EVALUATING EFFICIENT MARKET HYPOTHESIS WITH STOCK CLUSTERING

Graeme Allan Hitchman

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in fulfilment of the requirements for the degree of Master of Science in Engineering

Johannesburg 2014

DECLARATION

I, Graeme Allan Hitchman, declare that this dissertation is my own, unaided work. It is submitted in fulfilment of the requirements for the degree of Master of Science in Engineering at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

____

Graeme Allan Hitchman

_____ day of __________________ 20 ___

ACKNOWLEDGEMENTS

I would like to thank Dr. Campbell for assisting me with my project and always being available for consultation. Without his supervision this would not have been possible. Thank you to my family for allowing me to continue studying fulltime and supporting me throughout the first year. Finally to my varsity peers, for all the good times and encouragement throughout the project.

EXECUTIVE SUMMARY

This study investigates the validity of Efficient Market Hypothesis (EMH) by taking clusters of firms, generated using Self-Organising Maps (SOMs), and comparing their financial performance. Clusters were generated using 10 different financial variables as inputs to SOMs of different sizes. The effectiveness of the clustering was analysed using Silhouette Width, Davies-Bouldin Index and two Dunn’s Index metrics. The financial performance of the clusters was investigated using equal and value weighted returns and portfolio standard deviation. Market capitalisation was the only variable able to generate statistically significant results – in particular larger firms outperformed their smaller counterparts. It was concluded that this difference could be attributed to the volatile time frame chosen (2007-2012) which resulted in investors favouring larger firms. For future work it is recommended that researchers focus more on pre-processing the inputs, using different clustering methods (in particular fuzzy clustering) and conduct the analysis over a longer time frame.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

AMEX	American Stock Exchange
ANOVA	Analysis of Variance
B/ M	Book-to-Market
CAPM	Capital Asset Pricing Model
CMP	Number of Companies
CRISP-DM	Cross-Industry Process for Data Mining
CSP	Centre for Research in Security Prices
DB	Davies-Bouldin
DE	Debt/ Equity
DI	Dunn’s Index
EMA	Exponential Moving Average
EMH	Efficient Market Hypothesis
GMDH	Group Method of Data Handling
GUI	Graphical User Interface
IQR	Interquartile Range
JSE	Johannesburg Stock Exchange
M1	First Multivariable Test
M2	Second Multivariable Test
MACD	Moving Average Convergence Divergence
MC	Market Capitalisation
NASDAQ	National Association of Securities Dealers Automated Quotation
NRN	Neuron Number
NYSE	New York Stock Exchange
OOP	Object Oriented Programming
P/ E	Price/ Earnings

PB	Price/ Book Value
PC	Price/ Cash Flow
PCA	Principal Component Analysis
PE	Price/ Earnings
QR	Quick Ratio
RA	Return on Assets
RAFI	Research Affiliates Fundamental Index
RE	Return on Equity
RSI	Relative Strength Index
SMA	Simple Moving Average
SOM	Self-Organising Map
SW	Silhouette Width
U-matrix	Unified Distance Matrix
V	Volatility

1. INTRODUCTION

The following section has been broken down into a brief background where related studies have been mentioned. Thereafter, a motivation for the current work has been provided followed by a brief outline of this report. A more detailed and thorough literature has been completed in Section 2.

1.1 Background

Clustering algorithms have been used in a range of studies within the financial sector for data exploration. The studies presented here serve to provide a brief background to prior work which is necessary for the project motivation. Wang et al. [1] used fuzzy relations for the purpose of clustering due to there being no need for a predefined number of clusters. The study was designed to cluster financial ratios, thereby providing insight into which ratios should be considered the most significant.

Enke et al. [2] designed a three stage system for stock market prediction using statistical analysis, fuzzy type-2 clustering and neural networks. It was concluded that this proposed method of stock prediction was an improvement on older models. Gafiychuk et al. [3] used both the self-organising map (SOM) and group method of data handling (GMDH) algorithms to perform a cluster analysis using price data for the Dow Jones Index. The results showed that the clustering was successful for highlighting relationships between different companies and their respective industries.

Stock market clustering has also been researched on a variety of stock markets. For example Nanda et al. [4] completed a stock clustering analysis on the Indian stock market. This research required the clustering of stocks for the purpose of creating a diverse portfolio . The clustering methods used were K-means, fuzzy C-means and SOM. From this study it was concluded that the use of clustering may help with efficient portfolio generation and the K-means clustering produced the most compact clusters.

Liao [5] analysed the Taiwan stock market using two data mining algorithms. Initially an Apriori algorithm was used for association and the second method employed was K-means clustering. Finally the data provided could be used to present possible investment portfolios based on the information gathered. The results successfully presented different portfolio options and it was concluded that the work should be continued in further research.

Stock clustering completed by Wang [6] used K-means clustering to cluster stocks using a two stage method. Initially the stocks from the China Shanghai 180 were clustered based on their price-earnings ratio and turnover. They were then clustered using eight fundamental ratios and based on these results new clusters were formed. Wang concluded that the work was successful, however, further investigation into more variables, DuPont analysis and different clustering techniques were recommended.

Using 12 financial ratios Kelvin and Sian [7] clustered 470 stocks from the S&P 500 using the SOM algorithm. Two large SOMs (12x9 and 24x18) were used and the clustering was completed by visualisation of the U-matrix instead of each neuron defining a new cluster. In the results it was found that data normalisation and outlier removal played a significant role due to this visual technique. It was concluded that the results obtained would be useful to a financial analyst.

Silva and Marques [8] used a SOM to cluster 48 stock price time series. Missing data was interpolated (using the last known value) and normalised. It was found that the SOM was robust in terms of outliers and was capable of clustering stocks with only partially similar time series. It was concluded that the results could be applied to stock selection for portfolios by taking shares from different portfolios in order to diversify the portfolio.

In addition to the financial sector it is possible to incorporate data mining in other industries and much attention has been given to comparing the different clustering methods. Aguado et al. [9] completed multivariate data analysis on waste water processes using principal component analysis (PCA) and SOMs. These methods were used in order to assist in the visualisation of the variables. Afterwards K-means clustering (based on the Davies-Bouldin Index) was done and the clustering results achieved proved to be effective for the given data set. Aguado et al. also concluded that the visual results from PCA and the SOM were equally effective and assisted greatly in providing information regarding the relationships between variables.

Budayan et al. used traditional methods, fuzzy C-means and SOMs for clustering strategic groups within the Turkish construction sector. It was concluded that the SOM was visually superior to the other methods and fuzzy C-means overcomes the simplistic grouping obtained from traditional methods [10].

1.2 Motivation

The purpose of the proposed research dissertation is to investigate the possibility of using data mining and clustering to evaluate Efficient Market Hypothesis (EMH) and portfolio generation. The clustering work presented thus far has provided limited insight into the choice of financial variables. In comparison financial studies have focused on the predictive ability of specific variables and a large body of work exists regarding these variables and EMH. The motivation for this research is to provide more insight into the individual financial variables and their effect on the clustering process. The methodologies employed in this study are intended to provide results to bring together these previous studies in a complementary manner.

Numerous methods of clustering exist, however, the use of neural networks is relatively new. More specifically, SOMs are to be used for the purpose of clustering and offer a powerful method of reducing multidimensional data to a more manageable scope [11], [12], [13]. Neural networks are commonly used for prediction or classification of financial data, however, the proposed research is targeted at clustering the data, thereby resulting in a more descriptive approach. Recently SOMs have been used for clustering in a wide variety of fields and have proved to be relatively successful.

The majority of work has been completed on stock markets such as the New York Stock Exchange (NYSE). It is possible that smaller less developed markets, such as the Johannesburg Stock Exchange (JSE), could exhibit different behaviour to more developed markets. In addition to this the JSE is relatively smaller and could therefore make a more manageable scope for research of this nature, enabling a variety of variables to be analysed in more detail. With significant improvements in computing power and data storage the amount of available data, regarding share information for the JSE, has become abundant. Financial ratios as well as technical analysis are commonly used in an attempt to predict market behaviour; however the large number of possible variables creates a situation where the data becomes multivariate. For this reason it will be necessary to determine a manageable scope, whereby only the most significant input variables are selected and analysed in more detail.

1.3 Outline of the Study

The purpose of this study is to evaluate the validity of EMH using clustering. The clustering was completed using SOMs with a range of SOM sizes and financial variable inputs. The methodologies used are intended to build on traditional techniques while adding insight into the variables from a different perspective. The remainder of this report will be structured as follows:

Section 2 provides a review of related studies as well as a theoretical background. In Section 3 the objectives for this study have been identified. The data used for the purpose of this study has been presented briefly in Section 4 along with reasoning behind some of the assumptions made. In Section 5 the methodology has been explained and the results from the study are then shown in Section 6. In Sections 6 to 8 the results along with a discussion and conclusions are presented. Finally in Section 9 recommendations for future work are discussed.

2. LITERATURE REVIEW

An overview of the literature related to both finance as well as different aspects of data mining and clustering is presented in this section. Particular attention has been given to research regarding financial anomalies as well as clustering validity.

2.1 Stock Analysis

In order to attempt to profit from the buying and selling of stocks it is necessary to perform an analysis of the shares to determine whether they are under-priced or over-priced. Currently there are two main schools of thought on how shares should be analysed, namely fundamental and technical analysis. These methods may be employed together or independently in an attempt to determine the intrinsic value of shares.

2.1.1 Fundamental Analysis

Fundamental analysis refers to the interpretation of the financial ratios used to describe a firm's performance. Once all the financial ratios have been calculated it is necessary to compare the results to market benchmarks, past results, as well as industry averages to provide a relative scale of comparison. The financial statements required for the ratio calculations are broken into three distinct sections, namely the balance sheet, income statement and the statement of cash flows [14]. Gibson separates the financial ratios into four distinct categories (liquidity, long term debt paying ability, profitability and investor analysis), each explaining a different aspect of the firm [15].

When calculating the return on shares it is important to note that the dividends received must be included (Equation 1) [16]. In this equation R is the return over the chosen time (0 to *T)*; $P_0$and $P_T$are the prices at times 0 and *t* and D_T is the dividends paid out over the time frame.

$$R = \frac{P_T + D_T - P_0}{P_0} \tag{1}$$

The return shown in Equation 1 may also be referred to as a simple return because it does not account for the reinvestment of the interest acquired during time *0-T*. Continuous compounding return is commonly employed in financial studies [17], and it may be derived from simple return using Equation 2 [18], [19]. Here R is the return as calculated in Equation 1 and r is the continuous compounding return.

$$r = \ln(1 + R) \tag{2}$$

Continuous compounding returns are commonly used for statistical purposes [18], [19] and these returns may be assumed to follow an approximately normal distribution [20].

It is also important to note that companies are able to manipulate the income statement and balance sheet results [21], [22] however the cash flow statement will reflect these accounting inaccuracies. It is for this reason that several ratios compare the income statement and cash flow statement results.

2.1.2 Technical Analysis

Technical analysis is mostly concerned with the price and volume of shares traded, and employs many mathematical tools in the analysis of the share data. Murphy defines technical analysis as “the study of market action, primarily through the use of charts, for the purpose of forecasting future price trends” [18].

Trend Analysis

Trend analysis is considered to be central to technical analysis [23] and the concepts employed are relatively simple. In brief there are three directions a trend may move, (downward, sideward and upward) all of which are self-explanatory. In addition to this, trends may be further classified according to the time frame over which they occur [23], [24].

Many additional trends exist to explain a wide range of share price phenomena. However, this detailed trend examination is beyond the scope of this report.

Moving Averages

Moving averages is a technique developed from statistics whereby an average is calculated for the last n data points. As new data becomes available the average is updated, hence the average is continually changing [25]. By using moving averages unwanted noise may be removed and it can assist in defining a trend. These trends do however lag the actual data [26].

The three most commonly used methods of smoothing are simple, weighted and exponential [23], [26]. The simple method weighs all data points evenly. In weighted and exponential more significance is given to more recent data and it becomes more responsive [26], [27]. The formulas for a simple moving average (SMA) and exponential moving average (EMA) are given below [28].

$$SMA(t) = \frac{1}{n}\sum_{i=0}^{t} C_i \tag{3}$$

$$EMA(t) = \frac{2}{n+1}C_t + \left(1 - \frac{2}{n+1}\right).SMA(t) \tag{4}$$

In Equations 3 and 4 n is the number of recorded prices and C_i is the closing price on day – similarly C_t is the closing price on day *t.*

Oscillators

The Relative Strength Index (RSI) is shown in Equation 5. As with moving averages, RSI is greatly affected by the chosen time frame. Most technical analysts use 9 or 14 days for the calculation [23].

$$RS = \frac{average\ of\ x\ days\ when\ market\ closed\ up}{average\ of\ x\ days\ when\ market\ closed\ down} \tag{5}$$

$$RSI = 100 - \frac{100}{1 + RS} \tag{6}$$

Stochastics is based on the assumption that during downtrends prices close near the bottom end of the price range and the converse during uptrends. This oscillator consists of two lines, with %D simply being the moving average of the %K line. The %K line equation is shown in Equation 6, where the number of days (n) is typically 14 [23].

For Equation 7, C is the latest closing price; L_n is the lowest price for the previous *n* periods and H_n is the highest price for the last same period. It is also important to note that the period *n* can be measured in a chosen time frame, i.e. days, weeks, months, etc.

$$\%K = 100[(C - L_n)/(H_n - L_n)] \tag{7}$$

Moving Average Convergence Divergence (MACD) is the difference between the long and short term EMAs (as shown in Equation 7). The time frames used are usually 12 and 26 days and it is used for identifying trends as well as changes in trends [28] , [29], [30].

$$MACD = EMA_{short}(t) - \ EMA_{long}\ (t) \tag{8}$$

2.2 Fundamental Financial Theories

Financial and economic models are constantly being developed in an attempt to explain the behaviour of the financial market. Some of the fundamental theories are presented below and serve as an introduction to Section 2.3 where economic anomalies are discussed in more detail.

2.2.1 Efficient Market Hypothesis

The origin of efficient market hypothesis (EMH) can be found in research completed by Fama [31], who concluded that markets followed a random walk model. This is due to the market being information efficient and since the information is irregular and random the stock market follows a random walk [32], [33].

In brief EMH states that share prices correct themselves in accordance to any new data so rapidly that it is not possible to consistently yield greater than average returns on the stock market. The share price encompasses past information regarding the share, therefore implying that the market is operating efficiently. Before examining the EMH in more detail it is necessary to consider some of the fundamental assumptions related to this theory [34] [35].

1. There must be many investors who make rational decisions and they act upon new information as it becomes available.
2. Irrational decisions made by investors are unrelated and cancel out, thus having no net effect on the stock price, making the market rational.
3. New information becomes available randomly and must be independent of past information.
4. There are no taxes or transactional costs.

