerbound] \times t \tag{8}$$

Usually, the range of uniform noise is fixed at a predefined value by setting t to a fixed value. However, for the purpose of this paper the range of uniform noise is decreased slowly as the simulated annealing algorithm progresses. This is achieved by the following formula.

$$t = (1 - \frac{k}{k_{lim}})^{\frac{1}{x}} \tag{9}$$

Further elaboration on the formula above is presented in section 3.2. This tweaking method is derived from the Bounded Uniform Convolution method described by Luke [4]. The Tweak function further ensures that the new candidate solution satisfies the problem constraints as prescribed by the machine capacity, according to Equations (5) and (6).

Lines 9-17 decide whether to accept the new candidate solution R by replacing the old solution S with the new solution R. The decision is made based on the following criteria. If R pareto dominates S, then replace S with R. Else, if S pareto dominates R, then replace S with R with probability P. Then finally, if neither S nor R dominate each other, then replace S with R. This ensures that the algorithm does not stagnate on a reasonably good solution but is willing to jump to an alternative solution if neither one dominates the other. By Definition, a solution A pareto dominates another solution B if A is at least as good as B for all objectives and better than B for at least one objective.

The probability P, which largely controls the behaviour of the algorithm, is a function of the quality of the solutions S and R. Quality, is derived from Luke [8] and is a measured relative to a set of solutions. Firstly, strength defined as the number of solutions in a set that the original solution dominates. Quality is then determined according to the following:

$$Quality(B) = \frac{1}{1 + \sum_{g \in G\,that\,pareto\,Dominates\,B} Strength(g)} \tag{10}$$

Quality is thus defined as the inverse of the sum of all the solutions in the set that pareto dominates the original solution. This method provides an objective way of defining the quality of a solution relative to a set of solutions.

Line 18 adds the new candidate solution S to existing front F.

Lines 19–21 Every 100 iterations, the pareto Front is updated by removing any dominated solutions. The reason this happens every 100 iterations, is so that the user can see where and how the algorithm is searching in order to gain insight into the operation of the model.

Lines 22–24 are the closing of the main while loop where the counter k is increased and the current solutions in the set F are plotted.

3.2 Controlling the Algorithm - Exploration versus Exploitation

An important feature in any metaheuristic is the algorithm behaviour in terms of exploration versus exploitation. It is preferable to encourage more exploration in the early stages of an algorithm while later stages should be geared more towards exploitation. This allows the proper identification (via exploration) and clarification (via exploitation) of optimal solutions in the solution space. The behaviour of the algorithm should therefore shift from exploration to exploitation as the algorithm progresses.

Two mechanisms have been identified to control the behaviour of the algorithm in terms of exploration versus exploitation. These are the probability, P, in terms of a, and the half range noise, R_d, in terms of x. These are discussed presently.

The probability, P, is defined by the formulae

$$P = e^{fracQuality(R) - Quality(S)T_k} \tag{11}$$

$$T_k = a^k \times T_0 \tag{12}$$

The parameter, P, determines the probability that a worse solution will be accepted over a better solution. This effectively, controls the amount of exploration in the search, by allowing the search sometimes to move downhill. The higher the value of P, the more exploration in the algorithm, while the lower the value of P, the more exploitation. The parameter, T_0 (set to a value of 10), represents the approximate starting value for Tk where $k = 0$.

The parameter, a, in Equation (10) above controls the rate at which P decreases. In the context of simulated annealing, a is known as the cooling rate. It should be noted here that the parameter, k, refers to the algorithm iteration counter, as described in Section 3.1. Choosing a larger value for a, allows P to decrease at a slower rate, effectively allowing more exploration in the algorithm. Conversely, choosing a smaller value for a allows P to decrease at a faster rate, affording more exploitation to the algorithm.

The half range noise, R_d, is defined in terms of x according to the formula

$$R_d = [Upper\,Bound - Lower\,Bound] \times (1 - \frac{k}{k_{lim}})^{\frac{1}{x}} \tag{13}$$

where $Upper\,Bound$ and $Lower\,Bound$ represent the upper and lower bounds for each input variable respectively and k and k_{lim} represent the iteration counter and maximum number of iterations respectively.

The half range noise, R_d, is the maximum amount that any input variable can change when tweaking a candidate solution. The greater the value of R_d, the more noise can be added to input variables and, consequently, the more the output variables can jump around in the solution space.

The parameter x controls the rate at which R_d decreases. Choosing a larger value for x allows R_d to decrease more rapidly and, in so doing, affords more exploitation to the algorithm. Conversely, choosing a smaller value for x slows down the decrease of R_d, thereby affording more exploration to the algorithm.

Following the arguments above, the behaviour of the algorithm, in terms of exploration and exploitation, is controlled by the selection of parameter values for a and x. The correct selection of these parameter values is determined by experimentation and the judgement of the user. For the purpose of this problem, values of 10 and 0.9986 for x and a respectively, were found to yield good results and were used in the experiments following.

4 Analysis

4.1 Initial Analysis and the Nature of the Time-Cost Relationship

For the initial analysis, two configurations of model parameters and constraints were used. Figure 2 below shows the result of each configuration. The figure plots solutions found by the simulated annealing algorithm on a Cartesian plane with machining time and cost on the x and y-axes respectively. Further, the optimal solutions are shown in blue, while the green dots indicate non-optimal solutions found by the model. Notice here the only difference between the set of model parameters and constraints is the feed rate limit. It is 1664 and 4000 mm/min for graphs on the left and right respectively.

