Some Recent Developments in Statistical Theory and Applications

Some Recent Developments in Statistical Theory and Applications

Selected Proceedings of the International Conference on Recent Developments in Statistics, Econometrics and Forecasting, University of Allahabad, India, December 27-28, 2010

Edited by

Kuldeep Kumar & Anoop Chaturvedi

BrownWalker Press
Boca Raton

Some Recent Developments in Statistical Theory and Applications:
Selected Proceedings of the International Conference on Recent Developments in Statistics, Econometrics and Forecasting, University of Allahabad, India, December 27-28, 2010

BrownWalker Press
Boca Raton, Florida
USA • 2012

ISBN-10: 1-61233-573-X *(paper)*
ISBN-13: 978-1-61233-573-5 *(paper)*

ISBN-10: 1-61233-574-8 *(ebook)*
ISBN-13: 978-1-61233-574-2 *(ebook)*

www.brownwalker.com

CONTENTS

PREFACE

This book is part of the proceedings of The International Conference on Recent Developments in Statistics, Econometrics and Forecasting. Statistician is a noble profession. According to AMSTAT News (News letter of American Statistical Association), "Statistician is rated as third Best Job in the nation". According to Hal Varian, Chief Economist at Google "The sexy job of the next 10 years will be statistician. The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that is going to be hugely important skill in the next decades". As H G Wells said "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

There is a common belief amongst the students and researchers from other disciplines that subjects like Mathematics and Statistics do not change over time. However, this is a completely wrong view as the subjects like Statistics are changing very fast. Multivariate Analysis has been replaced by "Modern Multivariate Analysis", Econometrics by "Modern Econometrics" and Time Series Analysis by "Modern Time Series Analysis". For example, Box-Jenkins models developed in the early seventies have been replaced by more sophisticated ARCH, GARCH etc. models.

The goal of the The International Conference on Recent Developments in Statistics, Econometrics and Forecasting which was held in Alahabad, India on 27th and 28th December, 2010 was to provide opportunities for academics and researchers to share their knowledge on recent developments in this area. The conference featured the most up-to-date research results and applications in statistics, econometrics and forecasting. This book is part of the proceedings of this conference in which we have selected fifteen papers.

The book has fifteen chapters contributed by different authors and can be divided into five parts; Time Series and Econometric Modelling, Linear Models, Non-parametrics, Statistical Applications and Statistical Methodology. The first four chapters deal with Econometrics and Time Series modelling. In the first chapter Gepp and Kumar have reviewed various statistical techniques which are used in bankruptcy prediction including some of the cutting edge techniques such as Decision Trees. The second chapter by Sen discusses covariance between stochastic processes with an application to online auctions. The Third chapter by Rajaguru and Abeysinghe provides a quantitative assessment of the magnitude of the distortions created by temporal aggregation on Granger Causality. The fourth chapter by Mishra and Kumar deal with modelling of Indian growth series using ARMA models. The next three papers deal with linear models. In the fifth chapter Shalabh and Heumann have considered the simultaneous prediction of average and actual values of study variables in a linear regression model using Stein rule estimators. In the sixth chapter Lalitha and Kumar have discussed a procedure for detection and testing of multiple outliers in a linear model. In the next chapter Shrivastava et. al considered a dynamic model involving structural change, which may occur either due to shift in disturbances precision or due to shift in regression parameters using Bayesian analysis.

The next two chapters deal with non-parametric statistics. In chapter nine Steele et. al compared the power of the discrete Kolmogorov-Smirnov and Chi-Square goodness-of-fit tests. In the next chapter Mathur and Sadana developed a Rank Based Test for Detecting a Shift in Location and demonstrated its application in genomics. The next two chapters deal with statistical applications. In chapter ten Katyan and Singh have discussed Gender Differences In Health Care Expenditures Using Multilevel Modeling and in the next chapter Tiwari has developed a Statistical Model Predicting the Prevalence of Tuberculosis in Almora District of Uttarakhand, India in Relation to Age and Sex. The last four chapters deal with statistical methodology. In Chapter 12 Tomer et al. did an Analysis of Incomplete Data from Rayleigh Distribution under Competing Risk Model. In chapter thirteen Bhushan and Pandey discussed a successive sampling strategy for estimation of population mean using prior information about the correlation coefficient and in the next chapter the same authors have considered a Bayesian estimation of the modified Weibull distribution. Finally in the last chapter Chaturvedi and Srivastava have derived finite sample properties of Taguchi's Process Capability Index.

All the chapters in this book, which are part of the conference proceedings, were refereed. A "blind" paper evaluation method was used. This book is a good mixture of theory and practical applications. It may not be a suitable book for the beginners but definitely graduate and research students will enjoy it. Also the practitioners will find this book quite useful. This book will be helpful to graduate students, researchers and applied statisticians working in the area of time series, statistical and econometric modelling.

Finally we will like to thank all our speakers, sponsors, review committee members, organising committee members and especially all the conference participants for making this conference a success.

Kuldeep Kumar
Bond University
Gold Coast, Australia

Anoop Chaturvedi
Allahabad University
Allahabad, India

Business Failure Prediction Using Statistical Techniques: A Review

Adrian Gepp and Kuldeep Kumar[1]

Bond University, Gold Coast, Australia

Corresponding author's e-mail address[1]: kkumar@bond.edu.au

Abstract

Accurate business failure prediction models would be extremely valuable to many industry sectors, particularly in financial investment and lending. The potential value of such models has been recently emphasised by the extremely costly failure of high profile businesses in both Australia and overseas, such as HIH (Australia) and Enron (USA). Consequently, there has been a significant increase in interest in business failure prediction from both industry and academia.

Statistical business failure prediction models attempt to predict the failure or success of a business. Discriminant and logit analyses are the most popular approaches, and there are also a large number of alternatives. In this paper, the various techniques used in previous studies are presented and reviewed, including two alternative techniques that have produced promising results, namely survival analysis and decision trees.

Key words: bankruptcy prediction, Decision tree, survival analysis, Discriminant analysis

1. Introduction

The field of business failure prediction has many aliases, such as bankruptcy prediction, firm failure prediction and financial (di)stress prediction. Hereafter it will be referred to as business failure prediction (BFP). As the name suggests, BFP involves developing models that attempt to predict the financial failure of a business before it actually happens. Accurate BFP models would be extremely useful and valuable in the real world, as recently emphasised by the extremely costly failure of high profile businesses in both Australia (HIH and OneTel) and overseas (particularly Enron in the United States). Consequently, there has been a significant increase in interest in BFP, from both industry and academia.