EMH can be broken down into three primary forms, namely weak, semi-strong and strong [5]. In weak form EMH the stock price is said to contain all the price information related to the stock, which includes past price data. In semi-strong form EMH all the public data is contained within the current stock price. Finally in the strong form of the EMH all information, including private, is contained within the stock price [14].

Since the development of the EMH there have been numerous studies with many opposing opinions on the validity of EMH. Many papers have investigated the use of financial ratios as a proxy for stock returns, as well as price patterns and behavioural anomalies. Although EMH has undergone extensive studies it still remains relatively robust. More recent studies related to behavioural finance have shown support against the validity of EMH [14], [36].

2.2.2 Portfolio Theory

Reilly and Brown define two requirements for an investment portfolio [14].

- It must achieve above average returns for its assumed risk.
- It must be diverse to avoid unsystematic risk.

Portfolio theory was initially developed by Markowitz and uses the variance in the return on an investment to define its risk [14]. The variance, σ^2, for asset i is defined by Equation 9 and the covariance, $Cov(R_i, R_j)$, between two assets i and j is shown in Equation 10 [14], [20].

$$\sigma^2(R_i) = E[R_i - ER_i]^2 \; where \; EX = \sum_{i=1}^{n} P(X_i)X_i \tag{9}$$

$$Cov(R_i, R_j) = E\big[(R_i - ER_i)(R_j - ER_j)\big]^2 \; where \; EX = \sum_{i=1}^{n} P(X_i)X_i \tag{10}$$

The above equations refer to the ex-ante calculation for the variance and covariance. In these equations, $P(X_i)$ is the probability of an event, R_i is the possible return and ER_i is the expected return for i. This implies that the expected returns, of each investment, are determined based on the future expected returns. In contrast the above calculations can also be completed ex-post i.e. based on historic data [20].

In order to evaluate the standard deviation for a portfolio consisting of numerous investments it is necessary to account for the risk of each investment as well as the correlation between the investments. By doing so it is possible to define the risk of a portfolio according to the risk of returns as well as the degree of diversification (as shown in Equation 11 where σ_{port} is the portfolio standard deviation, w_x is the weighting for investment x and n is the number of investments) [20].

$$\sigma_{port} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j Cov(R_i, R_j)} \tag{11}$$

As previously mentioned, a portfolio must achieve a greater return for its associated risk. By using the above formula for defining risk, it is possible to determine a portfolio’s return in relation to its’ associated risk. A commonly used formula for relating the return and risk for a portfolio is known as the Sharpe ratio (S_h in Equation 12 below) [20].

$$S_h = \frac{\overline{R_p} - \overline{R_F}}{S_p} \tag{12}$$

The numerator can be viewed as the portfolio's excess return (over a risk free investment). It should be noted that since the standard deviation of the portfolio (S_p) is measured as a percentage, the Sharpe ratio is dimensionless [20].

2.2.3 Beta and the Capital Asset Pricing Model

The Capital Asset Pricing Model (CAPM) is a single variable model which is used to relate the expected return of an asset with its associated risk. The risk of the asset i is defined as beta (β_i) and is shown in Equation 13 where M refers to the market portfolio [14].

$$\beta_i = Cov(i, M)/{\sigma_M}^2 \tag{13}$$

Rather than going into the derivation of the CAPM in this paper one can refer to Brown and Reilly [14] where a thorough explanation is provided.

From the CAPM it can be seen that the expected return of asset i above that of the market is proportional to market risk (β) [37]. The effectiveness of CAPM has been frequently debated and numerous sources argue that it is no longer effective. However, other sources still regard it as a useful tool for asset pricing [38].

2.2.4 Fama and French Three Factor Model

Work completed by Fama and French evaluated the use of different variables (size, book-to-market value, leverage, earnings-price and market beta) for the purpose of stock return estimation [39]. It was concluded from this work that the asset pricing model (which only uses market beta) was valid for the period between 1926 and 1968; however from 1963 to 1990 there is no relationship. Furthermore firm size and book-to-market value may be used to explain stock returns from 1963 – 1990 [40]. In this period there is a negative relationship between return and size and a positive relationship between return and the book-to-market equity [39]. From this work Fama and French derived a model based on three variables, namely market beta (from CAPM), market value (size) and book-to-market value [41].

2.3 Additional Financial Studies

Numerous financial and economic relationships have been researched and of particular interest is the comparison between value and growth stocks. Value investing (which was originally popularised by Benjamin Graham) refers to the method of investing in shares which appear to be undervalued. This can be seen by their low price-earnings, market-to-book value and price-cash flow ratios. In contrast growth stocks have high price-earnings, market-to-book value and price-cash flow ratios [42], [22].

Fama and French [43] found that by taking stocks from 13 different markets, between 1975 and 1995, resulted in the value stocks, on average, outperforming their growth counterparts by 7.68% when sorted according to book-to-market values. Although it is widely recognised that value stocks have historically outperformed their growth counterparts there is still some debate as to why they are able to yield greater returns.

The following sections expand upon this initial comparison by comparing numerous studies which evaluated the various financial ratios and relationships.

2.3.1 Contrarian Investing and the Overreaction Hypothesis

As previously mentioned value stocks have generally obtained greater returns than growth stocks. It has been speculated that this is due to investors overreacting to positive and negative news, thus causing the stock prices, at either end of the spectrum, to overreact to information. In addition to this investors often extrapolate past information with little understanding, thus compounding the above mentioned overreaction. This forms the basis for contrarian investing, which requires investors to invest in contrast to the general market (naïve investors) [44], [45]. The overreaction hypothesis states that stocks which have performed poorly (losers) achieve greater returns than the market [46]. This has also been extended in the evaluation of previously well performing stocks (winners) attaining poor results.

De Bondt and Thaler [47] hypothesized that if the overreaction occurs systematically, then based on historic returns (and without the need for accounting data), abnormal movements in the stock price should be followed by an opposite price movement. In addition to this, the subsequent price movement should be proportional to the abnormality of the original movement. De Bondt and Thaler [47] evaluated the effect of choosing portfolios based on loser and winner stocks using monthly returns for the NYSE from 1926 to 1982. The results showed that the loser stock portfolios outperformed the winners by 25% 3 years later and the loser portfolios were also found to be less risky. Furthermore, the loser portfolios achieved significant gains over the January period (known as the January effect [48]). Similar results, which confirmed the overreaction hypothesis, were again achieved in a later study [49].

By analysing the UK market (from 1955 to 1990) Clare and Thomas [46] also found that the losers outperformed winner portfolios over the following two year period. When controlling for size it was found that the phenomenon, in the UK market, was simply due to small firm effect. This conclusion agreed with the research by Zarowin [50], which concluded that the results achieved by De Bond and Thaler can be mostly attributed to small firm effect which is discussed in more detail in Section 2.3.2.

The contrarian investment strategy has also been evaluated in the French and German markets where it was found that the contrarian method of investing was more effective over a one year time frame than two or three years [51].

2.3.2 Size Effect

Various studies on the size effect of stock returns have shown that there is a negative relationship between the return and size of the firm [52]. In addition, this effect has been evident in developing stock markets such as Singapore [53] and Mexico [54]. It has been hypothesized that investors are reluctant to invest in small firms due to a lack of information thus enabling them to produce greater returns [52]. In addition to this there is a limited ownership of small stocks by institutions, thereby increasing the likelihood for incorrect pricing [55].

In contradiction to this Horowitz et al. [56] found no evidence to support the hypothesis that the small firm effect exists between 1980 and 1996. In agreement with Horowitz, Chan et al. [57] noted that between 1984 and 1998 a large-cap index significantly outperformed (18% versus 11% annually) its' small-cap counterpart. Moor and Sercu [58] argue that most of the research has ignored the smallest shares, due to a lack of data, and conclusions regarding small firm effect are incomplete. This statement is based on prior work completed by Banz [52], where it was noted that the size effect is most noticeable amongst the smallest firms.

2.3.3 Stock Liquidity

Liu describes liquidity as "the ability to trade large quantities quickly at low cost with little price impact" [59]. Acharya and Pedersen [60] expanded on the CAPM model, by including liquidity tests on the NYSE and AMEX, and found that constant decreases in liquidity decrease the immediate returns, however predict greater returns for a stock over a longer time frame [60].

Most studies have been completed in the United States where the stock market is relatively stable and liquid. However the Hong Kong stock exchange, which was analysed by Lam and Tam [61], provided a different data source with smaller average firm size and increased volatility. The results showed that liquidity was a significant factor for the returns on the Hong

Kong market and a four factor model, derived by adding liquidity to the Fama and French Three factor model, outperformed other models.

In comparison a study by Lischewski and Voronkova [62] showed that liquidity was not a significant factor in asset pricing for the Polish market. Although liquidity had little effect it was found that market beta, size and book-to-market values did affect the asset pricing. Chang et al. [63] found that a negative relationship existed between stock return and liquidity for the Tokyo Stock Exchange.

2.3.4 Dividends

Traditionally dividend producing shares have been considered an integral part of a portfolio. Dividend paying policies have undergone much research and contradictory theories have been developed in order to determine whether they still offer a suitable proxy for share returns. Currently there is still little evidence to attribute any correlations between dividend payments and share prices.

LaBarge and Hamilton [64] completed research which compared dividend payments with share repurchases. It was concluded that in more recent years the amount of dividend yielding companies has been diminishing and more companies are focusing on share repurchasing. Similar work by Brav et al. [65] consisted of a survey of 384 companies as well as in-depth interviews with 23 companies in order to evaluate the financial executives' views on dividends and share repurchases. The following points briefly summarise some of their findings.

- Often companies are hesitant to begin paying dividends (or increasing them) due to the future expectations thereby established and as a result share repurchasing is often viewed as a more flexible option.
- Companies avoid lowering dividends (even at significant costs) due to how it could be interpreted. Reducing repurchases is viewed as having fewer consequences.
- Executives believe that institutions view dividend payments and repurchases equally from an investment perspective.
- Companies repurchase shares when it is felt that the shares are undervalued. However, the share price has little effect on the issuing of dividends.

Work by Fama and French [66] revealed that the proportion of Centre for Research in Security Prices (CRSP) industrial companies (from the NYSE, AMEX and NASDAQ) which paid dividends dropped from 66.5 % to 20.8 % over the period 1978 to 1999. It was concluded that this declining propensity was partly due to an increase in small listed firms. Large firms were also found to have decreased their dividend payments [66]. In contrast to this, work by DeAngelo et al. [67] using the same firm database revealed that although the proportion of dividend paying firms from 1978 to 2000 decreased, the amount of dividends paid increased. This increase in dividends was due to the largest dividend payments increasing by such a large degree that the effect of the decrease became insignificant. In 2000 it was found that the top 25 dividend paying firms accounted for 54.9% of the dividends. Furthermore a two tier system within the market existed due to the high concentration of earnings and dividend paying firms as well as a significant correlation between a firm's earnings and dividend payments.

Ferris et al. [68] evaluated the declining propensity of dividend payments in the UK and Japan markets from 1990 to 2001, in order to determine whether dividend payment policies are market dependent. It was found that there was a marginal decline in dividend paying propensity in Japan, however there was little evidence supporting this trend in the UK market. Furthermore Japanese firms were found to not exhibit a concentration of dividend paying firms. In contrast the UK market was found to have a two tier system (similar to that found by DeAngelo et al. [67]) which also exhibited the same correlation between earnings and dividend payments.

Miller and Modigliani [69] hypothesized that in an ideal market a firm's dividend policy was viewed as irrelevant by shareholders. Subsequent studies have revealed opposing results for actual markets. Although the above research shows that fewer firms are paying dividends, Fuller et al. and DeAngelo et al. [70], [71] concluded that a firm's dividend policy is not considered irrelevant to shareholders and it was found that dividend paying firms outperformed their counterparts by a greater margin in a declining market.

2.3.5 Price-Earnings Ratio

The price-earnings (P/E) ratio (or inverted as the earnings price E/P ratio) is well known as a significant investment tool for the valuation of firms. For this reason, notable research has been completed with regards to the effect of the P/E ratio and work completed by Nicholson [72], Basu [73] and Reinganum [74] found that low P/E stocks yielded greater returns than high P/E stocks. These studies were completed on NYSE and AMEX stocks. Bildersee et al. [75] compared the P/E ratios effectiveness when applied to the Japanese equity market and it was found that the P/E ratio is less important in Japan. This was mostly attributed to different accounting procedures and policies, and once accounted for, the differences are reduced.

Giannetti [76] used quarterly price-earnings as a proxy for stock returns and found that the predictive nature of this ratio declined from 1997 to 2002. Giannetti [76] noted that by using daily price-earnings ratios the measure variability would be caused by the price volatility and this could be misleading. For this reason quarterly data was chosen for the analysis.

2.3.6 Cash Flow

The analysis of a firm's cash flow continues to grow in popularity [77] and it is often regarded as an appropriate measure for a firm's performance due to a greater robustness to misleading accounting manipulations [78]. This is because with accrual accounting the various transactions are reflected based on estimates which may differ from the actual amounts, whereas the cash flow statement reflects the actual cash flows as they occur. The accrual method of accounting has been designed such that it can provide more financial information regarding the firm's performance over the given time frame [79]. By comparing the figures obtained from the accrual method of accounting to the results in the cash flow statement it is possible for the analyst to determine whether the firm's results are indicative of their actual performance [79].

2.3.7 Book-to-Market Value

The book-to-market (B/M) value of a stock is often considered to be the leading indicator of stock performance [80], [81] and it is commonly used to separate value from growth stocks. In addition to this B/M is considered to offer a proxy for risk which is not explained by the return variance [82].

In contrast to the above, Daniel and Titman [83] contended that there is no significant relationship between the firm's historic performance and future expectations. Jiang's [84] research revealed that current information has a significant effect on how institutions invest and this often leads to an overreaction and contributes to the B/M effect [84]. It has also been found that the book-to-market effect is more prevalent in smaller firms. Chen [85] expanded on this and found that the B/M effect was most significant amongst firms which have a short life expectation.

2.3.8 Capital Structure and Leverage

The capital structure and leverage of a firm are often considered to be very important because by changing the leverage the managers affect numerous aspects of the firm's future performance [86]. A simple method of determining whether the effect of the leverage was beneficial is by comparing the return on assets and the return on common stockholder's equity. If the latter is greater than the return on assets then the leverage may be viewed as having a positive effect on the firm [87].

The extent of financial leverage still remains a subject of debate since the additional debt can impose overwhelming financial distress. It is also important to consider that there are potential benefits from the additional debt by enforcing managers to make appropriate financial decisions [88]. Lang et al. [89] concluded that a firm's level of leverage does not necessarily effect its growth if the firm has sufficient investment prospects. If the firm does not have adequate investment opportunities, or if it becomes overwhelmed by debt, then leverage is negatively related to growth. This implies that firms must be careful when considering leverage if adequate investment opportunities are not available.