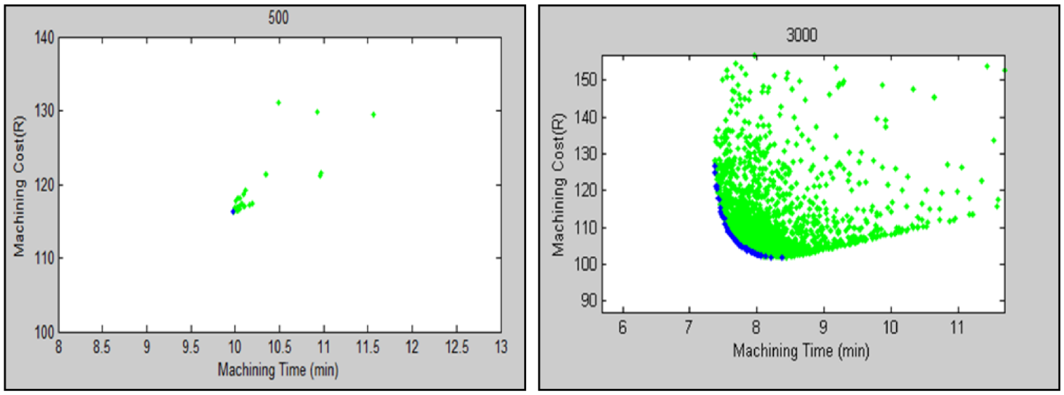


Figure 2: *Initial Analysis. Overhead Rate = R250/hr, Tool Cost = R70, Rotational Velocity Limit = 60000 rpm, Feed Rate Limit = 1664 mm/min (left), Feed Rate Limit = 4000 mm/min. (Right)*

The solutions found on Figure 2 (left) appear as though they tend towards the bottom left-hand corner of the graph above, without forming a pareto front. The algorithm, thus, tends towards one optimal solution in terms of both time and cost. This would indicate that the objectives are, in fact, non-conflicting for the particular constraints. In addition, the configuration on the left contains few feasible solutions indicating that algorithm struggled to accept new solutions in this instance. This configuration was run for 500 iterations only because it stagnated after only a few hundred.

The solutions found on Figure 2 (right) indicate a distinct trade-off curve or pareto front. Contrary to the initial analysis, this suggests that the objectives are in fact conflicting. In addition, this graph displays numerous feasible solutions indicating that the algorithm did not struggle to accept new solutions with this configuration. This configuration was allowed to run for 3000 iterations.

The difference between the shape of the graphs on the left and right of Figure 2 suggest an interesting phenomenon. It suggests that the nature of the relationship between machining time and cost is dependent on the range within which the cutting parameters are selected. This range of cutting parameters is effectively controlled by the feed rate constraint of Equation (6).

The change in the nature of the time-cost relationship is an interesting phenomenon and has to do with the way in which tool life is defined. It should be noted here, from Equation (6), that the feed rate is always increasing for increasing cutting speed and feed per tooth. Consider Figure 3, which shows a plot of tool life in terms of volume of material removed (mm3) on the z-axis, for a fixed depth of cut, with feed per tooth (ft (m)) on the x-axis and cutting speed (v (m/min)) on the y-axis.

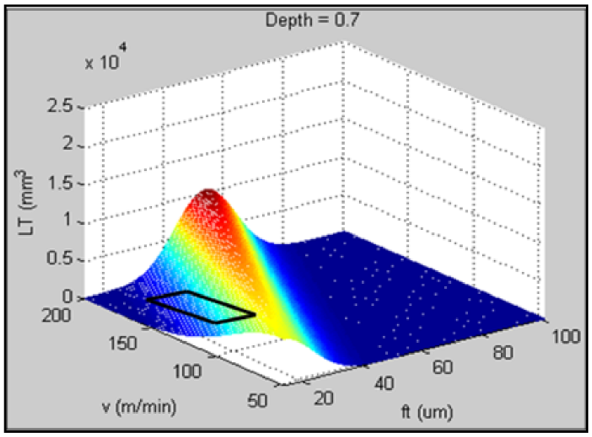


Figure 3: *Tool life Plot for Fixed Depth*

Consider now the behaviour of tool life within the narrow parameter range, demarcated by the rectangle, versus the behaviour of tool life over the whole parameter spectrum. Within the narrow range, tool life always increases for an increase in feed per tooth and cutting speed. However, over the whole spectrum, tool life can also be decreasing for increased cutting speed and feed per tooth. The inconsistency in the behaviour of tool life in terms of the cutting parameters is the reason for the changing nature of the time-cost relationship, illustrated by Figure 2. That is, the change from a non-conflicting to a conflicting relationship as the solution space is broadened.

4.2 Cost Parameter Sensitivity

The changing nature of the time-cost relationship is an interesting phenomenon from an academic perspective, but provides little practical value. This is because the feed rate required for a conflicting time-cost relationship is approximately 4000 mm/min, an infeasible value for micro milling machines, whose capacities range between 1000 and 2000 mm/min.

An important consideration for managers, however, is the effect that changing cost parameters have on the machining cost, as defined in Equation (2). Cost parameters include overhead rate (OR (R/hr)) and tool cost (TC (R/tool)).

Figure 4 demonstrates the effect of changing cost parameters on machining cost. This sensitivity analysis was performed by finding the optimal solution for each combination of overhead rate and tool cost using the simulated annealing algorithm. The changing colours of Figure 4 represents the changing minimum machining costs as the cost parameters are varied. The dark blue colour represents a low minimum cost while the red represents a high minimum cost. This relationship is formalised in Equation (12), which was determined using linear regression analysis.