Statistical BFP models attempt to predict the failure or success of a business based on publicly available information about that business, such as financial ratios from financial statements. In addition, some studies also include indicators of industry and economy wide performance to aid in the business failure predictions. Some benefits from accurate business failure predictions are:

- Banks, investment banks, credit unions, and other financial institutions could avoid lending to businesses that will fail, and thus never repay their loans.
- The financial investment sector could improve the risk return trade-off from investments by not investing in failing businesses.

- Businesses could establish long-term relationships with other businesses (such as suppliers) that will not fail in the future, and thus increase the longevity and viability of their business relationships.
- Regulatory bodies could make early identifications of failing businesses. This early identification assists regulatory bodies in ensuring that business failure is 'handled' legally and illegal activities, such as avoiding taxes or diluting debt holders' claims by issuing substantial common stock dividends prior to failure, are avoided.

Any individual or organisation dealing with businesses could profit from using accurate BFP models in order to plan to deal with solely successful businesses. Moreover, it is possible for businesses to use a BFP model to predict their own failure or survival, and use this information as a measure of financial health for management decision making. Overall, accurate BFP models would increase people's confidence in investment, lending and the development of profitable business relationships. Thus, lending and investment (in successful businesses) would increase, which would result in increased stable economic growth for the benefit of all involved.

There are three general types of research in the field of BFP (adapted from Laitinen and Kankaanpää (1999)); these are:

1. Theoretical modelling of the business failure process. Only a few studies have focused on this area, particularly Wilcox (1971) and Santomero and Vinso (1977).
2. Searching for the set of explanatory variables (usually accounting and financial ratios) that best explain and predict business failure. Research focusing on this area is common in the accounting literature.
3. Searching for the 'best' empirical method for BFP. Research into this area often studies the predictive power of different explanatory variables as a by-product.

Recently, the 3rd area of research has been the main contributor to the field of BFP. This has been further enhanced by the continuing improvements in computer technology that have enabled extraordinary advances in data analysis. The ramifications of this are that theorists, who develop mathematical and statistical models, no longer have to be concerned about the feasibility of being able to apply their model. Thus, the more computationally intensive models are becoming practically viable options for BFP, and hence the popularity of the 3rd area of research.

The remainder of this paper is structured as follows. A review of BFP models is followed by an analysis and review of decision trees, and then survival analysis, for BFP. Potential research in decision future research is then proposed and a conclusion is given.

2. Different Business Failure Prediction Models

Many different techniques have been applied to BFP since its beginnings in the 1960's. The field arguably started earlier, but the first statistical and mathematical models for BFP were published. The various techniques used since then are reviewed and referenced in this paper. Where the terms Type I and Type II Error are used, Type I Error refers to misclassifying a failing business as successful and conversely Type II Error refers to misclassifying a successful business as a failure.

Univariate Analysis

Early attempts to use financial ratios to predict business failure stem from the work of Patrick (1932). This work was later extended by Beaver (1966), who presented the first modern statistical model for BFP. Beaver developed a univariate model, and used a set of 30 financial ratios that were chosen with regard to the following three criteria:

- Prevalence in previous literature,
- Performance in previous studies, and
- Consistence with Beaver's logical (not theoretical) cash flow concept that placed a preference on ratios derived from cash flow statements. The logic behind this is that bankruptcy and insolvency is related to a lack of liquid assets (particularly cash) relative to liabilities due, rather than to accrued accounting assets and liabilities.

Beaver tested his model on 79 failed businesses and 79 similar successful businesses between 1954 and 1964. A cut-off point was identified for each financial ratio, which separated the predicted successful businesses from the predicted failure businesses. From the univariate analysis of each chosen ratio on the data set, a ratio set that consisted of ratios with the greatest predictive power was established. This set comprised (in declining order of predictive power):

1. Cash flow to total debt,
2. Net income to total assets,
3. Total debt to total assets,
4. Working capital to total assets,
5. Current ratio = current assets / current liabilities, and
6. No credit interval = (immediate (or quick) assets – current liabilities) / (operating costs – depreciation)

It is interesting to note that Beaver considered business failure to be much broader than just liquidation: a business was considered to have failed financially if any of the following were satisfied:

- the business had gone into liquidation,
- there was an overdrawn bank account,
- there had been default on debt (such as a bond), or
- preferred stock dividends had been missed.

Beaver's univariate approach did not contain an overall measure of financial distress, which led to the problem that different ratios made conflicting predictions about a given business. In addition, it was noted that one ratio alone could not encompass the complexity of business failure. This model's error was estimated at 22% Type I Error and 5% Type II Error for one year prediction intervals. In addition, although the Type II Error remained constant for longer predictions, Type I Error increased as the length of the predictions increased. Hence, Beaver's model performs better for shorter predictions. Overall, Beaver's work was pioneering in the BFP field, and sparked the development of using statistical models to predict the failure of a business.

Discriminant Analysis

Altman (1968) pioneered the use of discriminant analysis (DA), which was the first multivariate approach applied to BFP. This extended the work of Beaver (1966) by addressing the problem that various ratios made conflicting predictions, and incorporates the concept of a composite measure of business distress. Altman (1968) presented a DA model to predict business failure, in which the information from several variables (ratios) was combined into a single weighted score for each business. This score was calculated based on the following general discriminant function:

$$Z = a1x1 + a2x2 + ...+ anxn + c$$

where Z was the score, xi were the independent variables, and ai and c were the estimated parameters. The separating function for Altman's (1968) model was:

$$Z = 0.012x1 + 0.0141x2 + 0.033x3 + 0.006x4 + 0.999x5$$

where

x1 = working capital over total assets,

x2 = retained earnings over total assets,

x3 = earnings before interest and tax over total assets,

x4 = market value of equity over book value of total debt, and

x5 = sales over total assets.

It is interesting to note that no cash flow ratios were found to be significant, which contrasts the emphasis placed on cash flow ratios by Beaver (1966).

Cut-off scores are then generated based on sample results and used to classify each observation. Altman's MDA used 2 cut-off scores (1.8 and 2.7) to classify businesses into three categories as shown in the table below.