Although the above provides a logical explanation with regards to the effects of debt and equity there is still much debate with regards to how firms manage their capital structure. One of the predominant theories is known as pecking order theory. In this theory it was hypothesized that firms choose internal financing over external financing, after which debt is chosen over equity. In comparison static trade-off theory assumes that managers make financing decisions based on an optimal amount of debt to take advantage of tax benefits, while considering the implications of bankruptcy. Based on this, managers set a debt/ equity target and work towards achieving the set ratio [90]. Other variables which have been considered significant in capital structure are information asymmetry (managers knowing more than investors and therefore being able to take advantage of this when issuing debt or equity) as well as financial flexibility [90], [91].

2.4 South Africa and the Johannesburg Stock Exchange

The Johannesburg Stock Exchange is the stock market of South Africa and is largely comprised of basic resources (as shown in Figure 2-1), and 410 firms were listed as of 31 December 2009 [92].

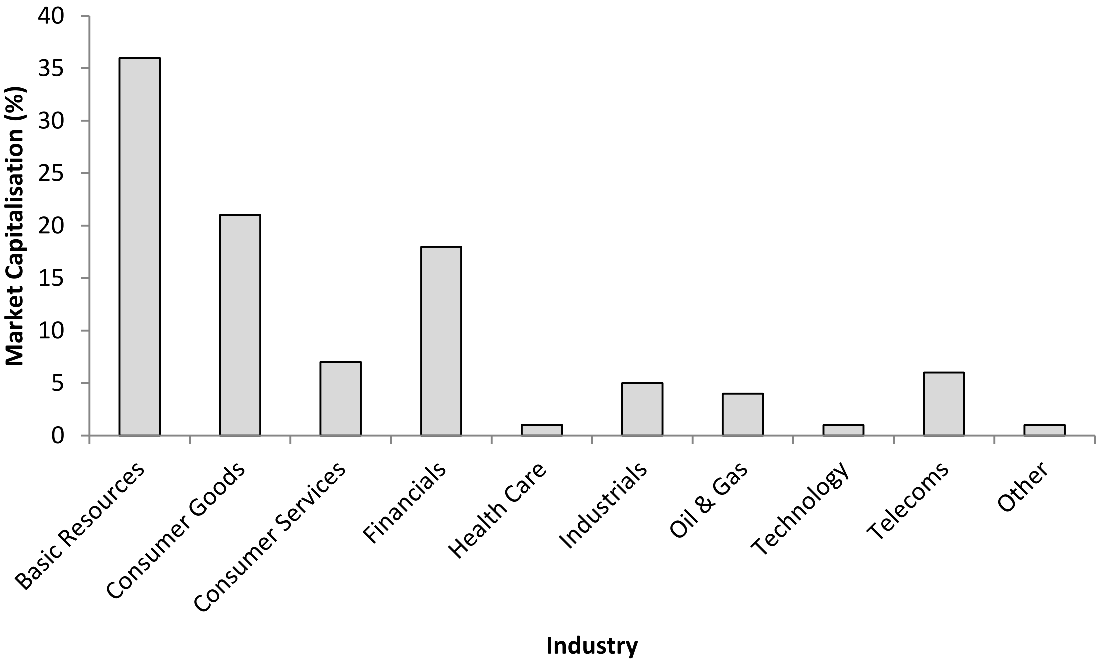

Figure 2-1: JSE Industry Market Capitalisation (Derived from [92])

2.4.1 South African Stock Indices

The JSE has several stock indices which have been selected to simulate market behaviour. A brief summary of several commonly used indices is presented below, with the data from 31 July 2012.

The JSE All Share Index and the JSE Top 40 have been designed to follow the market in a relatively similar manner. The JSE All-Share Index represents 99% of the market, where as the JSE Top 40 simply consists of the 40 largest firms, based on market capitalisation. In comparison, the net market capitalisation of the JSE All-Share Index is R 5 657 604 million, whereas the JSE Top 40 Index is R 4 697 367 million. This shows that the JSE Top 40 contributes to a significant portion of the JSE All-Share Index [93], [94].

The JSE Research Affiliates Fundamental Index (RAFI) evaluates firms based on four fundamental financial factors (dividends, cash flows, sales and book value) for the development of the index. The JSE RAFI All-Share Index consists of 137 stocks, of which the 10 largest stocks constitute over 50% of the value [95]. The JSE Dividend+ Index comprises the 30 stocks with the greatest expected dividend yields, and uses this information for the weightings. Although this index only consists of 30 stocks the top five only contribute to 23.3% of the index, which is smaller than the previously mentioned indices [96].

In addition to the above mentioned indices there are additional indices which follow other aspects of the market such as specific firm sizes, industrial groups as well as value and growth indices.

2.4.2 Risk Free Rate

The South African risk free rate can be calculated based on numerous government bonds and Figure 2-2 compares the bonds used as a proxy for a risk free rate in 2009/ 2010 [97].

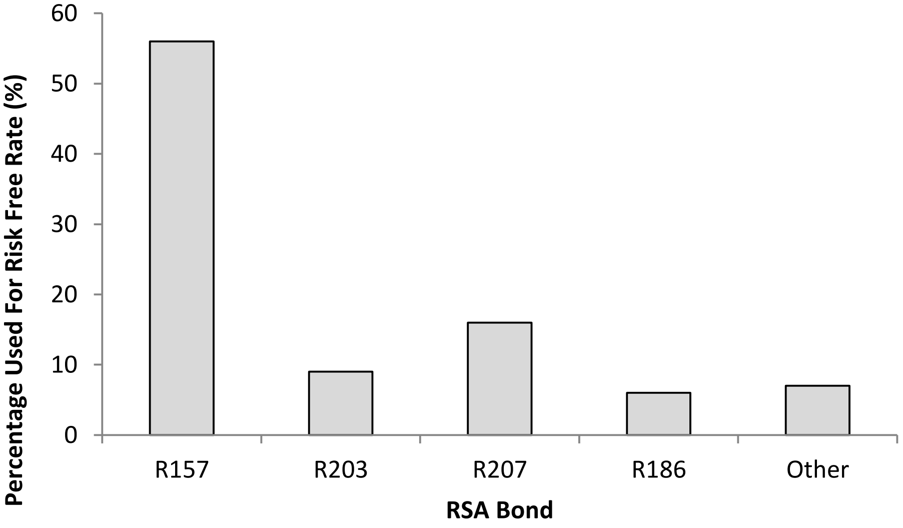

Figure 2-2: Proxies Used for Risk Free Rate (Derived from [97])

Table 2-1 lists the different bonds, which are commonly used as a proxy, and compares their various aspects.

Table 2-1: South African Government Bonds (Derived from [97])

Bond	Maturity (yyyy/mm/dd)	Coupon Rate (%)	Yield on 31 Jan 2010 (%)	Median Daily Traded Volume (2009)
R157	2015/09/15	13.50	8.38	11824
R203	2017/09/15	8.25	9.01	1784
R207	2020/01/15	7.25	9.19	1982
R186	2026/12/21	10.50	9.17	5787

2.5 Data Mining

Zaïane defines data mining as “the non-trivial extraction of implicit, previously unknown and potentially useful information in databases” [98]. In addition to this, other authors [99], [100] describe data mining as an automatic (or a semi-automatic) process of obtaining useful information. Data mining has been successfully applied to retail, banking, insurance and telecommunications as well as many other industries [101].

In general the results from data mining are either analysed in a descriptive or predictive manner. In predictive data mining the variables related to the samples are used to forecast future variables which are of importance. In comparison, descriptive data mining is primarily focused on highlighting patterns and relationships within the data which can then be further analysed by humans [102].

To assist with the task of data mining in a logical manner, data mining procedural standards have been developed.

2.5.1 CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a method of data mining and it can be broken into the steps shown in Figure 2-3. Many of the stages are iterative and require adequate preparation in order for the results obtained to be meaningful [101], [103].

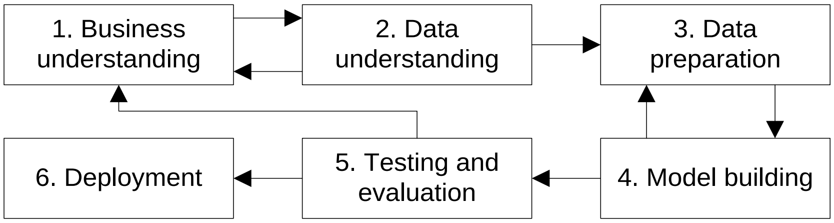

Figure 2-3: CRISP-DM Process (derived from [101])

1. As shown in Figure 2-3 the *business understanding* is initially required. This is to determine what the goals and objectives of the data mining process should be, and to assist in developing a project strategy.
2. *Data understanding* is completed with a statistical analysis of the data and relationships within the data may be identified. It is also essential to understand what is required from the data at this stage.
3. *Data preparation* must be completed by processing it into an appropriate form for the model building.
4. *Model building* refers to the use of data mining tools to interpret the data visually and with clustering techniques. Initially more basic methods may be employed and more advanced methods applied later.
5. *Testing and evaluation* requires that the results be evaluated in terms of the goals and objectives from the business understanding stage. This process may require that the objectives and goals be re-evaluated.
6. Finally the data may be *deployed* and features from the data mining process may be used within the business and should be constantly re-evaluated since the business environment is changing.

2.5.2 Data Mining Techniques

Patel separates data mining techniques into two categories, namely supervised and unsupervised techniques (shown in Figure 2-4) [104].

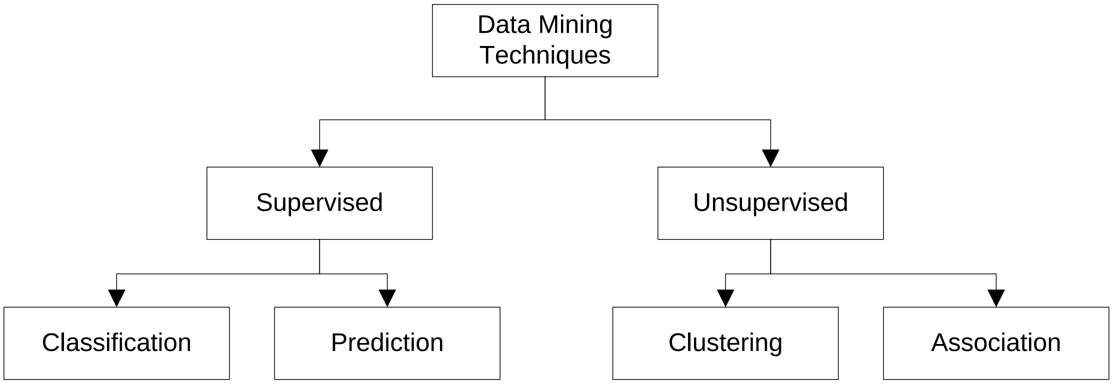

Figure 2-4: Data Mining Techniques

Each section in Figure 2-4 can be further broken down into the various methods employed, however only a brief description, of each technique, is presented below.

Classification

Numerous methods, such as decision trees, neural networks and support vector machines, are used for the purpose of applying a model which classifies data according to which predefined group it best matches. Applications of this technique range from fraud detection to classifying tumours [99].

Predictive

Predictive techniques, and in particular regression analysis, is a mathematical method which uses past data in an attempt to forecast future values. Its application in economics and finance has been widely researched and it can be applied to many other fields.

Association

Association analysis has been used in a wide range of applications, with one of the most common examples being the shopping basket analysis [99], [105]. The basic method of association involves computing a support and confidence for the occurrence of variables and using these values to predict the likelihood of occurrence x given y [99].

Clustering

Numerous methods of clustering exist, the most common of which are fuzzy C-means, K-means, hierarchical and SOMs, all of which have been applied to a variety of fields such as finance, biology and multimedia [4].

Clustering is a commonly used process within data mining and Chang et al. defines it as “an important unsupervised technique where a set of patterns, usually vectors in a multidimensional space, are used to identify groups (clusters) of similar characteristics” [106]. In order to achieve this it is therefore necessary to group the samples which are most similar, while maximizing the variation between different clusters [106], [107], [108].

Rapoport and Fillenbaum [109] note that one must be careful to not erroneously identify relationships within the data when in fact there are only random anomalies.

2.5.3 Data Normalisation

Data normalisation is required for pre-processing in many data mining applications. Three commonly used methods of data normalisation are min-max normalisation, z-score normalisation and normalisation by decimal scaling [110]. These three techniques are shown in Equations 14 to 16 where x' represents the x value after normalisation and x is the original value.

Specifically, in min-max normalisation the data is linearly transformed from an initial range (between min_i and max_i) to be within a new minimum (min_f) and maximum (max_f). This is shown in Equation 14 [110], [111].

$$x' = (x - min_i)\left(\frac{max_f - min_f}{max_i - min_i}\right) + min_f \tag{14}$$

Z-score normalisation (also known as zero-mean normalisation) uses statistical properties of the variables for the normalisation process [110], [111] (shown in Equation 15 where μ is the mean of the population and σ is the standard deviation).

$$x' = \frac{x - \mu}{\sigma} \tag{15}$$

Decimal scaling normalisation reduces the scale by powers of 10 (Equation 16). For the purpose of many applications the data range must lie within -1 and 1, therefore $|\max(X')| \leq 1$ is used to determine a value for c [111], [112].

$$x' = \frac{x}{10^c} \quad (16)$$

2.5.4 Statistical Analysis

Due to the large amounts of data analysed in data mining it is often necessary to incorporate statistical methods. Although numerous statistical tests are available to compare relationships between variables this study is primarily aimed at comparing groups or clusters. For this reason the statistical analysis methods presented in this section are related to the comparison of groups.

The tests for comparing groups can be broken down into two main categories namely parametric and non-parametric tests. In brief, parametric tests make assumptions regarding the sample distribution such as normality whereas non-parametric tests are often used for samples where the distribution is unknown or the data is ordinal. Some of the more common parametric tests are the t-test and ANOVA, which both have non-parametric equivalents such as the Wilcoxon Signed Rank test and the Kruskal-Wallis test.

2.5.5 Cluster Analysis

In order to interpret the results of the clustering process it is necessary to introduce measures of the clustering performance. These measures of clustering effectiveness are often referred to as clustering validity indices. Since the purpose of the clustering process is for the clustering algorithm to sort similar data points together, while simultaneously separating dissimilar points, it is logical to have measures which look at how similar and dissimilar points are.

The compactness or similarity of points within a cluster can be referred to as the homogeneity of the cluster. Numerous methods exist for measuring the homogeneity (H_{av}) and Equation 17 shows a simple method for this [113].

$$H_{av} = \frac{1}{N}\sum_{i} D(X_i, A_i) \quad (17)$$

In Equation 17, X_i represents a vector in the cluster, A_i represents the centre for that particular cluster and N is the number of vectors in the cluster. Therefore the homogeneity may simply be interpreted as the mean distance of the vectors within the cluster from their cluster centre [113].