$$C_{Mach} = 0.008 + 0.166 \times OR + 1.068 \times TC \tag{14}$$

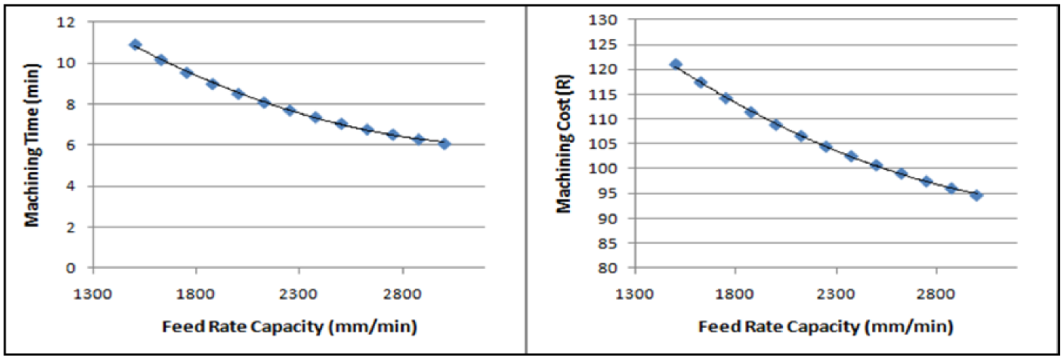


Figure 4: *Cost Parameter Sensitivity*

4.3 Feed Rate Limit Sensitivity

Another important consideration, from a management perspective, is the effect of the feed rate constraint, from Equation (6), on the optimal solution. The feed rate constraint is prescribed by the maximum achievable feed rate of the machine used. A more expensive machine would typically be able to achieve higher feed rates. Therefore, it would be useful to known what improvements in time and cost can be achieved by increasing the feed rate capacity.

To address this issue a sensitivity analysis was performed that relates the improvement in time and cost achieved to increased feed rate capacity. The results are shown in Figure 5. Using the simulated annealing algorithm, optimal solutions in terms of machining time and cost were determined for changing algorithm configurations. For each new configuration, the feed rate constraint was increased and the resulting optimal solutions recorded.

$$T_{Mach} = 21.819 - 0.0094 \times F_{RMax} + (1 \times 10^{-6}) \times F_{RMax} \tag{15}$$

$$C_{Mach} = 172.84 - 0.0438 \times F_{RMax} + (6 \times 10^{-6}) \times F_{RMax} \tag{16}$$

The relationship, obtained using linear regression analysis can be formalised by Equation (13) and (14). This result indicates that significant improvements in both time and cost can be achieved by increasing the feed rate capacity of the machine. This is an important managerial consideration because it affects profitability and production lead-time, both of which are strategically and operationally important.

5 Conclusions

This paper demonstrated the machining time-cost relationship in terms of cutting parameters for the micro milling of bipolar plates. Further, the influence of external factors such as overhead rates, tool cost and machine capabilities on machining time and cost were demonstrated. A few important issues that were identified during this process are discussed presently.

Tool cost plays an important role in machining cost. This parameter is determined by external factors, largely out of the control of the enterprise. Therefore, it is important to consider how tool costs might be affected by factors such as exchange rates, import duties etc. as micro tools are imported.

Further, careful consideration must be given to the selection of the cutting parameters. Failure to select the optimal combination will result in sub-par performance in terms of time and cost. It is not sufficient simply to maximise the feed rate. Rather, the correct combination of feed per tooth and cutting speed must be selected that considers both feed rate and tool life.

Still, further, it is possible to significantly improve machining time and cost by increasing the feed rate capacity of the milling machine. Purchasing a machine with a higher feed rate capacity will also affect the overhead rate because capital costs are typically apportioned via the overhead rate.

The decision maker should also bear in mind that extrapolation of Figure 5 is not possible past a certain point. That is, the nature of the time-cost relationship changes as the feed rate constraint is increased, as described in Section 4.1.

The decision to take regarding whether or not to manufacture bipolar plates using micro milling, and the decision on the correct level of capital investment is left to the decision maker. By no means does this paper form a definitive study. It is the opinion of the author that a holistic approach must be taken during decision-making. This requires the consideration of several factors outside the scope of this paper. The intention of this paper is to serve as a decision making guideline, providing insight into the nature of the problem. This paper further serves to initiate a broader investigation into the economic feasibility of manufacturing bipolar plates using micro machining.

Bibliography

[1] Essmann, E. and Van Schalkwyk, T. (2011). Micro milling of bipolar plates - a tool life model. In *ISEM 2011 Proceedings*, pages 102–113. Stellenbosch: ISEM.

[2] Filiz, S., Conley, C., Wassermn, M., and Ozdoganlar, O. (2007). An experimental investigation of micro-machinibilty of copper 101 using tungsten carbide micro-endmills. *International Journal of Machine Tools & Manufacture*, 1:1088–1100.

[3] HySA (2009). Hydrogen & fuel cell technologies research, development & innovation strategy. Retrieved February 9, 2011, from http://hydrogen.qsens.net/.

[4] Luke, S. (2010). Essentials of metaheuristics.

[5] Nam, D. and Hoon Park, C. (2010). Multi-objective simulated annealing: A comparative study to evolutionary algorithms.

[6] Prakash, J., Rahman, M., Senthil Kumar, A., and Lim, S. (2002). Model for predicting tool life in micro milling of copper. *Chinese Journal of Mechanical Engineering*, 15:115–120.

[7] Sreeram, S., Senthil Kumer, A., Rahman, M., and Zaman, M. (2006). Optimization of cutting parameters in micro end milling operations under dry cutting conditions using genetic algorithms. *International Journal of Advanced Manufacturing Technology*, 30:1030–1039.

[8] Talbi, G. (2009). *Metaheuristics: From Design to Implementation.* New Jersey: John Wiley & Sons Inc.

Proceedings of the 2011 ORSSA Annual Conference
pp. 137–145
http://www.orssa.org.za/XXXXX

ORSSA Proceedings
ISBN
978-0-7972-1351-7
©2012

The 2011 municipal elections in South Africa and new trends since the 2009 national elections

JM Greben* CD Elphinstone† JP Holloway‡

Abstract

The CSIR has been involved in South African election night predictions since 1999 using a cluster prediction model based on the segmentation of the electorate according to voting behavior. In this paper these clusters are exploited in another way. Different clusters are related to different demographic groups, and an analysis is made how these different groups change their affiliation between subsequent elections. The changes in affiliation are determined by calculating a trend matrix, a new tool in elections that was introduced by one of the authors a few years ago. By comparing trend matrices between municipal (2006, 2011) and national elections (2004 and 2009) one can establish whether the observed trends are incidental or have a more generic character. It is felt that a better understanding of the voter behavior through such analyses can enhance the value of elections and thereby promote democracy.