Z-score lookup	**Prediction**
$Z > 2.7$	Success
$Z < 1.8$	Failure
$1.8 \leq Z \geq 2.7$	Inconclusive

Although the probability of failure or success is not an explicit output of this model, a relative measure of probability can be obtained by calculating the difference between Z-scores and cut-off values. For example, a business with a Z-score of 1.0 is more likely to fail than a business with a Z-score of 1.7.

Altman's DA model outperformed Beaver's univariate model for one year prediction intervals; however, Altman's model was not as accurate for longer predictions. It is also interesting to note that Martin (1977) concluded that using DA with more than two classification groups (multiple discriminant analysis) was more accurate that only using two classification groups (linear discriminant analysis).

The use of DA has been further developed since 1968; principally, Deakin (1972) increased the number of independent variables to use the 14 variables used by Beaver (1966), whilst Edminster (1972) applied a model similar to Altman's to small businesses. However, standard DA has two major statistical assumptions that are often violated (Laitinen and Kankaanpää, 1999); they are that

1. The independent variables are multivariate normal, and that
2. The covariance of the two classification groups should be equal.

For example, the first requirement is not met if an independent dummy variable is introduced; however, techniques such as log or square root transformations and elimination of outliers can be used to aid consistency with the first assumption (Laitinen and Kankaanpää, 1999). Richardson and Davidson (1983) found that if the first assumption were violated the model was sensitive to the data used in the estimation of model parameters; however, the practical significance of the difference may be minor. Moreover, to overcome the second requirement quadratic discriminant analysis can be used, which involves squared independent variable terms in the discriminant function stated above. However, Altman et al. (1977) and Hamer (1983) found that a quadratic discriminant function did not consistently improve practical accuracy even though it improved statistical validity. Furthermore, all standard discriminant analysis models assume equal misclassification costs of Type I and II Error, which is usually violated.

Other Standard Multivariate Techniques

Kumar and Ganesalingam (2001) focused on predicting financial distress among a selection of major Australian companies. Seventy-one such companies were subjected to an analysis to determine various facts about the companies, in particular their long-term stability. Among these facts are the likelihood of each company becoming bankrupt and the classification of companies into distinct groups (clusters) based on ten financial ratios. This classification into various groups can specifically help investors maximise their risk-return trade-off by selecting companies from different clusters to maximise their portfolio's diversity. The multivariate techniques employed in this research were principal component analysis, factor analysis, discriminant analysis (DA) and cluster analysis. The three techniques other than DA, which was discussed above, have been briefly described below. Standard multivariate analysis textbooks such as Flury (1997) have more details on all of the above techniques.

- *Principal Component Analysis* involves attempting to express a system with p components (variables) in a linear combination of k principal components, where k<p and the k components are representative of the system. That is, the goal is to explain the variance-covariance matrix with a linear combination of variables that is less that the number of original variables. It is important to note that every principal component was previously an original component of the system. This analysis often reveals extra relationships between components that lead to new interpretations.

- *Factor Analysis* also attempts to express a system in a linear combination of a number of variables fewer than in the original system. However, with factor analysis the variance-covariance matrix is estimated by some underlying, but unobservable, random quantities termed factors. That is, factors are not components of the original system; for example, the intelligence of top management may be an underlying factor in the financial success of a range of businesses in different industries.

- *Cluster Analysis* is an exploratory technique similar to DA. The difference between the techniques is that cluster analysis makes no assumptions about the number of groups or group structure; that is, groups are created based on natural groupings in the data set being analysed. A 'human eye' cluster analysis is often used to obtain a useful, but not optimal, grouping. Overall, cluster analysis is more primitive than DA, but it is still a useful method to assess dimensionality, outliers, simplify data, and suggest hypotheses about relationships amongst data.

Logit and Probit Analysis

Logit Analysis (LA) generates a score for each business similar to DA. However, it is free from the normality and equal covariance assumptions of DA. LA is based upon the cumulative logistic function (CLF), and due to its non-linear nature the coefficients are usually estimated using the maximum likelihood method (Kumar and Tan, 2005). Furthermore, unlike the difficult interpretation of the Z-score in the DA model, the score in LA can be directly interpreted as the probability of failure (where the CLF ranges from 0 to 1). The general relationship between the two models is demonstrated in Figure 1, where Z was defined in the previous DA section.

$$\ln\left(\frac{P(Z)}{1-P(Z)}\right) = Z \quad or\ rearranged\ as$$

$$P(Z) = \frac{1}{1+e^{-Z}}$$

Figure 1: LA Model where P(Z) is the probability of failure.

Once each business has an associated probability of failure, n cut-off (or critical) values can be established to separate the businesses into n + 1 groups similar to DA. However, there is usually one cut-off value that separates businesses into failure and success groups. As in DA, the cut-off values in LA can be changed to cater for different misclassification costs: as Type I Error is more serious, the chosen critical value is often lower than 0.5. Furthermore, inherent in the CLF is that businesses with probabilities closer to 0.5 are more sensitive to changes in the independent variables, as shown by the sleeper slope of the CLF close to values of 0.5 in Figure 2. This behaviour is logical as short-term changes will have a smaller effect on extremely distressed or successful businesses compared with borderline businesses.

Ohlson (1980) is the seminal study for applying LA to BFP. He produced three separate LA models to predict failure for one, two and three years in advance. Fourteen ratios were used as predictors, consisting of standard accounting ratios, dummy variables based comparisons of balance sheet figures, and a variable representing the change in net income over the last year (which comprises time-lagged information). The empirical results for his model were disappointing, but he showed that LA is more statistically valid and easier to interpret than DA. In addition, subsequent studies on LA have shown that it is usually slightly empirically superior to DA in both classification and prediction accuracy (Laitinen and Kankaanpää, 1999). However, Martin (1977), Collins and Green (1982) and Hamer (1983) all stated that the overall classification accuracy of DA and LA is not significantly different. Collins and Green went on to further state that Type I Errors appear to be lower in LA, but the additional computational power required for LA is only justified if Type I Errors are very large. This condition may well be justified as, since 1977 when the statement was made, there have been significant increases in computational power and the high cost of Type I Errors has been highlighted by the failure of large businesses (such as HIH and Enron).

Overall, it is generally agreed that LA is at least as accurate as DA. Therefore, due to its improved statistical validity and straightforward interpretation, LA has become more popular than DA.