The dissimilarity between clusters (S_{av}), also referred to as separation, may be measured using Equation 18 [113].

$$S_{av} = \frac{1}{\sum_{i \neq j} N_{Ci} N_{Cj}} \sum_{i \neq j} N_{Ci} N_{Cj} D(A_i, A_j) \tag{18}$$

In Equation 28 A_i and A_j are the centres of clusters and N_{Ci} and N_{Cj} are the number of vectors in each cluster. D is again the measure of distance between two vectors (in this case A_i and A_j) and S_{av} may be viewed as the distance between the centres of the clusters i and j, which have been weighted according to the number of samples within each cluster [113].

The measures of clustering effectiveness (shown in Equation 17 and 18) provide an insight into methods of measuring cluster validity; however more commonly used methods are presented below.

Silhouette Width

The Silhouette width is a commonly employed method of measuring clustering validity and it can be used to provide insight into how effectively each individual vector within a cluster has been grouped. Equation 19 shows how the Silhouette Width $s(i)$ for sample i may be calculated, where $a(i)$ is the mean distance of sample i to other samples within the same cluster and $b(i)$ is the mean distance of sample i to other samples in the closest cluster [113].

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{19}$$

It can therefore be seen that the greater the numerator the better the clustering process [113], with the range of possible values lying between -1 and 1. To better understand this Figure 2-5 has been provided, where the Silhouette Width for each sample can be seen, as well as the average Silhouette Widths for each cluster and the overall clustering process [114].

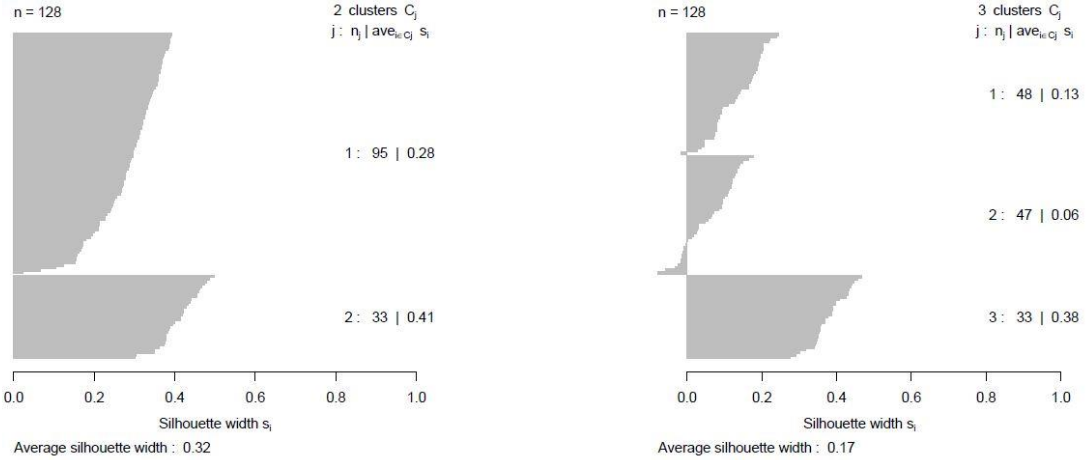

Figure 2-5: Silhouette Width Comparison [114]

Davies-Bouldin Index [115], [116]

The Davies-Bouldin Index (DB) was developed in 1979 by Davies and Bouldin to assist in determining the optimal number of clusters. Equation 20 shows the first step in calculating DB where X_j represents the individual vectors, A_i is the centre of cluster i and T_i is the number of vectors belonging to the cluster. S_i measures the dispersion within a cluster and if q is chosen to be 1 then S_i is the mean of the Euclidean distances of vectors from their cluster centre. Alternatively, if q is chosen to be 2, then S_i is the standard deviation of the distances of the vectors from their cluster centroid.

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} \left| X_j - A_i \right|^q \right)^{1/q} \tag{20}$$

The second value, known as the Minkowski metric (M_{ij}), is simply a distance measurement between centroids of two clusters (i and j). Equation 21 shows this measurement where each cluster centroid has N dimensions and a_{ki} simply represents the kth component of the centroid for cluster i. By assuming the value of p to be 2, M_{ij} becomes defined as the Euclidian distance between the centres of clusters i and j.

$$M_{ij} = \left(\sum_{k=1}^{N} \left| a_{ki} - a_{kj} \right|^p \right)^{\frac{1}{p}} \tag{21}$$

Using Equations 20 and 21 it is possible to derive a measure of the effectiveness of the clustering for cluster i (defined in Equation 22).

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{22}$$

This is then maximized (Equation 23) and the mean of all these values is determined – shown by $\bar{R}$ which is the Davies-Bouldin Index (Equation 24).

$$R_i = \max\left(R_{ij}\right) \tag{23}$$

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{24}$$

By evaluating the above steps it can be seen that optimum clustering will occur when $\bar{R}$ is minimised.

Dunn's Index [117], [118]

A variety of methods may be employed for determining the Dunn's Index for a set of clusters. However, the one provided here is one of the simpler methods. For defining the diameter (Δ) of cluster S the maximum distance between the vectors from within cluster S may be used, where $x, y \in S$ (Equation 25).

$$\Delta(S) = max\{d(x, y)\} \tag{25}$$

To determine the distance between two clusters, defined by $\delta(S, T)$, one may take the minimum distance between two vectors from different clusters where $x \in S$ and $y \in T$ (Equation 26).

$$\delta(S, T) = min\{d(x, y)\} \tag{26}$$

Dunn's Index (for c clusters) can then be calculated using Equation 27.

$$DI = \min_{1 \le i \le c}\left\{\min_{1 \le j \le c, j \ne i}\left\{\frac{\delta(X_i, X_j)}{\max\limits_{1 \le k \le c}\{\Delta(X_k)\}}\right\}\right\} \tag{27}$$

Care must be taken when interpreting the results from the Dunn's Index shown in Equations 25 to 27 because it may become distorted by outliers. To overcome this problem Bezdek and Pal [118] derived a general formula for Dunn's Index (Equation 28) with the following definitions: δ_i is any positive, semi-definite, symmetric set distance function and Δ_j is any positive, semi-definite diameter function.

$$DI = \min_{1 \le i \le c} \left\{ \min_{1 \le j \le c, j \ne i} \left\{ \frac{\delta_x(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta_y(X_k)\}} \right\} \right\} \tag{28}$$

Where $\delta_x(X_i, X_j)$ may be defined by one of five distance measures and $\Delta_y(X_k)$ is defined as one of three distance measures. Combining these variations creates 18 possible versions of Dunn's Index (refer to [118] for all the equations).

2.6 Neural Networks

The design of neural networks is based on the human brain and nervous system [119], [120]. The network consists of input and output layers, which are connected by synaptic weights. As data is fed into the network the synaptic weights change in order to better fit the data and the network adapts through this training method [119].

Figure 2-6 shows a basic schematic for a neural network model. The parallel nature of the neural network can be seen as well as the connections (synapses) between each layer within the network [121].

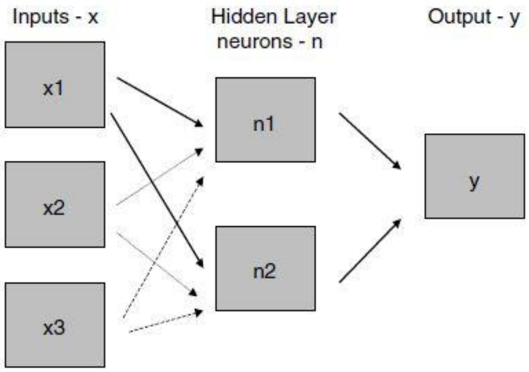

Figure 2-6: Basic Neural Network Layout [121]

2.6.1 Self-Organising Map

The SOM was created by Kohonen. This type of neural network utilises competitive learning, is trained in an unsupervised manner [122], [123] and is commonly used for data clustering.

The main reason for its use in data mining and clustering is its ability to compress high dimensional data to a low dimension [11], [12], [13].

Figure 2-7 shows the layout for a SOM, where the synaptic weights and 2D lattice can be seen.

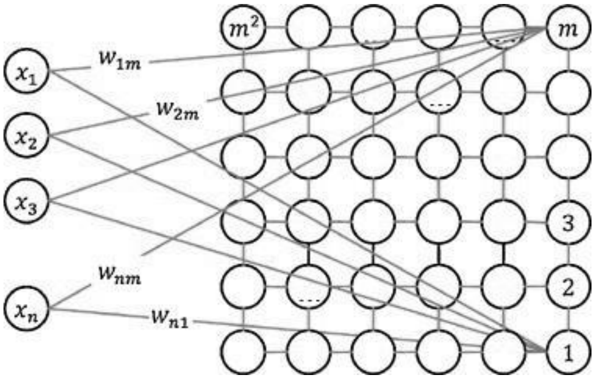

Figure 2-7: Basic Self Organising Map Schematic [122]

The model proposed by Kohonen is capable of managing an input and output space of different dimensions. This feature therefore enables the Kohonen model to reduce the dimensionality of the data, enabling data compression [13].

The learning algorithm completed by the SOM may be broken down into three distinct processes, listed below [13].

1. Competition: each neuron's weight is calculated and compared to the input and one neuron is chosen as a winner (based on the minimisation of the Euclidian distance between the neuron weight and the input vector [13] , [124]).
2. Cooperation: The neurons surrounding the winning neuron (neighbouring neurons) are also excited. The range (or distance) over which the neighbourhood encompasses is defined by the Gaussian function (Equation 29).

$$h_{j,i(x)} = \exp\left(-\frac{{d_{j,i}}^2}{2\sigma^2}\right), \qquad n = 0,1,2,\ldots \tag{29}$$

In Equation 29, i refers to the winning neuron, j is the excited neuron, $h_{j,i(x)}$ is the topological neighbourhood and $d_{j,i}$ is the lateral distance (Euclidean distance between the excited and winning neuron). The topological width (σ) decays exponentially with time, as shown in Equation 30. In this equation σ_0 is the topological width when the SOM is initiated, n is the number of iterations and τ_1 is simply a time constant. For two dimensional lattices σ_0 is set to the lattice radius and τ_1 as $1000/\log(\sigma_0)$.

$$\sigma = \sigma_0 exp\left(-\frac{n}{\tau_1}\right) \qquad n = 0,1,2,\dots \tag{30}$$

3. Synaptic Adaption: The excited neurons adapt their respective weights (w_j) in order to associate themselves more closely with the given input. Equation 31 shows how the synaptic weight of a neuron changes.

$$w_j(n+1) = w_j(n) + \eta(n)h_{j,i(x)}(n)\left(x - w_j(n)\right) \tag{31}$$

Where n refers to the time component and η defines the learning rate parameter and exponentially decays with time as shown in Equation 32.

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right), \qquad n = 0,1,2,\dots \tag{32}$$

The adaptive stage may be further divided into an ordering and convergence or tuning stage. The convergence stage occurs first and is responsible for the topographical mapping, whereas the tuning stage accounts for the fine tuning of the network [13]. Table 2-2 shows a comparison between these two stages.

Table 2-2: Comparison between Adaptive Stage Process Variables [13]

Recommended Variable	**Convergence Stage**	**Tuning Stage**
Initial learning rate η_0	~ 0.1	~ 0.01
Final learning rate η_n	> 0.01	$\sim 0.01 > \eta_n > 0$
Time constant τ_2	1000	$> 500 \times no.neurons$

Ultsch and Siemon [125] proposed the unified distance matrix (U-matrix) as a measure of the distances between neurons in the SOM. By using a colour scale to distinguish distances it is possible to visualise high dimensional data [126]. Wu et al. describes the visualisation of clusters, using the U-matrix, whereby when distances are small, clusters are present and are separated by regions of large inter neuron distances [123].

As it has already been mentioned the neurons in the SOM form a 2D lattice, however D. Wijayasekara et al. [127] proposed the use of a 3D output space, rather than the more common 2D mapping. By using a 3D output there should be better information preservation when the data is compressed to an output dimension [127].

A two-stage clustering process has been investigated in previous studies [123], [128]. The process consists of initially using a SOM (with a large number of neurons) on the original data and then a second clustering process on the SOM results.

2.6.2 Neural Network Software

MATLAB® has a built in Neural Network Toolbox™ [129] which is capable of easily implementing a wide variety of neural networks with the use of a basic GUI. A SOM toolbox for MATLAB was also developed by Alhoniemi et al. and is free for download at [130]. This software has been used in SOM research [131], however it has not been updated since 2005 [130].

2.7 Previous Studies

The previous studies which have been investigated in this section are related to both data mining research as well as purely financial studies. Unfortunately there is limited literature on the use of SOMs for clustering firms and many lack a strong financial view on the clustering. The work by Jago [22] has been included to provide a South African perspective on EMH.

Nanda et al. [4] used K-means, SOM and fuzzy C-means algorithms to cluster 106 firms using data from 2007 – 2008. The proposed methodology was aimed at using the clustering algorithms to generate clusters which would then be optimised using the firm weightings to minimise the portfolio variance. The study used the return data over different time frames as well as the firm’s price/ earnings, price/ book value, price/ cash flow, earnings per share, enterprise value/ earnings before interest, tax, depreciation and amortisation and market capitalisation/ sales as inputs. The validity of the clustering was investigated using numerous validity measures. It was found that the SOM performed best with seven clusters and achieved overall Silhouette widths ranging from -0.1498 to +0.331. The Davies-Bouldin Index values ranged from 1.3156 to 1.8038, however the K-means clustering was able to achieve more compact clusters. It was concluded that the work could be beneficial to investors wishing to develop portfolios and that future work could benefit from analysing additional financial variables.

Using data from the China Shanghai 180, Wang [6] clustered firms in several stages. The clustering algorithm used was K-means and the clustering process involved three stages.

Initially five clusters were generated using input vectors of five dimensions (primary earnings per share, net asset value per share, total assets turnover ratio, principal business growth rate and liquidity ratio). In addition to this, five clusters were also generated using price/ earnings and turnover as input variables. The firms which then occurred in the same clusters in both clustering tests were grouped together into one of eight clusters. The stocks were then investigated using technical analysis and it was found that the clustering results were beneficial for the purpose of stock selection. Although it was not mentioned in this study the cluster sizes which were generated were not necessarily even. In fact the cluster sizes achieved using price/ earnings and turnover were 138, 26, 5, 2 and 1. The study focused more on the clustering aspect of generating the portfolios and not the financial. As a result the financial impact of the small clusters was not looked at.