1 Introduction

This paper aims to provide detailed insights in the voter behavior in South Africa over the last eight years, using a combination of cluster and trend matrix techniques. Clustering techniques have been used extensively by the CSIR election research group since 1999 [3, 2]. This group is involved in electionnight predictions and has demonstrated the accuracy and effectiveness of the cluster segmentation of the South African electorate through its successful election night predictions in all elections since 1999. The clustering is based on prior election results, so that the clusters used in 2000 were based on the 1999 election results, the clusters used in 2004 were based on the 2000 election results, while since then the 2004 results have been used as a basis for the subsequent election analyses in 2006, 2009 and 2011. Using the census results from the 2001 South African census, one can also determine the demographic characteristics of these clusters, and thereby relate voter behavior to demographics. By keeping the clusters the same since 2004 one can compare the election behavior in different sectors of the population. This

[*] Corresponding author: CSIR Built Environment, South Africa, email: jgreben@csir.co.za

[†] CSIR Built Environment, South Africa, email: celhpin@csir.co.za

[‡] CSIR Built Environment, South Africa, email: jhollowa@csir.co.za

comparison is greatly facilitated and enhanced by the availability of trend techniques, which were recently developed by one of the authors [1]. In the next section, 4 of the 20 clusters will be described in detail as they represent representative and distinct parts of the South African electorate. After the introduction of the cluster technique in Section 3, the four segments of the electorate are the basis for four case studies using these trend techniques in Section 4. These studies provide interesting insights in the behavior of the South African electorate since the first democratic elections.

2 Segmentation of the South African Electorate

In the following the demographics of the individual clusters are illustrated for four of the twenty clusters of the 2004 cluster model. The demographic analysis was carried out on the basis of the 2001 census, but uses the latest information on the number of registered voters in the relevant voting districts. It also shows the equivalent 2009 election results for each cluster, rather than the 2004 results which were used in the construction of the clusters. The biggest cluster (33.2 percent of the electorate in 2009) shown in Fig. 1, is based on a large percentage (88.2 percent) vote for the African National Congress (ANC), which is considerably higher than the national average (66 percent) in the 2009 elections.

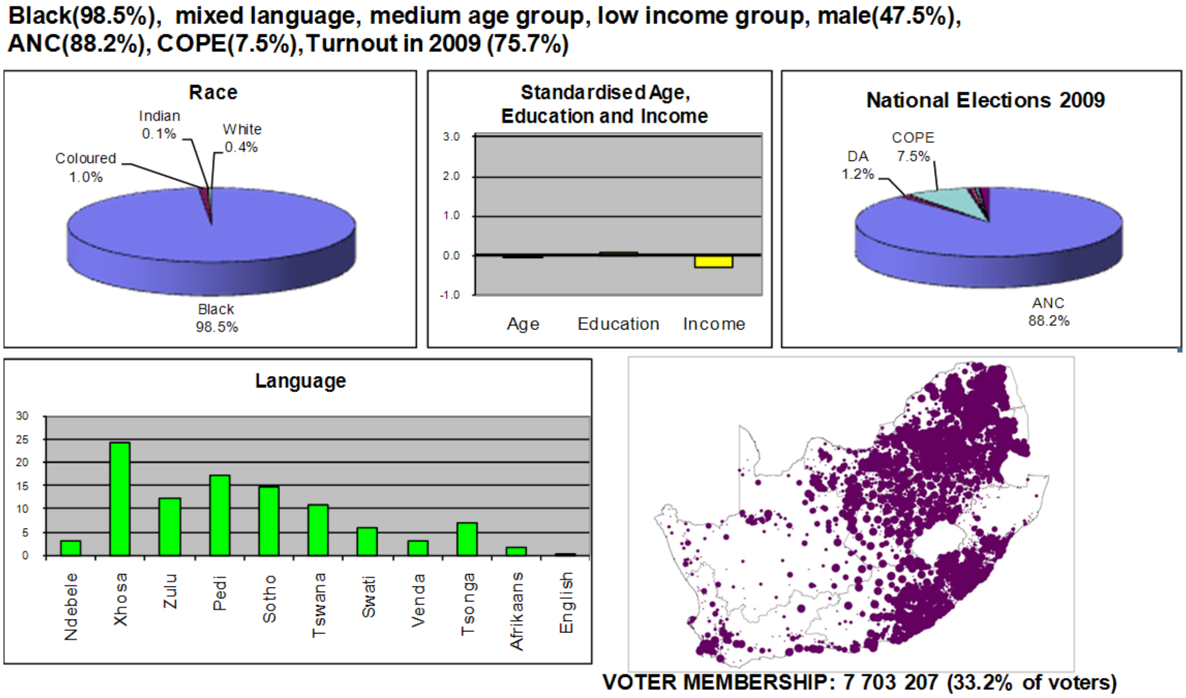


Figure 1: *Largest cluster characterized by strong ANC support. The turnout in 2011 was 54.8 percent, which is 1.7 percent below the average of 56.5 percent.*

The high percentage of Blacks in this first cluster confirms the strong support for the ANC under the black population.

The cluster shown in Fig. 2 is dominated by the white electorate. The main party represented is the DA (66.8 percent), however the ANC also has a substantial following (17.3 percent). The cluster shown in Fig. 3 is determined by a large IFP vote and has a large Zulu representation concentrated in Kwazulu-Natal. The dynamics of this group is of particular interest as there was a strong historical struggle between the ANC and the IFP in this province for many years. The ANC has made inroads in this IFP stronghold, partly through the rise to presidency of Zuma (ANC), who is of Zulu decent. In the current municipal election the IFP has been split in two parties, the original IFP and the new NFP (National Freedom Party).