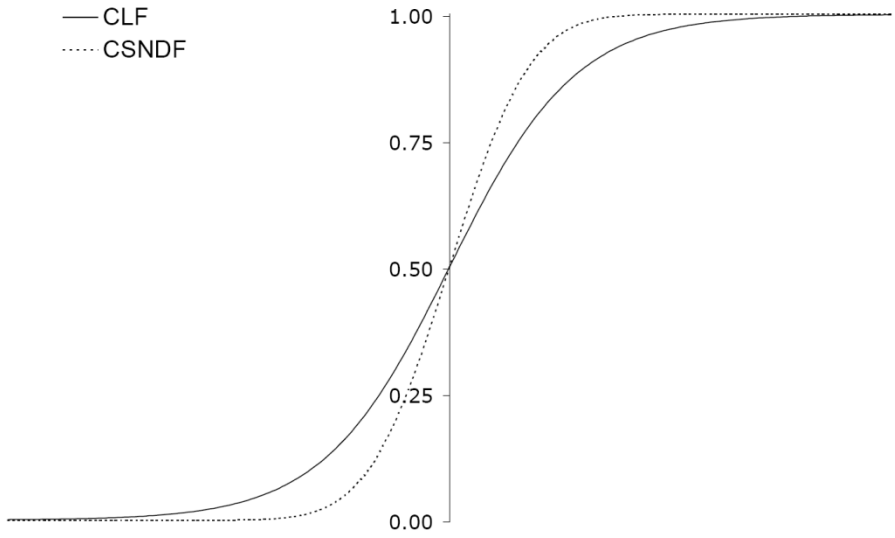

Figure 2: A comparison of the CLF and CSNDF, used in logistic and probit analysis respectively.

In addition to LA, probit analysis (PA) has also been applied to BFP. The only difference is that PA uses the cumulative standard normal distribution function (CSNDF) instead of the CLF, whereby the CSNDF is defined in Figure 3. Although more complicated, this function has a very similar shape to that of the CLF, as illustrated in Figure 2. This fundamental similarity means that both LA and PA usually produce the same conclusions for the same data. However, PA is more computationally intensive than LA due to the nonlinear estimation needed (Gloubos and Grammatikos, 1988); therefore, unlike LA, PA is not prominent within the BFP literature.

$$P(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{\frac{-Z^2}{2}} dZ$$

Figure 3: PA Model where P(Z) is the probability of failure

There is also an interesting paper by Pacey and Pham (1990), who state the following major problems with DA and LA:

- Arbitrary estimation of cut-off values,
- Assumption of equal misclassification costs in model estimation stage, and
- Bias in the selection of samples for model estimation.

Pacey and Pham even suggest that a naive model that assumes all businesses to be bankrupt may outperform both DA and LA. These points are valid, but realistically the cut-off values can be estimated mathematically for given misclassification costs. Therefore, the first problem is not related to the model, but rather caused by a lack of agreement on the actual misclassification costs.

In addition, the majority of studies have shown their DA and LA models to be as good as or better (although often not significantly better) than this naive model.

Human Information Processing

Human Information Processing (HIP) involves utilising the existing ability of human decision makers to use information (usually accounting information) to group companies according to their probability of future bankruptcy. Hence, HIP attempts to model the relationship between cues and decisions, rather than the information processing used to form the decisions. Consequently, interviewing and questionnaires are frequently used research techniques.

The pioneering studies in this area are by Libby (1975) and Casey (1980). Abdel-Khalik and El-Sheshai (1980) presented results of a quantitative comparison that found mathematical and statistical (quantitative) models to be superior decision makers compared with experienced professionals. Moreover, the HIP model produces decisions without a level of certainty. These are the main reasons, together with the recent dramatic increase in computational power, that the emphasis has shifted mainly to developing quantitative models rather than using HIP models. Nevertheless, HIP generates some interesting results that can be used to improve mathematical and statistical models. For example, conclusions such as (1) a greater amount of accounting information does not improve predictive power (Casey, 1980) and (2) the choice (rather than the processing) of information is most important (Abdel-Khalik and El-Sheshai, 1980), are both useful background knowledge when developing mathematical and statistical models.

Artificial Neural Networks

Although somewhat related to HIP, Artificial Neural Networks (ANNs) are far more powerful. ANNs were developed as a soft computing technique to model the inner-workings of the human brain as a type of artificial intelligence. The first major step towards developing a computer information processing system based on the human brain was made by McCulloch and Pitts in 1943. These models have since been applied to BFP and become somewhat popular in the field, although not as popular as DA and LA.

The topology of an ANN (in BFP) involves connected layers of neurons:

- One layer of input neurons (usually financial ratios),
- One or more hidden layers of interconnected neurons, and
- One output layer of neurons (usually just one boolean fail or success neuron).

Each connection between two neurons has an associated weight that indirectly relates to the likelihood that it will be used in the decision. Prior to its use, the ANN is 'trained' like a human brain by processing data with known results. A common ANN with back-propagation is then 'rewarded' for a correct classification by increasing the weight of the neuron connections that led to the correct classification. Similarly, the weights of neuron connections that lead to an incorrect classification are reduced. Eventually, a suitable weighted interconnection of neurons is established and can be used for making predictions. This is known as supervised learning, but there are also different training techniques such as an unsupervised or graded learning. Laitinen and Kankaanpää (1999) provide a brief discussion of these training methods.

The training of an ANN plays a large factor in its success. Training an ANN is a difficult and complex process that involves many decisions and refinements: the training method must be cho-

sen along with many other parameter values such as the learning rate, momentum and input noise. Optimising the various parameters of an ANN is an art rather than a science. Furthermore, the process of optimising the optimal number of hidden layers and neurons in those layers is also an art. The 'trial and error' style of process associated with training an optimal ANN for prediction causes the model quality to vary substantially, depending upon the individual undertaking the training process. Thus, an ANN approach is more time consuming than alternative approaches and is usually only used by individuals familiar with them, which has been a barrier to increasing the popularity of ANNs.