Sian and Kelvin [7] used the financial ratios of 470 stocks from S&P 500. The data was taken from 2001 and the financial ratios were the current ratio, debt/ equity, dividend yield, earnings before interest and tax growth, net income growth, price/ book value, price/ cash flow, price/ earnings, price/ sales, return on assets, return on equity and sales growth. Although 470 stocks were used the SOM dimensions were 12x9 and 24x18. This form of cluster formation required that the U-matrix be analysed and groups of neurons with low neighbourhood distances were regarded as clusters. Using this methodology however only resulted in 98 stocks being clustered with the smaller SOM and 278 stocks being clustered with the large SOM. In addition to using the U-matrix, Sian and Kelvin [7] visually analysed the resultant weights for each of the input planes. The analysis of the weight planes revealed that outliers within the data were problematic and resulted in the colour scales becoming distorted by these outliers. This rendered the visual analysis of the SOM input weight planes ineffective because the majority of the weights would fall into the same colour range, except for a few outliers. When the outliers were removed it was found that the SOM still had difficulties and these were attributed to the input vectors becoming so similar. This made it difficult for each neuron to become unique and led to an increase in quantisation error. The SOM did prove to be successful at separating the companies into clusters predominantly comprised of similar industries or with similar financial inputs.

Although not related to finance, the study by Aguado et al. [9] provides insight into the benefits associated with using the SOM for clustering. The data consisted of 328 samples, each containing 11 variables related to waste water treatment. The component planes of the SOM network were analysed in this study to investigate whether the input variables were related and the U-matrix used to for generating clusters. This was done by investigating the

U-matrix and analysing where similar neurons have been grouped. This was then combined with results from the Davies-Bouldin Index to determine an ideal number of clusters.

Numerous financial studies have been focused on developing portfolios based on financial ratios and the work by Fama and French [39] focused on using size, market beta, book-to-market value and earnings/ price. The study was completed over the 1963 to 1990 period and the different year ends of the companies were found to have little impact on the results. The portfolios were generated using the individual financial ratios and separating the firms into deciles, thus producing 10 portfolios. The financial data was calculated for the previous year (t-1) and the returns were determined from July to June of years t and t+1 respectively. In some of the tests the top and bottom deciles were split again resulting in 12 portfolios being generated. In addition to this, portfolios were generated by using two steps, firstly by taking size deciles and secondly taking beta deciles, thus making 100 portfolios. In this work it was concluded that market beta does not suitably explain market returns and for the period from 1963 to 1990 the market capitalisation and book-to-market values can be used to explain returns. In addition to this it was found that these variables account for the information contained within the earnings/ price and leverage as well.

Jago [22] completed a study using JSE financial data from 1990 to 2008. Portfolios were generated using both single variables as well as multiple variables. Of particular interest to this dissertation is the methodology used for generating portfolios with only one variable. The variables used were market capitalisation, book-to-market value, earnings/ price, cash flow/ price and dividends/ price. The portfolios were then generated by dividing the firms into two groups, with one representing the top 50% and the other representing the bottom 50%.

3. OBJECTIVES

The objectives for this research are aimed at evaluating the SOM clustering process as well as various financial variables.

1. Use the SOM and data mining techniques for the purpose of clustering companies from the JSE.
2. Evaluate the individual variables used for clustering and determine which are appropriate for clustering algorithms.
3. Investigate the variables from a financial perspective to determine whether clusters of companies with different performance can be obtained.
4. Use the results from the SOM clustering to analyse the validity of EMH for the JSE.

4. DATA

The data for this research has been taken from the McGregor BFA Database, a South African financial database. The data consisted of both the standardised financial ratios from McGregor BFA and the related price data. In the following sections the choice and frequency of the variables used, as well as the time frame have been discussed.

4.1 Financial Variables

One of the most important aspects of the data is the choice of financial variables to use as inputs into the clustering algorithm. When selecting variables it is important to consider their availability (on the McGregor BFA Database) as well as their predictive ability. In order to evaluate all appropriate aspects of a firm's performance a variety of financial variables have been chosen.

The literature reveals that financial studies place significant emphasis on the financial variables while clustering studies have not investigated the relationship between the variables and the clustering process. For this reason the selected variables have been chosen to incorporate both previous financial and clustering studies. The number of variables has been kept to a minimum in order to reduce redundancy.

Additional consideration was given to the variety of firms to be analysed which makes the use of some variables, such as sales figures, irrelevant to some industries. Sian and Kelvin [7] used price/ sales along with other variables and the clusters obtained consisted of companies from similar industries. Rather than only choosing ratios, which would transparently cluster stocks based on their industry, the purpose of this data mining research is to examine underlying factors.

Dividend yield would have been included in the analysis but many companies did not have dividend yield data. This would limit the companies used for clustering and it was decided to rather exclude this variable. This lack of dividend yield data is also apparent in the study completed by Jago [22] which looked at the JSE.

Table 4-1 provides a brief reason for the choice of each financial variable as well as a code which will be used to refer to each variable for the remainder of this report.

Table 4-1: Financial Variables for Clustering Input

Variable	Code	Description
Debt/ Equity	DE	Provides insight into corporate structure, which can be beneficial or induce excessive risk [89]. Could also provide interesting insights into its relationship with other variables, especially firm size. Was also considered in work by Fama and French [39].
Price/ Earnings	PE	Often considered a leading indicator for future returns and has significant amounts of related information. Traditionally low price/ earnings ratios have been used to define value stocks which have historically outperformed their growth counterparts [43], [72], [73], [74]. Also, since this variable is commonly used by investors it is less likely to possess outliers which affect the clustering process. Used in various studies [4], [6], [7] for clustering stocks as well as by Jago in a financial study [22].
Price/ Book Value	PB	Should be able to separate growth from value stocks [43]. Has been considered a proxy for risk, which is not accounted for with volatility [82]. Also should provide insight into the relationships between the various price ratios. The Fama and French Three Factor model also found that the book-to-market value of stocks was successful in explaining stock returns [40] and this ratio would underline the same features.
Price/ Cash Flow	PC	Not as thoroughly researched as other price ratios, however cash is more difficult to manipulate and could provide a good proxy for risk and underlying behaviour. Could provide a good comparison with other price ratios, in particular the price/ earnings ratio. Has been used in previous studies related to both clustering [4], [7] and pure financial analysis [22].

Quick Ratio	QR	Measures a firm’s ability to pay short term (current) debts, while not including the inventories since they are not necessarily liquid [14]. This ratio is likely to be highly correlated to the current ratio so it was decided to take only this ratio because it is stricter when evaluating liquidity. Should therefore also provide a measure of possible risk and may be related to leverage. Liquidity was used in the clustering completed by Wang [6]. Sian and Kelvin [7] used the current ratio however the quick ratio offers a stricter measure of liquidity.
Return on Assets	RA	Used by Sian and Kelvin [7] for clustering with a SOM. May provide an insight into how specific industries rely on assets. Can be compared to the return on equity.
Return on Equity	RE	Sian and Kelvin [7] used the return on equity for SOM clustering. Often used by investors and can therefore be regarded as a ratio which managers view as significant. Due to its popularity it is likely that this ratio will not contain many outliers, making it suitable to clustering.
Market Capitalisation	MC	Provides the ability to distinguish between small and large firms and may also provide insight into growth. Smaller firms are sometimes regarding as being information inefficient [52], [55]. Could provide insight into whether financial performance (measured with financial ratios) is related to firm size. Used in various financial studies for the purpose of generating financial portfolios [22].
Volatility	V	A different risk measure, to any of the above listed ratios. Could provide additional insight into share performance, which is not apparent from the traditional financial ratios. The only ratio which solely looks at price movements and since it is dependent on share prices it will be readily available.

4.2 Data Range

In order to select the time frame for the analysis it was necessary to consider the following:

1. There must be sufficient data to analyse a wide range of firms so as to not exclude companies which may be required for various financial relationships.
2. The data must be taken over a sufficient time frame to enable several years of historic data (used for the clustering) as well as several subsequent years of additional data for the purpose of cluster return analysis.

Considering the above mentioned criteria it was decided that the 10 years prior to this study (2002 – 2012) would provide the most comprehensive data over an adequate time frame. By separating the 10 years into two sections it is possible to use five years (2002-2006) for the inputs and the remaining five years (July 2007- June 2012) for the analysis of the performance of the results. The final period started in July to allow a six month gap to avoid look-ahead-bias.

4.3 Data Frequency

The next component of the data collection involves the frequency of the data. The data is available over a daily, weekly, monthly, quarterly and annual basis. Although some financial variables (such as the P/E ratio and market capitalisation) are available on a daily basis, not all of the variables are available with such frequency. By examining the McGregor BFA database it was found that annual year end results were the most abundant and would enable the largest inclusion of different firms. In addition to this it was decided that ratios, such as P/E, when taken on a daily basis are predominantly controlled by the share price. The inclusion of several variables on a daily basis would therefore measure the price movements [76], rather than underlying fundamental aspects of firm performance. Previous research by Fama and French [39] and Jago [22] only used a single annual value for the financial variables. These annual values were used to determine the clusters and therefore it should not be necessary to include data with higher dimensionality. For these reasons only the annual data has been used as an input into the SOM.

4.4 Data Collection

The data was downloaded from the McGregor BFA Database in Excel sheets and processed in MATLAB using object oriented programming (OOP). This reduced the possibility of errors, before processing the data back into an Excel sheet for inspection. Figure 4-1 shows the simplified data collection process – this has been broken into 8 steps which have been discussed in more detail.

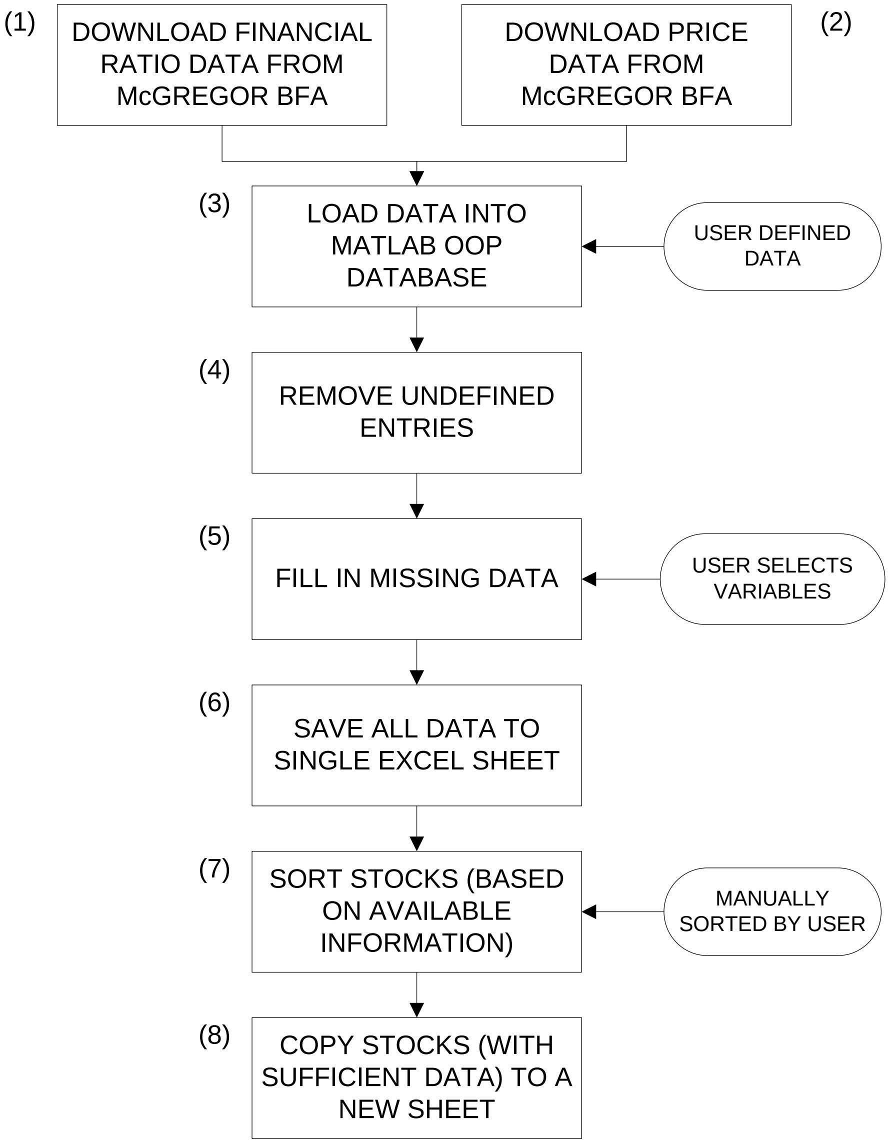
(1)
DOWNLOAD FINANCIAL RATIO DATA FROM McGREGOR BFA
(2)
DOWNLOAD PRICE DATA FROM McGREGOR BFA
(3)
LOAD DATA INTO MATLAB OOP DATABASE
USER DEFINED DATA
(4)
REMOVE UNDEFINED ENTRIES
(5)
FILL IN MISSING DATA
USER SELECTS VARIABLES
(6)
SAVE ALL DATA TO SINGLE EXCEL SHEET
(7)
SORT STOCKS (BASED ON AVAILABLE INFORMATION)
MANUALLY SORTED BY USER
(8)
COPY STOCKS (WITH SUFFICIENT DATA) TO A NEW SHEET

Figure 4-1: Data Collection (Illustrative)

When retrieving the information from the McGregor BFA database it was necessary to download the annual standardised financial ratios (step 1) for each firm and then download the price information separately (step 2). Within the price data file was the market capitalisation, dividend information and volatility.

The majority of this data manipulation was completed by developing an object oriented programing (OOP) system, whereby each firm was a new instance of an object (step 3). The financial ratios were separated into general, cash and growth ratios (as defined by the McGregor BFA Database) and each section was defined as a new property for that instance. This was done since all the financial ratio information had been downloaded and stored as separate Excel sheets. Table 4-2 lists all the properties of the OOP database, along with a description of where the data came from. For clarity the fundamental ratio Excel sheets refer to step 1 in Figure 4-1 and the price information Excel sheets refer to step 2 in Figure 4-1.

Table 4-2: Object Oriented Instance Properties

Property	Type	Description
Name	String	The name of the firm from the fundamental ratio Excel Sheets.
Ticker	String	The ticker which was obtained from the name of the price information Excel sheets.
Year	Array	The years of interest as input by the user.
Raw ID	Double	A number assigned according to when the firm was originally added to the database (order comes from the fundamental ratio Excel Sheets).
Month	String	The month extracted from the firm heading line in the fundamental financial ratio Excel sheets
Cash	Cell Array	An array containing the chosen annual cash ratios extracted from the cash fundamental ratio Excel sheet.
General	Cell Array	An array containing the chosen annual general ratios extracted from the general fundamental ratio Excel sheet.
Growth	Cell Array	An array containing the annual growth ratios extracted from the growth fundamental ratio Excel sheet.
Price	Cell Array	The selected columns from the price Excel sheets. This required the opening, closing and date columns for later calculations.
Dividend	Cell Array	The selected columns from the price Excel sheets. This required the LDR date column and the amount.