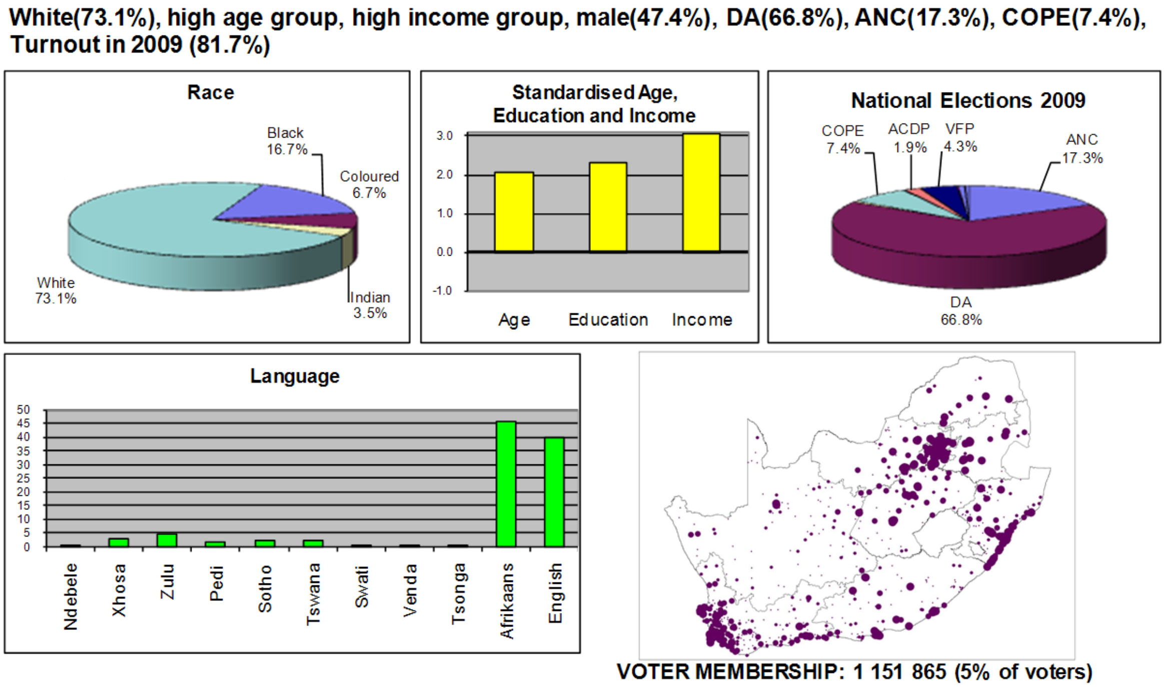


Figure 2: *4th cluster (in size of the electorate) dominated by the white population. The turnout in 2011 was 64.2 percent, which is 7.8 percent above the average of 56.5 percent.*

Finally Fig.4 shows a cluster which is representative of the colored population in the Western Cape. This province is the only province where the DA has successfully challenged the ANC, and the 2006 demise of the NP and the more recent demise of the ID (Independent party) has contributed to the increased support for the DA. This cluster was characterized in 2004 by a considerable DA following of 34.8 percent, which in 2011 has increased to 64.6 percent, partly by absorbing the NP support of 12 percent in 2004 and by absorbing the ID following, which was 16 percent in 2004. The latter party had increased to 20 percent in 2006 (within this part of the electorate), but suffered a considerable loss in 2009 (reduced to 5.8 percent). The party has now combined with the DA.

3 Trend matrices in elections

A trend matrix relates the election results of and "old" and a "new" election by counting how voters for a particular old party voted in the new election. If there are P_{old} parties in

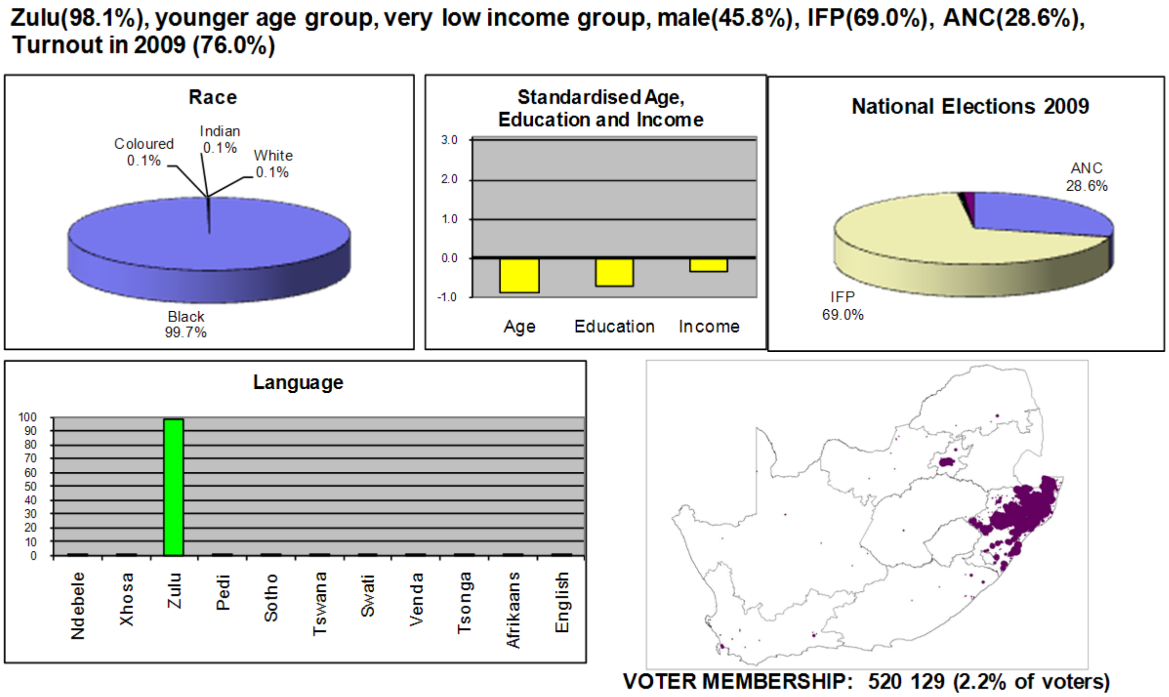


Figure 3: *10th cluster (in size of the electorate), dominated by the IFP. The turnout in 2011 was 62.6 percent, which is 6.1 percent above the national average of 56.5 percent.*