The major advantage of ANNs is that they do not have the same restrictive assumptions as traditional statistical methods, such as normality, linearity and independence among input variables. Elliot and Kennedy (1988) provide a comprehensive review of the assumptions made in other conventional techniques that are not made with ANNs. Due to their flexibility, ANNs can also deal with outliers, missing data and multicollinearity better than traditional techniques. Despite their statistical validity, ANNs have produced mixed results. Laitinen and Kankaanpää (1999) claim that ANNs have not produced significantly superior results compared with other well-known techniques, and show that in their study DA and LA were comparable with ANNs. However, Tan (2001) points out that there have been published applications of ANN in BFP that have outperformed DA and LA. The vast majority of the ANN studies with a positive result used forward feeding back-propagation neural networks, such as Odom and Sharda (1990), Coleman et al. (1991), Coats and Fant (1992), Tam and Kiang (1992), and Fletcher and Goss (1993). In addition, Salchenberger et al. (1992) demonstrated an ANN model that, at a minimum, performed as well as LA. Furthermore, Salchenberger et al. showed that if the cut-off value was decreased in an attempt to reduce Type I Errors, then the subsequent increase in Type II Error was less for their ANN model when compared with LA. Overall, the successes that ANNs have had has been in classification and short term (mostly one year) prediction, which is mainly due to the difficulty of adding a time element to ANNs (Kumar and Tan, 2005).

The major criticism of ANNs is that they are a black-box approach. Although they output a continuous score that can be compared with cut-off values to generate failure/success predictions, the internal logic is hidden from users. That is, it is not possible to gain a full understanding of the significance of each ANN input variable and model interpretation is nearly impossible for complex ANNs such as those needed for BFP. However, there has been recent research into methods for converting ANNs into transparent models after they have been trained. For example, RULEX is a rule-extraction ANN program, developed at Queensland University of Technology, that is designed to enable the extraction of rules from an ANN model [RULEX source code available at http://sky.fit.qut.edu.au/~andrewsr/rulexsoftware.html]. The inter-related rules extracted from RULEX can then be efficiently implemented as a decision tree or other logical structure.

Genetic algorithms (GAs) have been proposed to overcome the difficulties associated with training ANNs by automating the training process. GAs are an optimisation technique based on Darwinian evolution. Anandarajan et al. (2001) presented a GA trained ANN model that had greater predictive ability than both a back-propagation ANN and DA model. Fanning and Cogger (1994) also presented a similar model that was comparable to LA and superior to a back-propagation ANN in most instances.

These soft computing models are important as they offer qualitative methods that traditional quantitative tools in statistics and economics can not quantify due to the complexity in translating

the systems into precise functions. Further information on ANNs, and their application in BFP, is provided in the excellent book by Tan (2001).

Sequential Procedures

Healy (1987) presented cumulative sum (CUSUM) procedures that detect a shift in a series of variables' values from a 'good' distribution to a 'bad' distribution. CUSUM procedures, which date back to 1954, are a set of sequential procedures based on likelihood ratios. These procedures attracted the CUSUM name as they reduce to calculating cumulative sums for many common distributions. CUSUM procedures detect the optimal starting point of the shift and then provide a signal of the shift as soon as possible after the shift occurs. Healy demonstrated the application of CUSUM procedures for detecting shifts in the mean and covariance matrix of a multivariate normal distribution. He noted that this method is very efficient at detecting a shift in the mean, when the mean of the 'good' and 'bad' distributions are known. However, it was also noted that the CUSUM procedures were still applicable in situations when at least the direction of slide towards the 'bad' distribution was known, as is the case in BFP.

This CUSUM concept had obvious applications in BFP, but were not noted by Healy. Nonetheless, Theodossiou (1993) further developed Healy's model and outlined a sequential procedure for detecting when a business shifts from 'good' financial performance to 'bad' financial performance. In terms of BFP, the CUSUM model is a dynamic time-series extension of DA. That is, the model explains a business's tendency towards financial failure over time. A desirable feature of the CUSUM model is that good performance is forgotten quickly in comparison to poor performance, which is remembered longer. This is a desirable feature as it inherently treats Type I Error correctly as more serious. Empirically, Theodossiou convincingly showed that this model was superior to DA with his data set.

Kahya and Theodossiou (1999) further extended the work of Theodossiou (1993) by overcoming the problems of non-stationary variables and improving their definition of financial success. In addition, several minor statistical refinements were made to their CUSUM procedures. The best CUSUM model was then chosen by a neural network search procedure from an initial set of 54 explanatory variables (comprising 27 popular variables and their first differences). Interestingly, models using the most popular explanatory variables were found to be non-stationary with deteriorating forecasting performance over time. The final CUSUM model chosen had superior predictive power compared to both DA and LA. Despite these positive empirical results, CUSUM procedures have not been widely used in BFP. The exact reason for this is unknown, but the greater complexity of the CUSUM procedures appears to be the most likely reason for its low popularity.

Other Techniques

In addition to the techniques already presented, there are many other lesser-known techniques that have been applied to BFP.

Together with LA and PA, the Linear Probability Model (LPM) was suggested as an alternative to DA, because its score is bounded between 0 and 1 and can be directly interpreted as a probability of failure. Nevertheless, LPM is rarely used in BFP due to its inferior empirical performance. Theodossiou (1991) applied all three techniques to BFP and found that LA and PA had very similar classification and predictive ability that was considerably superior to LPM. More recently, econometric modelling has been applied to BFP (Kumar and Tan, 2005). This technique involves combining a large number of predictor variables, both exogenous and endogenous, into many

computationally intensive inter-related regressions. As the number of variables is very high the level of parsimony and predictive ability of econometric models is poor; however, the main goal of econometric modelling is furthering the understanding of the failure process.

In addition to being used to optimise the training of an ANN, Genetic Algorithms (GAs) have been independently applied to BFP. GAs are a type of evolutionary algorithm that involve searching for an optimised solution based on Darwinian Evolution, and its concept of survival of the fittest. A comprehensive review of GAs in general has been conducted by Bornholdt (1998). When applied to BFP, GAs search for the optimal set of rules to classify and predict business failure. The final optimised set of rules can easily be extracted from the GA model and then interpreted similar to rules in an expert system. Shin and Lee (2002) applied GAs to BFP and concluded that this GA rule extraction approach has significant promise.

Wilcox (1976) applied the Gambler ruin model taken from probability theory to predict business risk. Other lesser-known, and more complex techniques that have been applied to BFP include, chaos (or catastrophe) theory (Scapens et al., 1981), multicriteria decision aid (Zopounidis and Dimitras, 1998) and rough sets (Dimitras et al. (1999); Zopounidis and Dimitras (1998)). These alternative techniques, together with the more popular techniques, have appeared in the recent literature reviews by and Aziz and Dar (2006) and Balcaen and Ooghe (2004). Both of these papers have included excellent summary tables of the literature they have reviewed. Moreover, Balcaen and Ooghe (2004) addressed the question of whether these more complex alternative techniques produce better models. The findings were not definitive, but indicated that there may not be any benefits associated with using more sophisticated alternative techniques.