Market Cap	Cell Array	The selected columns from the price Excel sheets. This required the date and market capitalisation columns.
Volatility	Cell Array	The selected columns from the price Excel sheets. This required the date and volatility columns.
Technical	Cell Array	This is used for storing the chosen data from the price, dividend, market capitalisation and volatility properties in annual form.

Step 4 in Figure 4-1 refers to the removal of data points which were not available (presented as N/A in the excel sheets downloaded from BFA McGregor). Missing values were then replaced using linear interpolation and the last known value method described by Silva and Marques [8]. Companies with missing values for three or more input years were excluded and only data from 2000-2006 were used for interpolation. The MC was assumed to be zero for years the companies were not listed, hence it did not benefit from the last known value method. The MC and V were taken for the month end corresponding to the other financial ratios.

Previous financial studies have excluded companies from the finance sector (Fama and French [39]) and others have excluded companies with negative price ratios (Gaffney [132]). For this study it was decided to include all companies listed as of July 2007 with three years input values. The listing date of July 2007 is required for the financial analysis which begins six months after the final input year (2006). This six month gap has been included to avoid look-ahead-bias. Companies delisted after July 2007 have been included to prevent survivorship bias. In order to complete all the tests with the same set of companies those without sufficient data were excluded.

The differences between the market capitalisations for the JSE and the database were then compared. In Figure 4-2 it can be seen that the sample taken for the research adequately represents the JSE.

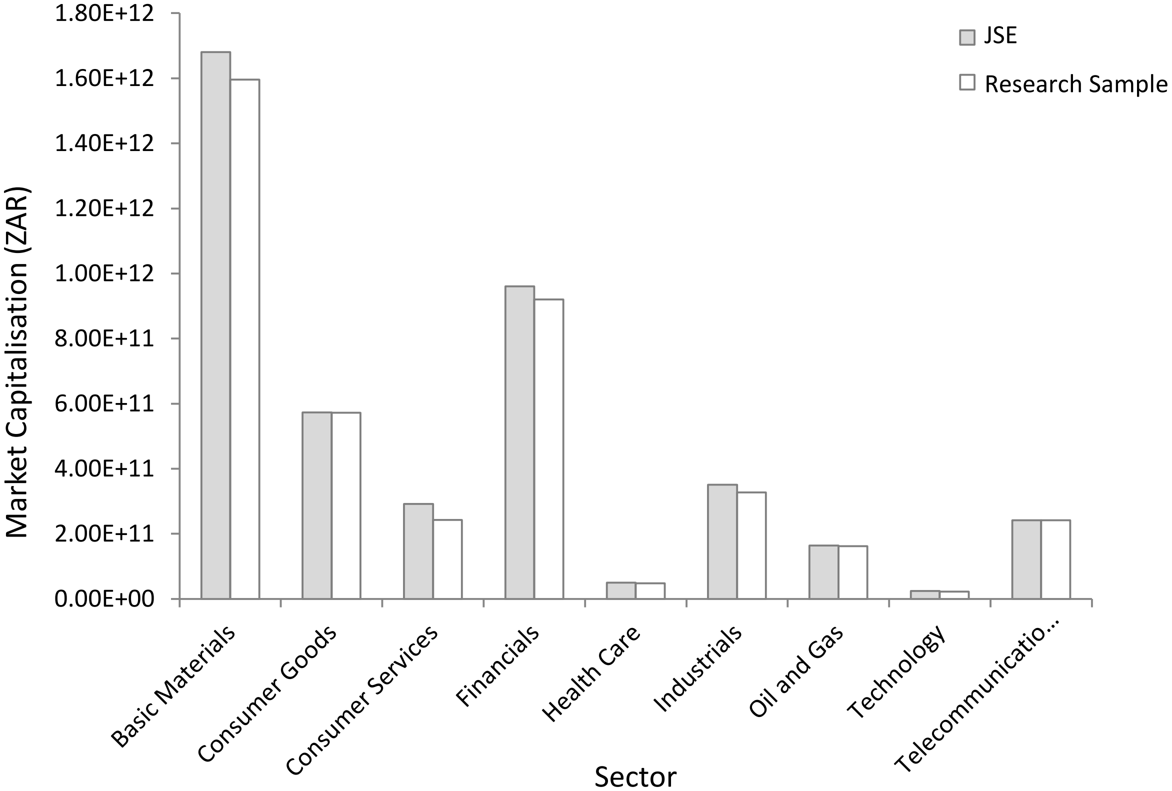

Figure 4-2: Database Composition

5. METHODOLOGY

To complete research into the validity of SOMs for stock clustering it is necessary to evaluate various parameters in the clustering process. Two parameters which are seen to play a significant role in the clustering process are the size of the clusters and the choice of financial variables. For clarity a brief explanation of the methodology chosen as well as the notation used for the remainder of the report will first be discussed in this section.

The first method of clustering (Section 5.2) used single values for all the financial ratios. In comparison to the study completed by Nanda et al. [4] the research presented here was aimed at the inclusion of many more companies. The inclusion of a greater number of companies was done so as to not limit the scope of the research, as mentioned by Moor and Sercu [58]. Moor and Sercu found that some relationships, such as small firm effect, were only apparent in the smallest of companies. Therefore by only taking the larger companies it is possible that some relationships may not become apparent. This inclusion of more companies did introduce more outliers which will have an impact on the clustering. The work by Wang [6] simply looked at the success rate of the clustering and it did not analyse the validity with tools such as the Davies-Bouldin Index and it is therefore difficult to interpret the validity of the clustering.

The second method of clustering employed in this study was aimed at providing insight into the individual variables. Previous work [4], [6] has focused on clustering firms with several financial variables but it is not apparent which individual variables are having an effect on the clustering process. This is in part due to the lack of component analysis. By taking the single variable approach applied in financial studies [39], [22], [133] it will provide more insight into how each variable may be a proxy for stock returns. The methodology specifically related to these tests is shown in Section 5.3 where it has been explained in more detail.

Section 5.1 refers to the general clustering process, which was applied to the clustering methodologies (Sections 5.2 and 5.3). This provides insight into the procedures followed by the algorithms designed to automate the clustering process for data mining.

5.1 Clustering Process

This section refers specifically to the clustering which was applied to Sections 5.2 and 5.3. The general process of clustering is explained in this section before expanding into more detail in the remaining sections.

The methodology (shown in Figure 5-1) is aimed at analysing how various parameters affect the clustering process. The remainder of this section will refer to Figure 5-1 as the description of the methodology is expanded on in more detail.

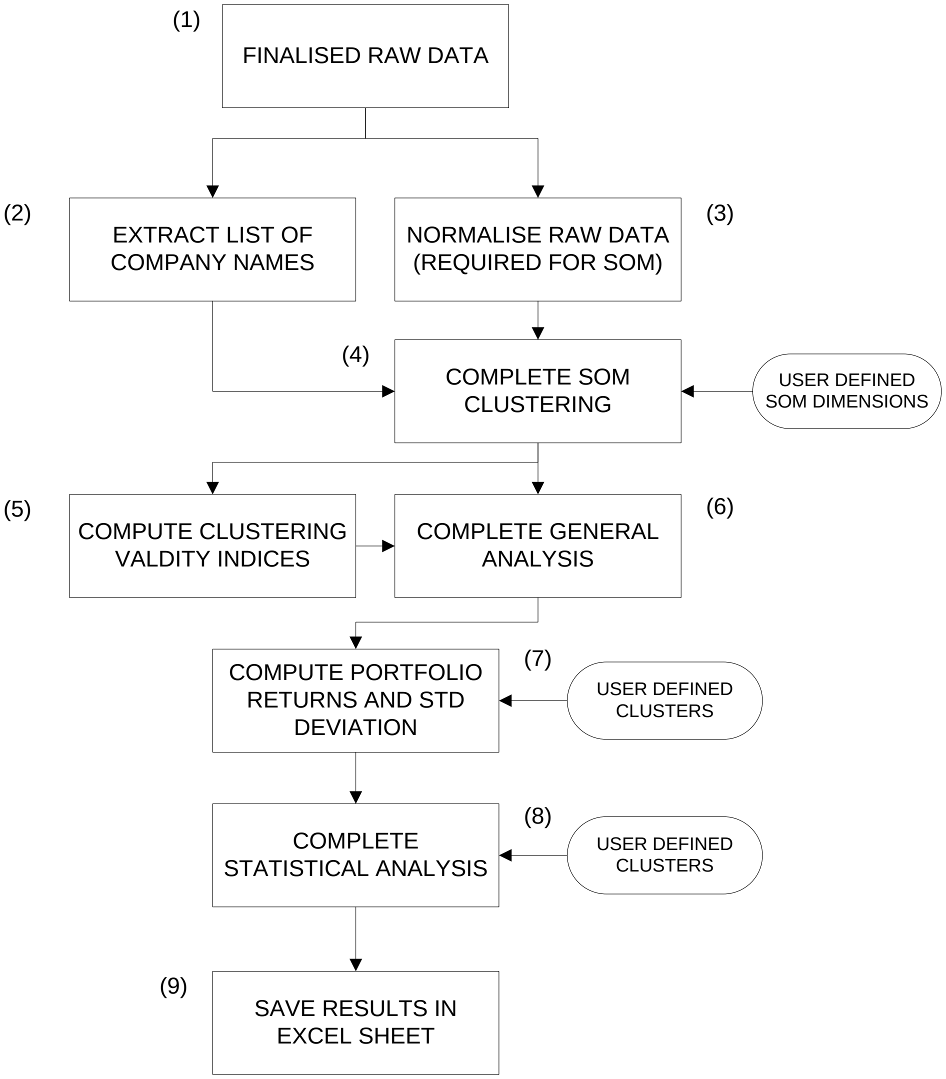
(1)
FINALISED RAW DATA
(2)
EXTRACT LIST OF COMPANY NAMES
(3)
NORMALISE RAW DATA (REQUIRED FOR SOM)
(4)
COMPLETE SOM CLUSTERING
USER DEFINED SOM DIMENSIONS
(5)
COMPUTE CLUSTERING VALDITY INDICES
(6)
COMPLETE GENERAL ANALYSIS
(7)
COMPUTE PORTFOLIO RETURNS AND STD DEVIATION
USER DEFINED CLUSTERS
(8)
COMPLETE STATISTICAL ANALYSIS
USER DEFINED CLUSTERS
(9)
SAVE RESULTS IN EXCEL SHEET

Figure 5-1: Clustering Methodology

5.1.1 Data Pre-processing

The data pre-processing refers to the preparation of the data for use in the SOM. In particular it deals with stages (1-3) in Figure 5-1. The names of all the companies were extracted from the raw data file and the financial data was normalised.

The presence of outliers (which greatly impacted on the clustering regardless of the clustering algorithm) was caused by the inclusion of small companies and the time frame chosen. Sian and Kelvin [7] found the presence of outliers problematic and that the use of normalisation increased the quantisation error because the input vectors became too similar. This work, as well as other studies, revealed that the two most likely causes for the poor clustering are:

1. The presence of outliers within the data which distorted the input vectors and was not overcome in the normalisation process.
2. The limited number of input vectors, which makes it difficult to generate sufficiently large clusters.

To overcome this problem an algorithm was developed based on Winsorising. Winsorising refers to the process of replacing outlying values with a new value considered to not be an outlier [134]. Winsorising can be easily applied by using percentiles to define the limits of the outlying range. This requires the analyst to determine a suitable cut-off region but for this research it was decided to use two criteria as shown in the algorithm below.

```
If (variable > Q3+3xIQR && variable > 95th percentile)
        variable = max{Q3+3xIQR, 95th percentile}
elseif (variable < Q1-3xIQR && variable < 5th percentile)
        variable = min{Q1-3xIQR ,5th percentile}
end
```

The choice of the upper and lower quartiles (Q1 and Q3) with three times the inter-quartile range (IQR) is based on the definition for extreme outliers ([135] , [136]). This criterion was combined with the percentile limits to create a system which was less likely to suffer from any bias. This algorithm was not applied to MC and V. MC and V were transformed using $ln(1 + variable)$ due to the log normal distribution present in the market capitalisation and volatility data. The study presented here had several zero values for MC and V and therefore

$ln(1 + variable)$ would result in zero values after the log transformation without distorting the original values too much. Once the outlying values had been replaced the data was linearly normalised between -1 and +1 using min-max normalisation.

5.1.2 SOM Clustering

The SOM clustering (stage 4 in Figure 5-1) was considered one of the simpler steps in the methodology process since most of the required algorithms are built into the MATLAB environment. It should be noted than when using the *nctool* method in MATLAB it is only possible to generate square SOMs; however by taking the sample code generated after this process it is possible to define each SOM dimension independently. This was required to generate SOM networks which were not square, as shown in Table 5-4 in Section 5.2.

5.1.3 Clustering Validity Indices

The clustering validity stage occurred after the clustering had been completed and was not dependent on the financial results because it deals purely with the input variables. Although the financial analysis should provide more evidence regarding the effectiveness of the clustering process (in terms of portfolios and returns) the clustering validity is intended to provide information regarding the use of the input variables.

For the purpose of measuring the validity of the clustering process the four measures (mentioned in Section 2.5.4) were chosen. The Davies-Bouldin Index is commonly employed in cluster analysis and is regarded as an adequate measure of clustering performance. It is most commonly used to determine the optimal number of clusters and for this study it should provide insight into how the SOM dimensions affect the clustering. DB values related to single firm clusters have not been considered in the overall DB index.

The second measure, Silhouette Width (SW) is incorporated within the MATLAB environment, therefore making its application relatively simple. The overall SW has been determined using the cluster sizes to determine the weighted average. In addition to this it has been used to investigate whether individual shares were clustered appropriately.

The third and fourth clustering validity indices refer to Dunn's Index (DI). This measure has been used in previous studies and it was decided to use both traditional DI as well as the alternative DI. The difference between the two methods has been discussed in Section 2.5.5. The MATLAB algorithms for these two metrics come from the algorithms developed by Ramosall [143]. The alternative DI was expected to produce less biased results with DI being more influenced by outliers. For this work the Alternative Dunn's Index with the following metrics for distance and diameter between two clusters (S and T) were used [118]. As with the equations described in Section 2.5.5 the distance between two clusters is defined by $\delta(S,T)$ and the diameter of a cluster is defined by $\Delta(S)$. For Equations 34 and 35 $|S|$ and $|T|$ refer to the size of clusters S and T.

$$\delta(S,T) = \delta_{avg}(S,T) = \frac{1}{|S||T|}\sum_{x\in S, y\in T} d(x,y) \tag{34}$$

$$\Delta(S) = 2\left(\frac{\sum_{x\in S} d(x,v_s)}{|S|}\right), \text{where } v_s = \frac{1}{|S|}\sum_{x\in S} x \tag{35}$$

All these measures were calculated using the normalised input values. Euclidian distance has been used to be consistent with the SOM distance metrics. Refer to Appendix B Sections B-1 to B-3 for additional references and sample calculations for the Davies-Bouldin Index, Silhouette Width and Dunn’s Index.