the old election and P_{new} parties in the new election then a matrix with $P_{new} \times P_{old}$ elements is necessary to characterize this behavior. To calculate such a matrix exactly, one needs all individual election results of the two elections. Clearly this information is not available, and one has to derive an approximate trend matrix from the information available on the finest level (i.e. voting districts). One possibility is to replace individual results by the average voting district results. However, this eliminates important correlations between old and new results within the individual voting districts. Instead one can derive a correlation matrix by minimizing an objective function which relates the old and new election results via a correlation matrix [1]. This leads to a "trend" matrix that has many negative elements, and therefore is also not acceptable. Formal mathematical methods, such as the Kuhn-Tucker approach, can be used to eliminate these negative elements. However, this results in a large proportion of zero matrix elements, which are implausible and are not supported by studies using questionnaires of individual voters [1]. In Ref. [1] a heuristic "positivization" method is introduced which yields more acceptable results and can be implemented fairly easily. However, it is still a heuristic, whose validity is hard to verify. In this paper we use a new simulation method to mimic each individual voter, subject to the knowledge of the individual voting district results and guided by a global trend matrix (which could initially be one of the options discussed above). After constructing a new global trend matrix from the simulated results, one can redo the simulation with the new global matrix, until convergence is reached. Since the matrix constructed in this way is insensitive to the choice for the initial global matrix, its results look superior to the heuristic and formal methods. Hence, we will present results using this method, leaving a detailed discussion of this method to a future publication.

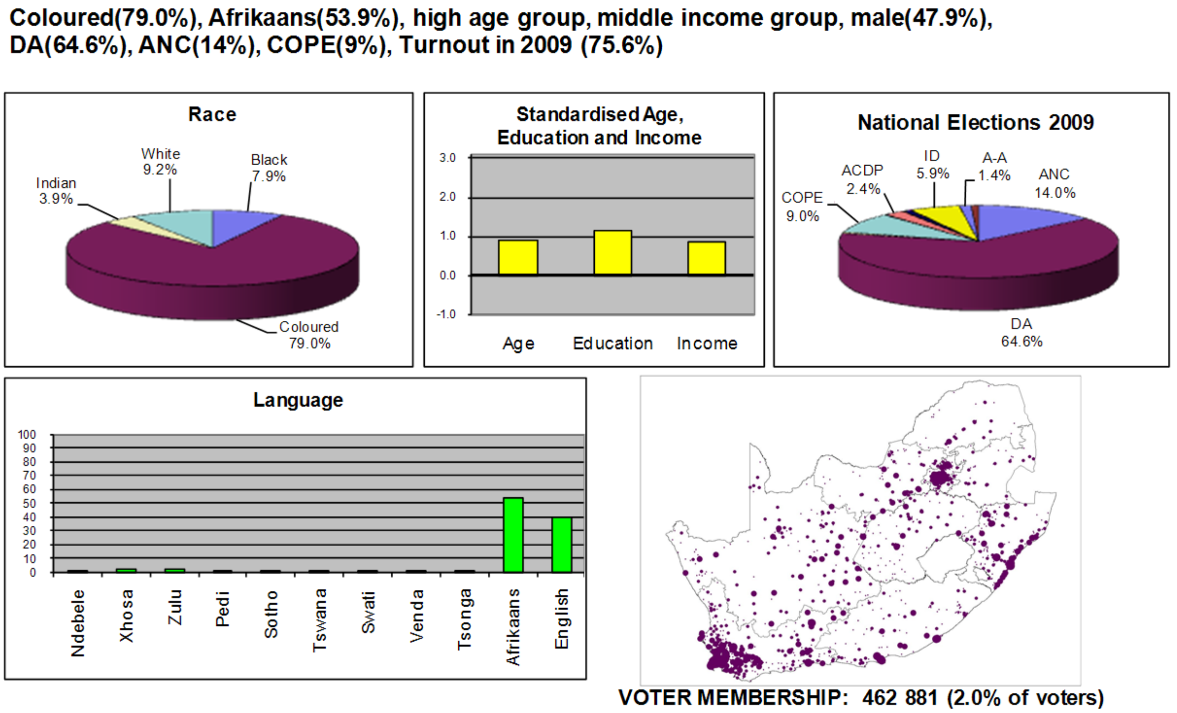


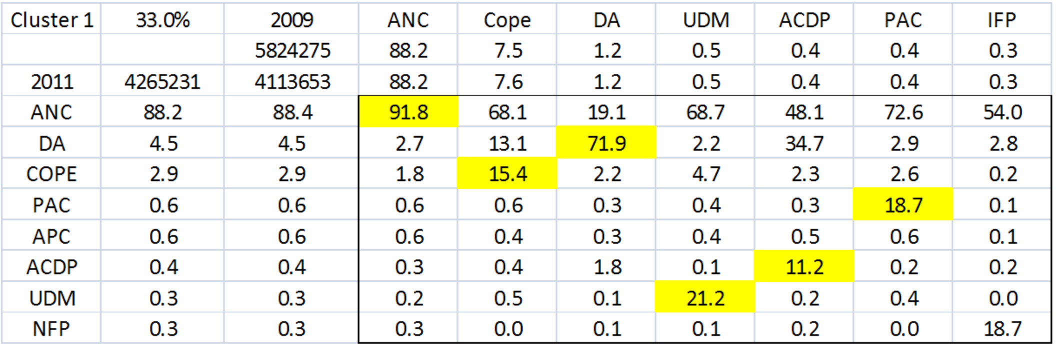
Figure 4: *15th cluster (in size of the electorate) dominated by the DA and a 9 percent COPE vote in the 2009 elections. The turnout in 2011 was 62.6 percent, which is 6.1 percent above the national average of 56.5 percent.*

Since the trend matrix only captures the electorate common to both elections it ignores new voters entering the system and old voters leaving the system. This component is captured by the difference between the (common) trend results and the actual result in the columns (new results) and rows (old results) in the tables shown in the next section. We explain the first trend matrix shown in detail so that the reader gets further insight in the usefulness of this new tool in election analysis.