Altman and Saunders (1988) present an excellent review (from an accounting viewpoint) of the BFP studies up to 1998 and Cybinski (2003) has produced a book containing a more recent review of the BFP field. (Zopounidis and Dimitras, 1998) also contains an excellent literature review that discusses most of the techniques discussed here. Laitinen and Kankaanpää (1999) is another recommended summary paper that presents a good theoretical analysis and empirical comparison of DA, LA, ANNs, HIP, survival analysis and recursive partitioning (a decision tree approach).

3. Decision Trees

Decision Tree (DT) techniques generate a set of tree-based classification rules and construct a decision tree (also known as a classification tree). DTs assign input objects to a group from a predefined set of classification groups: in the case of BFP, a DT usually assigns businesses to either the successful or failing group. In general, DTs are binary trees, which consist of a root node, non-leaf nodes and leaf nodes connected by branches, whereby each non-leaf node has two branches leading to two distinct nodes as shown in Figure 4.

When applied to classification problems such as BFP, leaf nodes that represent classification groups (fail or success) and the non-leaf nodes each contain a splitting (or decision) rule. Thus, the tree is built by a recursive process of splitting the data when moving from a higher level of the tree to a lower level. The splitting rules at each node define the details of the split. The splitting rules comprise an expression (usually containing one financial ratio) that is evaluated for each case (business) and compared to a cut-off value. For example, a splitting rule might be to classify a business into the

- Left sub-tree if current ratio < 2.1, or

- Right sub-tree if current ratio ≥ 2.1.

Splitting rules are usually univariate as shown above, but the same variable can be used in zero, one or many splitting rules.

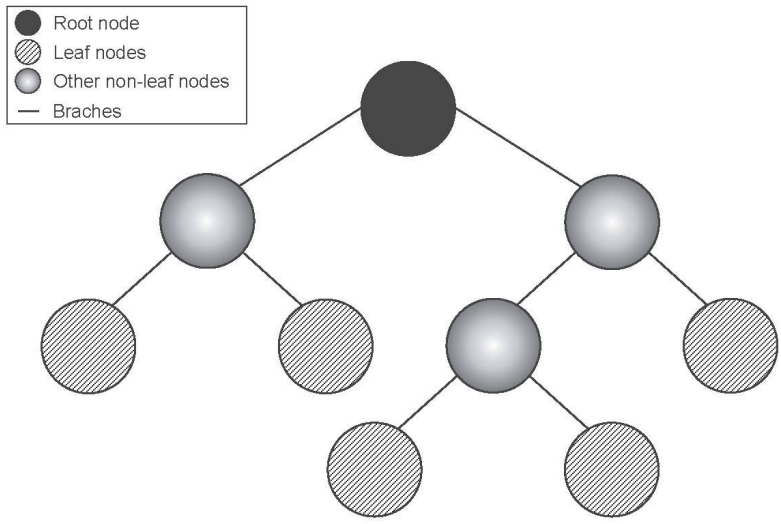

Figure 4: The basic structure of a binary tree

Similar to supervised learning with ANNs, DT building algorithms are used to manage the creation of DTs. There are two main tasks a DT building algorithm performs:

1. Choosing the best splitting rule at each non-leaf node that discriminates between successful and failing businesses; and,
2. Managing the complexity (number of nodes) of the DT, which includes the decision of when stop the process and use the current DT as the best DT.

There are different DT building algorithms that can be used to build the decision tree, each possessing its own method to determine the best splitting rule and best decision tree. Although similar tree structures are created by all building algorithms, the choice of which algorithm to use often has a large influence on the accuracy of the final DT model. The important DT building algorithms from a BFP viewpoint are the recursive partitioning algorithm, and entropy algorithms such as Classification and Regression Trees (CART) and See5.

Theoretical Analysis

Unlike the parametric discriminant and logit analysis methods, the major advantage of DTs is that they are non-parametric. This means that DTs make no distribution assumptions about the underlying data, and consequently there is no violation of distribution assumptions and there is no need to consider transforming variables (such as a log transformation). The only assumptions that DTs possess are that the successful and failing groups are discrete, non-overlapping and identifiable, which are common assumptions in all other statistical approaches. DTs can also handle missing values and qualitative data, as well as being easily represented in a user friendly graphical format (Joos et al., 1998). Another advantage of DT models is that they can take different misclassification costs for Type I and Type II Error as inputs, which can then be incorporated into the DT

building process at all stages. This is preferred to adjusting the cut-off values after the model has been generated, as is the case for the more popular techniques of DA and LA. However, a disadvantage of many DT building processes is that they require prior probabilities of successful and failed businesses as inputs. These prior probabilities are usually arbitrarily estimated, which adds an element of imprecision to the DT.

DTs are easy to interpret as the splitting rules are usually univariate. This also allows for easy identification of significant variables, where the root node contains the most significant variable. The relative significance of other variables can be found by comparing their proximity to the root node, with the closer nodes containing more significant variables. This process is accurate even when variables appear in more than one splitting rule, whereby the significance of a variable is measured by the proximity of the root to the closest splitting rule that contains the variable. However, DTs only identify the relative significance of variables, unlike DA and LA that provide quantified statistical figures that represent each variable's significance. Nevertheless, DTs still have the predictive power of a multivariate approach as there are sequences of nodes (univariate rules) that lead to each classification leaf-node. Furthermore, these sequences of splitting rules (in nodes) can naturally model interactions between variables without including an interaction term in the model, as has to be done with DA and LA. Although linear combinations of variables could also be used in splitting rules the potentially increased predictive power is not thought to outweigh the reduction in simplicity, ease of interpretation and identification of significant variables.

The DT approach has a discrete scoring system whereby the probability of group membership (failure or success) can not be calculated, and in addition, there is no way to compare and distinguish businesses in the same classification. These are the main disadvantages of DTs compared with the more popular DA and LA models that have a continuous scoring system. However, there is no need for determining cut-off values with a discrete scoring system like DTs. This is an advantage for DT techniques as the arbitrary assignment of cut-off values has been a major criticism of both DA and LA. Although it can be argued that DTs have cut-off values at each non-leaf, these values are calculated as part of the decision tree building process and are not arbitrarily set. Thus, the discrete scoring system of DTs has both advantages and disadvantages.