5.1.4 General and Financial Analysis

The general and financial analysis has been completed in iterative steps because it requires user inputs. The general analysis in stage 6 in Figure 5-1 refers to a brief investigation of the clusters formed. In particular this stage analyses the industry composition, cluster size, validity and number of delisted companies in each cluster. The validity of the individual clusters has been measured using Silhouette Width as discussed in Section 5.1.3. Using this information, only clusters suitable for financial analysis are investigated in more detail (stages 7 and 8, Figure 5-1).

The financial analysis of the portfolios refers to stage 7 in Figure 5-1. This step has been broken down into two stages, the computation of the individual share returns and the more involved portfolio analysis.

During the clustering process it was found that the clusters produced were varied in size, with several small clusters. When analysing very small clusters it is likely that differences in returns would become apparent simply due to the random nature of individual firm performance. It was therefore decided that a cut-off cluster size would need to be determined prior to financial analysis in order to achieve meaningful results. By making the cut-off cluster size too large many companies would be excluded, thus increasing the likelihood of overlooking relationships in the data. If however the cut-off is set too low the results become less reliable. In the work completed by Nanda et al. [4] the financial performance of clusters with only five companies was analysed. In comparison the double variable analysis by Jago

[22] generated four portfolios, with the smallest single year portfolio containing 21 companies. In addition to this, the work completed by Jago [22] was completed over a much longer time frame to prevent biased results. The main difference between these two studies is that Nanda et al. [4] was primarily focused on clustering whereas Jago [22] completed a purely financial study. Since the work presented here has a greater emphasis on financial performance than Nanda et al. [4] it was decided that the inclusion of clusters as small as five companies would create a bias. This bias would be generated because such small clusters would inherently have different performance. So, by taking larger portfolios the possibility of having false findings is minimized. Following a methodology more similar to Jago [22], it was decided to make the cluster cut-off 20 companies. To further understand the impact of the cut-off on the data collected for this study Table 5-1 has been provided. Each of the cells in Table 5-1 show the percentage of companies included in the analysis for the various tests presented in Section 6 (for more details with regards to the size of the SOMs used refer to the relevant tests in Section 6 where the SOM sizes have been discussed). Making the cut-off much greater results in a large portion of the companies being excluded and in order to include more companies it would be necessary to decrease the cut-off to 15 companies per cluster. Considering the significance placed on having large clusters by Jago [22] it was decided to rather keep the cut-off at 20 companies, thereby making the results achieved more conservative.

Table 5-1: Minimum Cluster Size Data Loss Analysis

Cut-off	Clustering Test (% companies included)										
	DE	M1	M2	MC	PB	PC	PE	QR	RA	RE	V
1	100	100	100	100	100	100	100	100	100	100	100
2	100	100	100	100	100	100	100	100	100	100	100
3	99	100	100	100	100	99	100	100	100	100	100
4	98	100	100	99	99	99	99	100	98	99	99
5	96	100	100	99	99	98	99	98	97	97	99
6	95	100	100	99	95	98	97	96	93	95	99
7	95	100	100	99	90	93	95	94	93	95	96
8	95	100	100	99	85	93	95	91	93	95	96
9	91	100	100	99	85	93	91	91	90	92	93
10	88	100	100	99	85	93	91	91	90	92	93
11	84	100	100	99	85	85	84	91	90	84	93
12	80	100	96	99	85	81	84	78	85	80	93
13	80	100	96	99	85	81	84	78	85	80	93
14	75	100	96	99	80	76	73	73	85	70	93
15	75	95	96	99	80	76	73	73	85	70	93

16	75	89	96	99	80	76	73	73	85	64	93
17	75	89	96	99	80	76	73	73	85	64	93
18	75	89	89	99	80	76	73	73	85	64	93
19	75	89	89	99	80	76	66	73	78	64	93
20	75	81	89	99	80	76	66	73	78	64	93
21	75	73	81	99	80	76	66	73	78	64	93
22	75	73	81	99	80	76	66	73	78	64	85
23	75	73	81	99	80	76	66	73	78	64	85
24	75	73	81	99	80	76	66	73	78	64	85
25	75	64	81	99	80	76	57	73	78	64	85
26	75	64	81	99	80	76	57	73	78	64	75
27	75	64	71	99	80	65	57	73	78	64	75
28	75	64	71	99	80	65	57	73	78	64	75
29	75	64	71	99	80	55	57	73	78	64	75
30	75	64	60	87	80	55	57	73	78	64	75
31	75	64	60	87	80	55	57	73	78	64	75
32	75	64	60	75	80	55	57	73	78	64	75
33	75	64	60	63	80	55	57	73	66	64	75
34	75	64	60	63	80	55	57	73	53	64	75
35	75	51	60	63	80	55	57	73	53	64	75
36	75	51	60	63	66	55	57	73	53	64	75
37	75	51	60	63	66	55	57	59	53	64	75
38	60	51	60	63	66	55	57	45	53	64	75
39	60	51	45	63	66	55	57	45	53	64	75
40	60	51	45	32	66	55	57	45	53	64	75
41	60	51	45	16	66	55	57	45	53	64	75
42	60	51	45	16	50	55	57	45	53	64	75
43	60	51	45	0	50	55	57	45	53	64	75
44	60	51	45	0	50	55	57	45	53	64	75
45	60	51	45	0	33	55	57	45	53	64	75
46	60	33	45	0	33	55	57	45	53	64	75
47	60	33	45	0	33	55	57	45	53	64	57
48	60	33	45	0	33	55	57	45	53	64	57
49	60	33	45	0	33	55	57	45	53	64	57
50	60	33	45	0	33	55	57	45	53	64	57

In addition to the cluster size cut-off it was necessary to consider the frequency of returns to be calculated. As it is common practice to use monthly returns ([22], [39]), and to not generate excessive data, it was decided to use monthly returns. The returns were calculated using the continuous compounding formula (Equation 2). To avoid look-ahead bias a 6 month period after January 2007 was allowed, as explained in Section 4.4. This resulted in the mean monthly returns being calculated from July year *t* to June year *t+1*. Dividends were included in the return calculations using the last day to trade as an associated date. In the event that a firm delisted in year *t* its returns were assumed to be zero after the delisting date. The firm was then removed from the cluster for year *t+1*. This is similar to the study by Jago [22] where delisted companies were given 0% return for that year, but this research only assumes 0% return after delisting and not for the whole of year *t*. For all portfolios transaction costs were not included.

So as to not overlook possible trends in the data, value weighted and equally weighted financial results were calculated. Value weighted results were calculated using each companies' market capitalisation from the last trading day before July year *t*. These values were assumed unbiased because market capitalisation data is readily available on a daily basis. In comparison the equally weighted results were calculated giving equal weighting to all the companies.

It was found that the time frame chosen for this study suffered from poor overall market performance. Rather than using the R157 as a benchmark it was decided to compare the monthly returns to the JSE All Share Index. The portfolio standard deviation was calculated using the monthly returns for year *t* and Equation 21 from Section 2.2.2. Appendix B contains the sample calculations related to this financial analysis.

Stage 8 in Figure 5-1 refers to the statistical analysis of the selected clusters. For this analysis it was necessary to use parametric and non-parametric tests. These statistical tests were included to evaluate the hypothesis that the clustering could produce clusters of companies with different financial performance. The parametric test used was one-way ANOVA and the non-parametric test used was one-way Kruskal-Wallis. The non-parametric test was included because several of the tests presented non-normal distributions as well as a limited sample size ($n < 30$). This inclusion of non-parametric tests is not unique and a similar method was used by Patel [137].

When completing the statistical analysis the monthly portfolio returns were used over three time frames (one year, three years and five years). Although the financial results were calculated and presented on an individual yearly basis it was decided to rather complete the statistical analysis over different time frames. This has been done to increase the sample sizes and improve the quality of the statistical analysis – i.e. the financial results for year three would only be for the 12 months in year three, whereas the statistical results for year three would run from year one to year three and would contain 36 months. The reason for using the portfolio returns as a measure of the variance in the returns (rather than the individual firm returns) was to keep in-line with the methodology introduced by Graham and Uliana [138] as well as deFusco et al. [20]. Each portfolio was then generated and the monthly returns taken. The returns were then compared using SPSS for the statistical analysis.

5.2 Multiple Variable Clustering

The first clustering method used all the financial variables. With this approach to clustering, the 2006 values for each financial variable were used as inputs into the SOM. This generated a system whereby each input vector (firm) had nine dimensions (financial variables). This approach was chosen because it accounts for all financial aspects of a firm's performance. Table 5-2 shows an example of how four input vectors (companies A to D) would each have nine dimensions (financial variables).

Table 5-2: Multiple Variable Clustering Input Example 1

Firm	DE	PB	PC	PE	QR	RA	RE	MC	V
A	var1	var2	var3	var4	var5	var6	var7	var8	var9
B	var1	var2	var3	var4	var5	var6	var7	var8	var9
C	var1	var2	var3	var4	var5	var6	var7	var8	var9
D	var1	var2	var3	var4	var5	var6	var7	var8	var9

The second multiple variable clustering test was completed using five financial variables (MC, PB, PC, PE and V) as shown in Table 5-3. These five financial variables were expected to yield the most significant results based on how often they have been investigated in financial studies and were therefore grouped together.

Table 5-3: Multiple Variable Clustering Input Example 2

Firm	PB	PC	PE	MC	V
A	var1	var2	var3	var4	var5
B	var1	var2	var3	var4	var5
C	var1	var2	var3	var4	var5
D	var1	var2	var3	var4	var5

For the remainder of the report it is important to note the notation which will be used to distinguish between these two tests. The first multiple variable test (nine variables) will be referred to as M1 and the second test (five variables) will be referred to as M2. In addition to this, a code which refers to the respective SOM dimension may be included when referring to a specific test. Table 5-4 lists the codes related to each SOM size, e.g. the first multiple variable test, with SOM size 3x3, would be referred to as M1-09.

In order to determine whether the size of the clusters has an effect on the generation of portfolios it was decided to complete numerous clustering iterations with different SOM dimensions. Table 5-4 lists the different iterations where it can be seen that SOM dimensions for prime numbers result in 1-dimensional clusters.

Table 5-4: SOM Dimensions

Number of Clusters	Code	SOM Dimensions
2	02	2x1
3	03	3x1
4	04	2x2
5	05	5x1
6	06	2x3
7	07	7x1
8	08	2x4
9	09	3x3
10	10	2x5
11	11	11x1
12	12	3x4
13	13	13x1
14	14	2x7
15	15	3x5

5.3 Single Variable Clustering

In order to evaluate a different method of clustering, the SOM algorithm was applied to each financial variable separately. To achieve this it was originally considered to take the financial inputs related to 2006. The use of a one dimensional input would not take full advantage of the SOM network and it was decided to rather use the previous five years financial results. This should enable the SOM to detect companies with irregular behaviour and be more beneficial than simply using 2006 values. Each test was completed using the five years of data for only one variable. In Table 5-5 it can be seen that four input vectors (companies A to D) would each have five dimensions, each related to a different year for the chosen financial variable.

Table 5-5: Single Variable Clustering Input Example 1

Firm	DE				
	2006	2005	2004	2003	2002
A	var1	var2	var3	var4	var5
B	var1	var2	var3	var4	var5
C	var1	var2	var3	var4	var5
D	var1	var2	var3	var4	var5

6. RESULTS

The results presented here have been separated into two sections. Section 6.1 looks at the clustering completed using multiple variables and Section 6.2 looks at the single variable clustering. For completeness a sample calculation has been included in Appendix B.

Although the naming convention has been discussed in Section 5 it has been explained in more detail in Table 6-1 because it has been used extensively in the results.

Table 6-1: Naming Convention for Clustering Test

Code	Description	Dimensions
M1	2006 values for all financial variables	9
M2	2006 values for five financial variables	5
DE	2006 – 2002 values for debt/ equity	5
PB	2006 – 2002 values for price/ book value	5
PC	2006 – 2002 values for price/ cash flow	5
PE	2006 – 2002 values for price/ earnings	5
QR	2006 – 2002 values for quick ratio	5
RA	2006 – 2002 values for return on assets	5
RE	2006 – 2002 values for return on equity	5
MC	2006 – 2002 values for market capitalisation	5
V	2006 – 2002 values for volatility	5

The first column (Code) in Table 6-1 refers to the clustering and when referring to a specific test, the code is followed by a SOM size. For clarity a few examples have also been given below and should provide sufficient understanding with regards to the naming convention.

- M2-08: multiple variable clustering with input values for 2006 for five variables completed using a 2x4 SOM network.
- PB-02: single variable clustering with input values for 2006–2002 for price/ book value completed using a 2x1 SOM network.

The results related to each test have been presented in the order listed below:

1. Clustering validity analysed with Davies-Bouldin (DB) Index, Silhouette Width (SW), Dunn's Index (DI) and Alternative Dunn's Index.
    - DB Index: lower values are more desirable
    - SW: values below 0.25 being regarded as poor [139]
    - Alternative DI & DI: Larger values indicate better clustering

    It should be noted that several validity indices were used because one alone would not necessarily yield accurate results. In fact the different validity measures can yield contradictory results due to them weighting outliers differently and using different measures of inter and intra cluster distances.
2. General results for a chosen SOM test which looks at neuron numbers, cluster sizes, individual Silhouette Widths, industry composition and delisted companies.
3. SOM plane weight diagrams for each input variable. Lighter colours indicate a greater input value. When referring to a neuron number in the SOM weight plane diagrams the numbering starts in the bottom left and moves across to the top right as shown in the sample below:

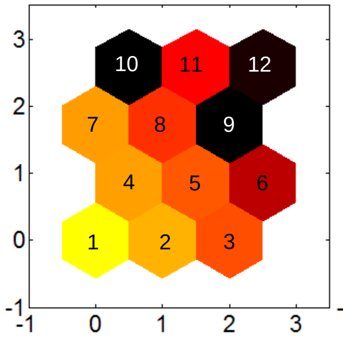

Figure 6-1: SOM Weight Plane Numbering Sample

4. Value weighted and equally weighted financial results of chosen clusters. Including comparison to the JSE All Share index and portfolio standard deviation.
5. Statistical results comparing returns for selected clusters using both parametric and non-parametric tests.

It is important to note that the returns presented are for the individual years (i.e. year 1 or 2 or 3 etc.) whereas the statistical results are presented over a time frame of years (i.e. year 1 or year 1 to year 3 etc.). For a full explanation regarding this refer to Section 5.1.4 where the statistical analysis has been discussed.

When investigating the specific clusters related to a SOM the terms cluster and neuron are often used interchangeably. A brief summary of the column headings is presented in Table 6-2.