4 Trend results for the four sectors of the electorate

To explain the use of the trend matrix, the trend matrix for cluster 1 is displayed in Table 1. The nature of this segment of the electorate was already explained through Fig.1.

Table 1: *Relative trend matrix for cluster 1, which covers 33% of the electorate. Diagonal entries are highlighted.*



Cluster 1	33.0%	2009	ANC	Cope	DA	UDM	ACDP	PAC	IFP
		5824275	88.2	7.5	1.2	0.5	0.4	0.4	0.3
2011	4265231	4113653	88.2	7.6	1.2	0.5	0.4	0.4	0.3
ANC	88.2	88.4	91.8	68.1	19.1	68.7	48.1	72.6	54.0
DA	4.5	4.5	2.7	13.1	71.9	2.2	34.7	2.9	2.8
COPE	2.9	2.9	1.8	15.4	2.2	4.7	2.3	2.6	0.2
PAC	0.6	0.6	0.6	0.6	0.3	0.4	0.3	18.7	0.1
APC	0.6	0.6	0.6	0.4	0.3	0.4	0.5	0.6	0.1
ACDP	0.4	0.4	0.3	0.4	1.8	0.1	11.2	0.2	0.2
UDM	0.3	0.3	0.2	0.5	0.1	21.2	0.2	0.4	0.0
NFP	0.3	0.3	0.3	0.0	0.1	0.1	0.2	0.0	18.7

On the top row are the main parties for cluster 1 in 2009, while on the leftmost column are the main parties in 2011. The cluster results for these parties are shown below/next to the party names in the respective years. These numbers add up to nearly 100 percent, the defect being due to smaller parties which have not been shown. The trend matrix is shown in the marked box. The number 91.8 on the top left means that of the ANC voters in 2009, 91.8 percent voted again for the ANC in 2011 (the party shown in the relevant row), while 2.7 percent of those voters voted for the DA in 2011. Of the 2009 COPE voters 15.4 percent voted again for COPE. By multiplying the old (2009) election results in Table 1 by the percentages in one row corresponding to a specific new party, and adding them up over all columns, one obtains the 2011 trend result shown in one of the left-hand columns. This result is different from the full (actual) election result, as the trend result is based on voters common to the old and new election. The difference between actual and trend results which is due to added or removed voters is virtually negligible in the current case. However, in forthcoming cases considerable shifts due to added or removed voters can be observed. In addition to trend matrices between 2009 and 2011 (such as shown in Table 1), trend matrices between 2006 and 2011 and between 2004 and 2009 will be used. Although results are quoted from the relevant trend matrices in the following sections, only one further trend matrix is shown for space limitations.

4.1 Cluster 1 dominated by the ANC

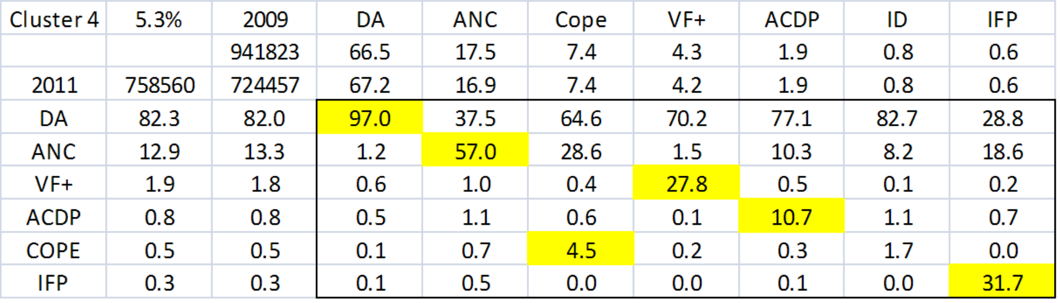
The loyalty of the ANC electorate for cluster 1 is very high (≈ 92 percent) whether we consider the transition from 2006 or 2009 to 2011. Compared to 2006 the remainder went in equal numbers to COPE (a breakaway from the ANC prior to the 2009 election) and the DA (namely 2.6 percent). However, compared to 2009 only 1.8 percent of the 2009 ANC voters voted for COPE, while 68 percent of the 2009 COPE voters went back to the ANC in 2011, so that COPE was reduced from 7.5 percent to 2.9 percent. For the whole country the defection of COPE voters back to the ANC was less, namely 55 percent. Despite this return of COPE voters, the ANC did not register a net percentage gain for cluster 1, as this increase cancelled by ANC voters that went to the DA (2.7 percent). Overall, the behavior of the electorate in this cluster shows that the loyalty of the black population in black areas remains high and that the defection to the DA is not (yet) significant. Despite many service protests against the sitting ANC dominated councils, which would relate to this cluster, the loyalty of ANC voters has not suffered. Also the percentage of spoilt votes (a possible vehicle for voicing protest) in this cluster (1.15 percent) is not significantly higher than the national average (0.99 percent). However, the turnout (54.8 percent) is 3 percent lower in relative terms than the national average of 56.5 percent, so this may indicate a protest vote. For the second cluster, which is also dominated by the ANC (84.3 percent), the turnout is even down to 52.6 percent, which is 7 percent lower than the norm. An in depth discussion of the effect of turnout in the 2009 election is given in [4].