DTs have been criticised for their forward variable selection process. This criticism is for the case of a single variable being used in multiple splitting rules, when the DT building algorithms do not review the previous rules containing the variable when determining future rules containing that variable (Zopounidis and Dimitras, 1998). This is definitely a theoretical weakness of DTs, but all forward stepwise approaches, such as those commonly used in DA and LA, have this weakness. Furthermore, there is no evidence to suggest that this weakness will significantly reduce the classification and predictive ability of DTs.

Review of DTs in BFP

Following the successful application of DTs, using the recursive partitioning algorithm, to medical decision making by Goldman et al. (1982, 1988), they were first applied to BFP in the seminal work of Frydman et al. (1985). Since then, other types of DTs have also been applied to BFP with varying success.

Frydman et al.'s RPA DTs

The application of DTs to BFP by Frydman et al. (1985) was the first application of a non-parametric technique to BFP. As stated above, Frydman et al. used the recursive partitioning algo-

rithm to build the DTs, a method that has also become known as recursive partitioning analysis (RPA). The RPA takes as inputs the prior probabilities of businesses failing or being successful, the misclassification costs, and a set of training data (including the actual group classifications). The RPA algorithm then constructs the tree to minimise the expected misclassification cost (Jones, 1987), which is referred to as the resubstitution risk by Frydman et al. The resubstitution risk is calculated using probability theory and Bayesian reasoning.

With the aim of assessing the suitability of DTs (specifically RPA) to BFP, Frydman et al. compared two RPA built DTs with two DA models. The difference between the two DT models and the two DA models was the model size, in relation to the number of explanatory variables incorporated. The larger RPA-DT was created using the approach outlined in the paragraph above, while the smaller tree was chosen as the sub-tree of the larger tree with the best cross validation performance. V-fold cross validation models are calculated by dividing the data into V approximately equal groups and generating a DT for each (V-1) group (Frydman et al. used V=5). Each DT was then used to classify the group left out, and the best DT is chosen as the tree with the lowest average resubstitution risk. Similarly, the smaller and larger DA models were constructed using the 4 and 10 most significant variables respectively, according to a forward stepwise method.

The prior probabilities were set at 2% for failing and 98% for successful businesses, but the techniques were compared over eight different misclassification costs. All the models were formed from a data set of 200 randomly selected manufacturing and retail businesses, of which 58 went bankrupt. The predictors used in this study were 20 standard financial variables (mostly ratios) found to be significant in previous studies. The RPA-DT models were found to be superior classifiers of business failure in the original training data compared with the DA models. In addition, as may be expected, the more parsimonious (smaller) models outperformed their more complex counterparts on the cross-validation analysis. More importantly, the more parsimonious RPA-DT model was found to be a superior predictor to the DA models for almost all misclassification costs on the cross-validation analysis. However, the RPA-DT was found to be the worst technique for prediction in the cross-validation, which highlights the potential risk of over-training a DT model. Overall, after considering their theoretical analysis and empirical results, Frydman et al. concluded that decision trees using RPA were useful tools for predicting business failure.

Other Relevant Studies

Surprisingly, there have only been very few papers that have focused on decision trees since Frydman et al. (1985), despite the existence of many different DT building algorithms, including ID3, See4.5, See5 and CART. See4.5, which is an earlier version of See5, was presented as an improved version of ID3. Thus, See5 is considered to be superior to both See4.5 and ID3; however, an overall 'better than' comparison can not be made with CART as it implements a slightly different approach. All of these DT building algorithms are entropy algorithms based on the concept of information entropy. Entropy algorithms select splitting rules that produce the most informative split, which is equivalent to minimising entropy (or noise), rather than minimising expected misclassification costs as occurs with RPA. The tree size or complexity is then managed by choosing a minimalist DT that retains only splitting rules based on general patterns in the data. There is also another related technique that has been applied to BFP, known as Multivariate Adaptive Regression Splines (MARS). MARS is an extension of the CART technique that incorporates the DT concept of splitting the data into subregions using stepwise regression. Thus, MARS can be thought of as a procedure that searches for the optimal piecewise regression model (which may include higher order terms).

Joos et al. (1998) used decision trees to predict credit classifications for one of Belgium's largest banks; specifically, LA was compared with See5. Three models were created for both techniques based on three different data sets comprising a full set of financial variables, a reduced smaller set of financial variables and a set of qualitative variables. Unsurprisingly, Joos et al. concluded that regardless of the technique used, the models based on these three data sets were ranked according to classification ability in the order mentioned in the previous sentence. Furthermore, LA was found to outperform DTs on the full data set. However, as DTs are better able to handle incomplete and qualitative data, See5 outperformed LA on both the reduced and qualitative data set. Finally, Joos et al. analysed the classification accuracy once different misclassification costs were introduced. See5 maintained its superior performance on the reduced and qualitative data sets. In addition, for the eight different misclassification costs trialled, See5 was superior for five of these trials with less Type I Error. However, for the trials with the cost of Type I Error set the highest LA was superior. Thus empirically, there was no obvious overall superiority between See5 and LA.

Huarng et al. (2005) also used the See5 package, as well as being the first to apply CART to BFP. They found that CART was empirically superior to See5; however, the data sets comprised less than 12 businesses and 5 variables, which is too small to obtain reliable results. It should also be noted that Shirata (1998) used CART prior to Huarng et al. (2005), but it was simply to select the significant variables to use in their study of Japanese bankruptcies using DA. Shirata used CART for variable selection because it automatically calculates a variable importance score for each variable.

Although decision trees have not been a focus of many research papers, various studies have used decision trees as a comparison technique, because Frydman et al. (1985) showed it to be superior to DA. There have been mixed results from these studies. Tam (1991) and Tam and Kiang (1992) found that artificial neural networks (ANNs), DA and LA all outperformed ID3 decision trees. Fernández and Olmeda (1995) also found ANNs to be slightly superior to See4.5 and MARS, which produced similar results. Nevertheless, despite lower classification accuracy on the initial data set, See4.5 and MARS were found to have superior predictive ability over both LA and DA on the hold-out data set. In addition, MARS had superior classification accuracy on the initial data set compared to DA. Unlike most studies, Fernández and Olmeda (1995) also empirically tested combining the techniques: ANNs and LA alone and with other techniques were compared. Overall, models produced by combining techniques usually produced more accurate classifications and predictions. In contrast to these studies, Martinelli et al. (1999) found that See4.5 outperformed ANNs. Furthermore, Laitinen and Kankaanpää (1999) showed that RPA produced similar results to DA and LA (as well as survival analysis). However, all these studies have not considered differing misclassification costs.

4. Survival Analysis

A survival analysis technique is the term applied to a dynamic statistical tool used to analyse the time till a certain event. Thus, the SA approach to BFP is fundamentally different from the other approaches mentioned above. While other techniques model BFP as a classification problem, SA models BFP as a timeline, where businesses are represented by lifetime distributions. Lifetime distributions are distributions with a non-negative random variable that represents the lifetimes of individuals (or businesses) in some population. Lifetime distributions can be characterised by a number of descriptor functions, the most common being the survival or hazard function. The sur-

vival function S(t) represents the probability that a business will survive past a certain time t, while the hazard function h(t) represents the instantaneous rate of failure at a certain time t. The interpretations of these two functions is very different, but either one can be derived from the other.

There are many different SA techniques to estimate the survival and hazard descriptor functions. These techniques use past data to calculate the functions at each specific time, but they do not have the ability to make future predictions. Thus, they can be used to analyse past failure to help further the understanding of the failure process. The most popular of these is a non-parametric technique known as the Product-Limit, or Kaplan-Meier, estimator. There is also a less-popular technique called the Nelson-Aalen Additive Estimator. This technique has some statistical advantages over the Kaplan-Meier estimator, which are briefly discussed by Harrell (2001) in Chapter 16. In addition to these techniques, there are also different SA models that define relationships between one of the descriptor functions (usually the survival or hazard function) and the set of explanatory variables. These models can also be used for prediction and are estimated using regression.

Regression-based Estimation

The basic difference between various SA models is the assumptions about the relationship between the hazard (or survival) function and the set (vector) of explanatory variables (X). Thus, the general regression formula can be written as

$$h(t) = g\,(t, X^T\beta),$$

where X^T is the transpose of X, β is the vector of explanatory variable coefficients (also known as covariates) and g is an arbitrary function. In SA models estimated from regression it is customary to estimate the hazard rate, and then derive the survival rate as required. Traditionally, SA has been divided into two main types of regression models. These types are the proportional hazards (PH) and accelerated failure time (AFT) models, both of which have fully parametric and semiparametric versions (refer to Prashanthi (2005) for more details). Due to its flexibility, the most prominent model applied in the medical and business failure field is the semi-parametric PH model defined by Cox (1972). Cox's PH model (Cox, 1972) is defined as

$$h(t) = h_o(t)\,\exp(X^T\beta+c),$$

where:

- $h_o(t)$ is termed the baseline hazards function and describes how the hazard function changes over time and is the non-parametric part of the model; and,
- $\exp(X^T\beta+c)$ describes how the hazard function relates to the business specific explanatory variables and is the parametric part of the model, where c is an estimated constant. Note that some or all of the explanatory variables can be time dependent.

The regression coefficients β are calculated by an efficient method very similar to the maximum likelihood method (detailed in Kalbfleisch and Prentice (1980)). Furthermore, as with traditional regression techniques, the best explanatory variables are chosen from a starting set by forward or backward selection methods.

Theoretical Analysis

SA techniques are more sophisticated than the traditional popular techniques of discriminant analysis (DA) and logit analysis (LA). Except for sequential CUSUM procedures, SA is the only well-known technique that incorporates the time-series (or longitudinal) nature of BFP data into its model. Thus, SA does not assume that the failure process remains stable over time. All other cross-sectional models are only valid if the underlying failure process remains stable over time, which is a problem as the steady failure process assumption is usually violated in the real world (Laitinen and Luoma, 1991). This fundamental difference between the time-series SA models and cross-sectional traditional models also makes empirical comparisons between the techniques difficult. For example, a single SA model can make predictions of varying length; however, a single DA model can only make predictions of a fixed length based on its training data. Therefore, a single SA model is usually compared with many traditional models. This is an advantage in itself as one SA model is clearly more powerful in making different predictions than one traditional model.

The built in time factor in SA models allows them to model time-dependent explanatory variables. Zavgren (1985) found that in BFP the signs of the explanatory variable coefficients may change in different years before failure. Laitinen and Luoma (1991) went further and added that the values of the coefficients may also change relative to time before failure. Thus, an advantage of SA is the capability to model these changes, which can not be done with cross-sectional models. Therefore, SA models appear to be more suited to modelling a dynamic process, such as business failure, than cross-sectional models. This also means that theoretically, the predictive accuracy of SA models should be greater than that of both DA and LA.

Almost all well-known approaches assume that the data (businesses) comes from two distinct populations, which are those either going to succeed or fail. SA models do not make this assumption, but rather assume that all businesses come from the same population distribution. In SA models, the successful businesses are distinguished by treating them as censored data, which indicates that their time of failure is not yet known. This assumption more accurately models the real world (Laitinen and Luoma, 1991). SA models can also deal with the delayed entry and early exit of businesses from a study, which is likely to happen in business failure studies. Furthermore, SA does not make any of the restrictive distribution assumptions inherent in DA and LA, such as linearity. The semi-parametric and parametric SA models make some distribution assumptions, but they are not so commonly violated.

In addition to the easily interpretable probability of success or failure, SA models also produce the interpretable hazard function that is not available in other techniques. Analysis of the hazard rate can aid understanding of any process of death or failure (Harrell, 2001). Thus, SA is able to provide more information than other techniques, which is a significant component of any good model (Chatfield, 1995). SA also has the relative advantage over DTs of a continuous scoring system similar to that in DA and LA. However, this also means it has the disadvantage relative to DTs of modelling misclassification costs by varying cut-off values that are not incorporated into the model estimation process.

There are also a few disadvantages associated with the use of SA. There is evidence to suggest that the sample construction, specifically the proportion of failing and successful businesses, may affect the estimation of the SA model. However, this problem seems to be minor as most randomly selected BFP data sets contain a mixture of failing and successful businesses. Some researchers also identified that SA techniques (particularly the Cox model) are subject to multicollinearity problems, but these can be easily avoided by using standard forward and backward variable selection procedures. A more important disadvantage is that SA is designed to focus on determining the