4.2 Cluster 4 dominated by the DA

This cluster is dominated by the white population (73 percent) and its characteristics were displayed in Fig.2. It will be analyzed together with the 5th cluster (not shown), which is also dominated by the white population (83 percent) and features similar behavior. The main difference between these two clusters is that the 2009 ANC vote in cluster 5 is less (8 percent) than that in cluster 4 (17 percent). The turnout in 2009 for cluster 4 (5) was 82 (83) percent, while the 2011 municipal elections featured turnouts of 64 (69) percent. While the municipal turnout in 2011 is considerably less than the 2009 turnout, these turnouts for cluster 4 and 5 are still substantially larger than the average over the whole electorate (56 percent). In fact, the reduction between national and municipal election turnout is considerably less for clusters 4 and 5 than for the main cluster 1, confirming the tendency that the ANC electorate voters abstain more easily in the municipal election(s). In Table 2 the relative trend matrix is shown for this cluster between 2009 and 2011. Showing this comparison is more insightful than the transition from 2006 and 2001, as many developments took place after 2006 (in particular the creation of COPE).

Within this cluster 37 percent of the ANC voters crossed over to the DA, while 65 percent of the COPE voters crossed over to the DA. In addition to this clear trend the DA profited from added new voters and leaving old voters by adding 1 percent in total. This increase seems to go fully at the expense of the ANC. Cluster 5 shows a similar picture as cluster 4, with the ANC even less loyal (50 percent instead of 57 percent in the upper case). Clearly, there is a reluctance of COPE voters in DA dominated areas to vote for the ANC (only 29 percent of its original support voted for the ANC). In ANC dominated areas (e.g. cluster 1) the situation is exactly opposite. For that cluster 68 percent of COPE voters went back to the ANC, while

Table 2: *Relative trend matrix for cluster 4, which covers 5.3% of the electorate Diagonal entries are highlighted.*



Cluster 4	5.3%	2009	DA	ANC	Cope	VF+	ACDP	ID	IFP
		941823	66.5	17.5	7.4	4.3	1.9	0.8	0.6
2011	758560	724457	67.2	16.9	7.4	4.2	1.9	0.8	0.6
DA	82.3	82.0	97.0	37.5	64.6	70.2	77.1	82.7	28.8
ANC	12.9	13.3	1.2	57.0	28.6	1.5	10.3	8.2	18.6
VF+	1.9	1.8	0.6	1.0	0.4	27.8	0.5	0.1	0.2
ACDP	0.8	0.8	0.5	1.1	0.6	0.1	10.7	1.1	0.7
COPE	0.5	0.5	0.1	0.7	4.5	0.2	0.3	1.7	0.0
IFP	0.3	0.3	0.1	0.5	0.0	0.0	0.1	0.0	31.7

only 13 percent made the transition to the DA. Further insights in the voting behavior can be obtained from studying the matrix in Table 2 in detail.

4.3 Cluster 10 dominated by the IFP

This cluster was dominated by the IFP in 2004, but has since undergone some significant changes. In 2004 and 2006 the IFP support was 87.4 percent and very stable. The only competing party was the ANC with 8.2 percent in 2004 and 9.4 percent in 2006. Then in 2009 the IFP lost considerable support and was reduced to 68.7 percent, nearly the whole loss due to the ANC which increased to 28.9 percent. For this cluster the election in 2011 was also dramatic as the NFP split off from the IFP. As a result the IFP lost further support (reducing to 49 percent) while the NFP scored 24 percent, attaining half the strength of the IFP. The increase of the ANC came to a stop in this sector as they only scored 25 percent in 2011. Other parties hardly feature in this sector. In particular, the DA has not managed to obtain any significant amount of the disgruntled IFP voters, although it has doubled in size from 0.4 percent to 0.9 percent in 2011.

4.4 Cluster 15 dominated by the colored voters in the Western Cape

In subsequent election years 2004, 2006, 2009 and 2011 the DA support in this cluster has steadily increased from 33.5, 46.3, 64.5 to 77.8 percent in 2011. The increase from the 2004 to the 2009 national elections was mainly due to the demise of the NNP (76 percent of the NNP votes went to the DA), and the reduced support for the ID, which decreased from 16.0 to 5.8 percent in this period. The trend matrix between 2009 and 2011 shows that the support for smaller parties like the ACDP and the VF+ also has diminished, again benefitting the DA in the Western Cape. At the same time the support for the ANC has eroded over the years, going from 26.9 percent in 2004 to 22.1 percent in 2006, 14.1 percent in 2009, and 13.5 percent in 2011.

5 Conclusions and summary

The new tool of trend matrices gives detailed insight in the movement of voters between parties. This is illustrated by comparing the 4 last elections in South Africa. The turnout in the municipal elections is on average only 74 percent of that of the national elections, although this reduction is more prominent for ANC dominated clusters (72 percent for cluster 1) than for DA dominated clusters (83 percent for cluster 5). Since the provision of services plays an important role in municipal elections one could interpret the refusal to vote as a protest tool of the electorate in ANC dominated regions. Nonetheless the loyalty of ANC voters in these regions is still very high (92 percent in cluster 1 between 2009 and 2011). However, in this same cluster the small DA support has increased threefold, from 1.2 percent in 2009 to 4.5 percent in 2011, so that the traditional voting patterns along racial lines may show some cracks.

Bibliography

[1] Greben, J. (2007). A theory of quantitative trend analysis and its application to south african general elections. *South African Journal of Science*, 103:232–238.

[2] Greben, J., Elphinstone, C., and Holloway, J. (2006). A model for election night forecasting applied to the 2004 South African elections. *ORiON*, 22:89–103.

[3] Greben, J., Elphinstone, C., Holloway, J., De Villiers, R., Ittmann, H., and Schmitz, P. (2004). National elections in South Africa. *South African Journal of Science*, 101:157–161.

[4] Kimmie, Z., Greben, J., and Booysen, S. (2010). The effect of changes in registration and turnout on the results of the 2009 South African election. *Politeia*, 29:98–120.