Table 6-2: Results Notation

Heading	Variable	Description
NRN	Neuron number/ cluster number	Refers to the location of the neuron (cluster) in the SOM network.
CMP	Number of companies	Number of companies related to the specific neuron (cluster).
SW	Silhouette Width	Clustering validity measure related to the specific neuron (cluster).
Mean Monthly	Arithmetic mean portfolio monthly returns	Calculated using firm continuous compounding returns
Mean Excess	Arithmetic mean portfolio monthly excess returns	Difference between the portfolio monthly mean returns and monthly JSE Top 40 Index
Std Dev	Portfolio Standard Deviation	Portfolio standard deviation consistent with Markowitz portfolio theory

6.1 Multiple Variable Clustering

The clustering results, using multiple variables, can be broken down into two tests. The first uses all nine variables analysed in this study and the second test only uses the variables which were considered most significant. The cluster sizes achieved in these tests were relatively even over the range of cluster tests.

6.1.1 All Variables

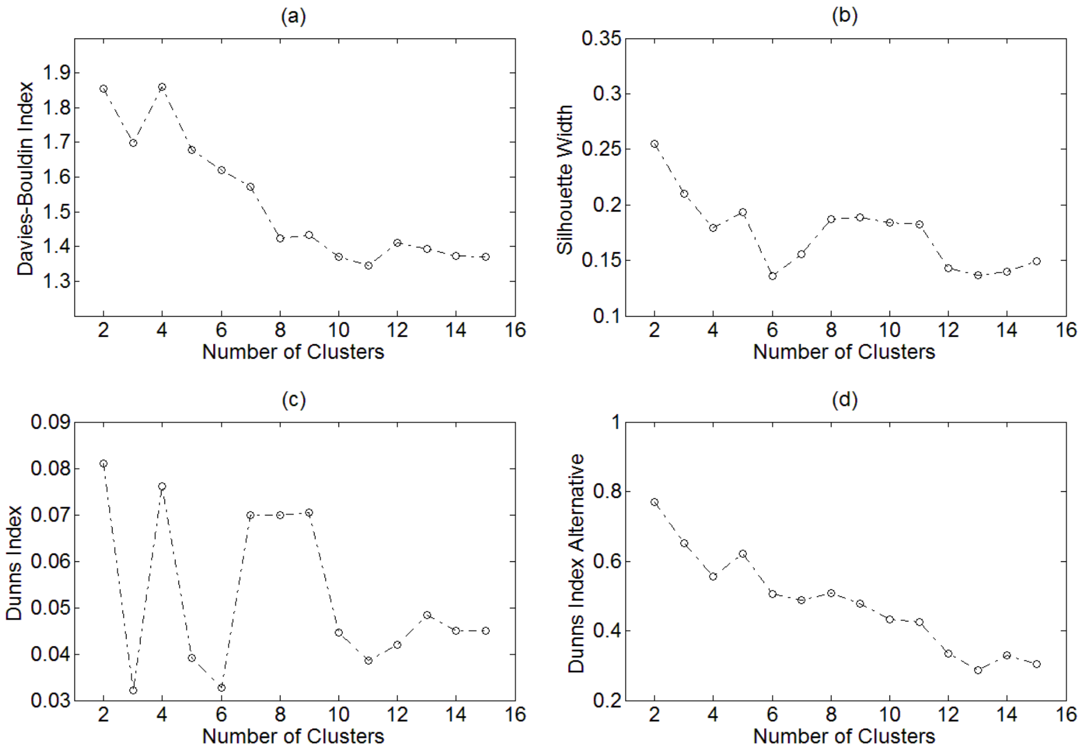

Figure 6-2: All Variables Clustering Validity

The clustering validity (Figure 6-2) related to the use of all the financial variables (M1) shows mixed results. Figure 6-2 (b) – (d) suggests that the smallest SOMs were the most effective whereas Figure 6-2 (a) suggests that the 11x1 SOM was the most effective. Overall the clustering is adequate and the SWs achieved do not fall below 0. Although the smallest SOMs could be considered the most effective, the larger SOMs achieved smaller clusters which are more appropriate for further analysis and for this reason the 2x4 SOM has been investigated in more detail.

Table 6-3: All Variables 2x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	24*	0.11	Financials	54	1	2	3	1	0
2	19	0.09	Basic Materials	37	0	2	0	2	2
3	45*	0.05	Industrials	36	2	3	1	1	1
4	14	0.30	Basic Materials	36	2	1	0	1	1
5	84*	0.31	Industrials	30	2	2	2	0	1
6	34*	0.07	Financials	62	2	0	3	2	3
7	20*	0.16	Basic Materials	25	2	0	0	0	1
8	15*	0.31	Financials	87	0	1	0	1	0
* Investigated in financial analysis									

The general information, regarding the clustering completed with all the input variables, can be seen in Table 6-3. In general the clusters achieved are evenly sized, excluding the single large cluster at neuron 5. None of the clusters have a small number of companies which is a desirable result from the clustering for practical reasons. All the SWs for the individual clusters are above zero; however several of the values are near zero. Cluster 8 consists of mostly financial variables and is also the best cluster from a validity aspect. The reason for this unique cluster can be seen by the high market capitalisation of the financial companies in input 8 (Figure 6-3). None of the clusters have a particularly high number of companies who delisted. To provide complementary information regarding the clusters Figure 6-3 has been provided.

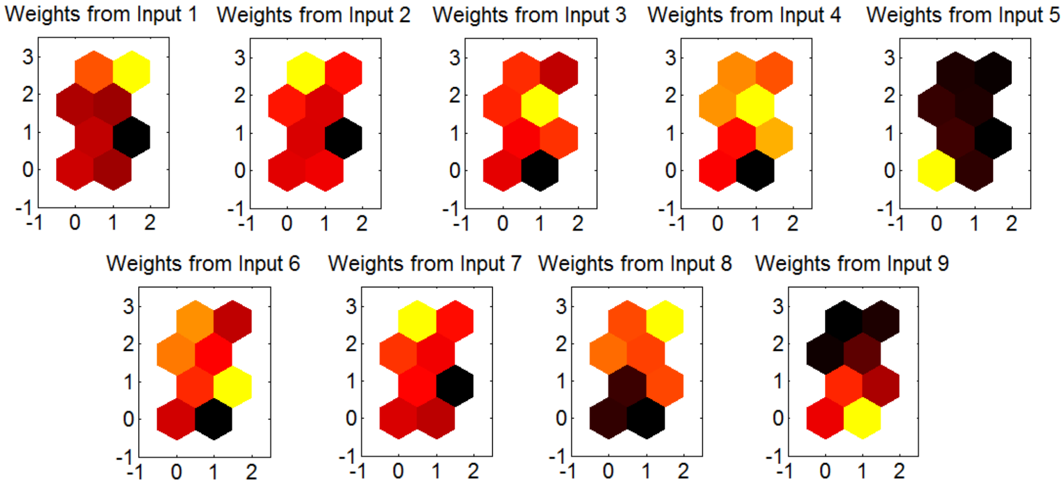

Figure 6-3: All Variables 2x4 SOM Weight Planes

The weight planes in Figure 6-3 show how the different financial variables were responsible for forming specific clusters. QR (input 5) played a significant role in forming cluster 1 which can be seen by the significantly lighter weight input colour at neuron 1. In general DE, PB, PC and RE (inputs 1, 2, 3 and 7) appear to have relatively similar weights when ignoring the clusters with outlying values (shown by black and yellow neurons). The most apparent feature of Figure 6-3 is the fact that each variable appears to have played a role in defining the clusters in a unique manner. PB, PC and RE (inputs 2, 3 and 7) were most affected by the presence of outliers which can be seen by the colour scale showing little detail. This is because the Winsorising was only capable of reducing the number of outliers. Had a more aggressive Winsorising approach been taken the outliers could have been completely removed. In contrast PE, RA and MC (inputs 4, 6 and 8) have resulted in the most evenly distributed colour scales.

In general cluster 5, which is the largest cluster in Table 6-3, achieved relatively medial values for all the inputs and this cluster could be considered closely related to the market when considering the input variables. Cluster 8 from Table 6-3 consisted of mostly financial companies and the most apparent feature of this cluster is its high DE and MC. The clusters occurring at neurons 1, 3, 5, 6, 7 and 8 have all been analysed in more detail in Table 6-4.

Table 6-4: All Variables 2x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1	24	1	-0.0178	-0.0237	0.0653	-0.0238	-0.0297	0.0382
		2	0.0094	0.0362	0.0696	-0.0003	0.0265	0.0519
		3	0.0155	0.0009	0.0562	0.0024	-0.0121	0.0299
		4	-0.0175	-0.0336	0.1052	-0.0126	-0.0288	0.0666
		5	0.0095	0.0049	0.0322	0.0108	0.0061	0.0274
3	45	1	-0.0251	-0.0310	0.0807	-0.0265	-0.0324	0.0490
		2	-0.0710	-0.0442	0.1418	-0.0324	-0.0056	0.0457
		3	0.0166	0.0020	0.0481	0.0069	-0.0077	0.0305
		4	0.0115	-0.0046	0.0301	0.0075	-0.0086	0.0296
		5	-0.0031	-0.0078	0.0494	0.0144	0.0097	0.0282
5	84	1	0.0042	-0.0017	0.0559	-0.0191	-0.0250	0.0543
		2	-0.0472	-0.0204	0.1115	-0.0166	0.0102	0.0617
		3	0.0155	0.0009	0.0472	0.0204	0.0058	0.0354
		4	0.0174	0.0013	0.0436	0.0130	-0.0031	0.0295
		5	0.0032	-0.0015	0.0395	0.0091	0.0044	0.0206
6	34	1	-0.0090	-0.0149	0.0385	-0.0160	-0.0219	0.0433
		2	-0.0438	-0.0170	0.1076	-0.0225	0.0043	0.0711
		3	0.0201	0.0055	0.0288	0.0150	0.0004	0.0194
		4	0.0130	-0.0031	0.0351	0.0008	-0.0153	0.0401
		5	0.0076	0.0029	0.0487	-0.0059	-0.0106	0.0806
7	20	1	0.0247	0.0188	0.0696	-0.0217	-0.0276	0.0454
		2	-0.0372	-0.0104	0.1217	-0.0112	0.0156	0.0714
		3	0.0144	-0.0001	0.0638	0.0206	0.0060	0.0438
		4	0.0183	0.0022	0.0549	0.0032	-0.0129	0.0543
		5	-0.0049	-0.0096	0.0722	0.0081	0.0034	0.0358
8**	15	1	-0.0206	-0.0265	0.0794	-0.0248	-0.0307	0.0653
		2	0.0030	0.0298	0.0854	-0.0026	0.0242	0.0775
		3	0.0090	-0.0056	0.0410	0.0196	0.0051	0.0495
		4	0.0155	-0.0007	0.0503	0.0108	-0.0053	0.0406
		5	0.0128	0.0081	0.0272	0.0102	0.0056	0.0308

** Not included in statistical analysis

With regards to the mean monthly returns, the general trend of poor first year performance is apparent in all the clusters besides the value weighted cluster 7. None of the clusters achieved exceptionally high or low returns consistently; however it should be noted that the largest cluster (neuron 5) did not achieve the lowest standard deviation. In addition to this cluster 8, which consisted of predominantly financial companies (Table 6-3), was not the least diversified cluster.

In Figure 6-4 (a) it can be seen that cluster 8 follows a similar trend to that of the All-Share Index and cluster 7 achieved high value weighted returns over the first 12 months. Cluster 3 had the worst overall performance and cluster 1 had highly irregular performance, which is most evident in the value weighted results. This has been attributed to a single large firm distorting the weighting of the other companies. The greatest deviation in performance can be seen in the first 12 months and after 24 months most of the clusters (value and equally weighted) follow a similar trend to that of the benchmark.

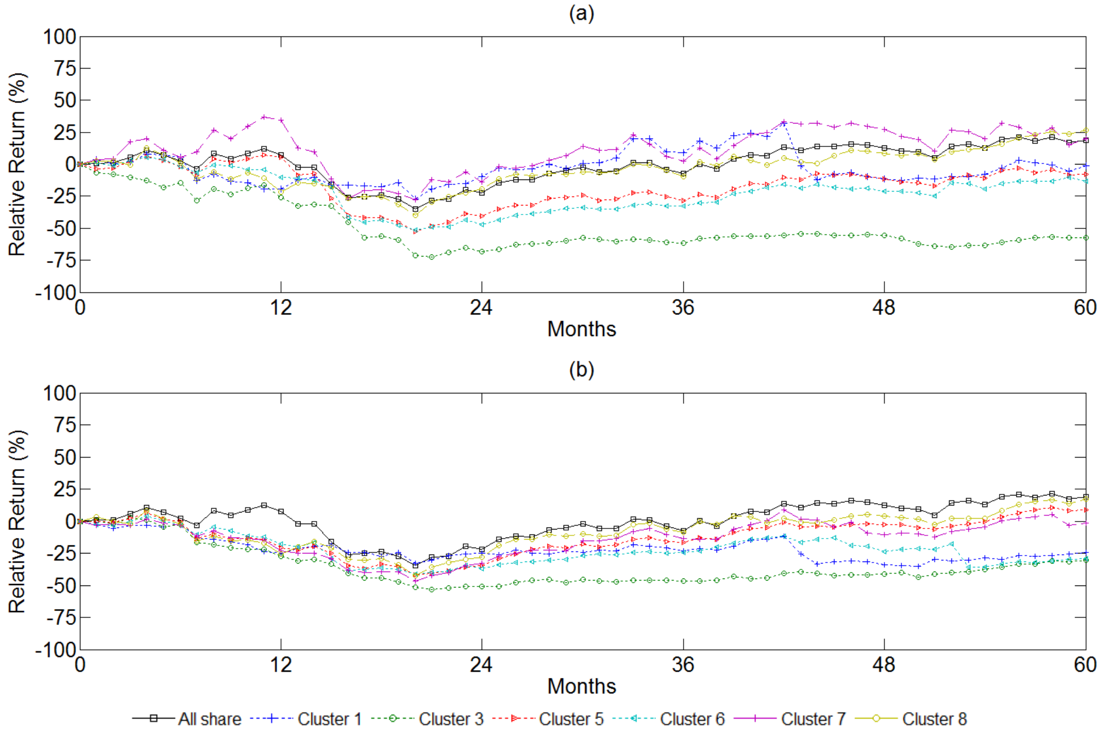

Figure 6-4: All Variables 2x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The statistical analysis shows that there is no evidence of the clusters having different performance. Even in the first 12 months, when Figure 6-4 revealed the greatest difference in returns, there is no evidence of the result being statistically significant. The results for the parametric and non-parametric tests (Table 6-5) indicate strong evidence in support of the returns being the same.

Table 6-5: All Variables 2x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.342764	0.576837	0.984334	0.988285
Years 1 – 3	0.592339	0.76126	0.804922	0.668708
Years 1 – 5	0.743797	0.885017	0.876417	0.769507

6.1.2 Primary Variables

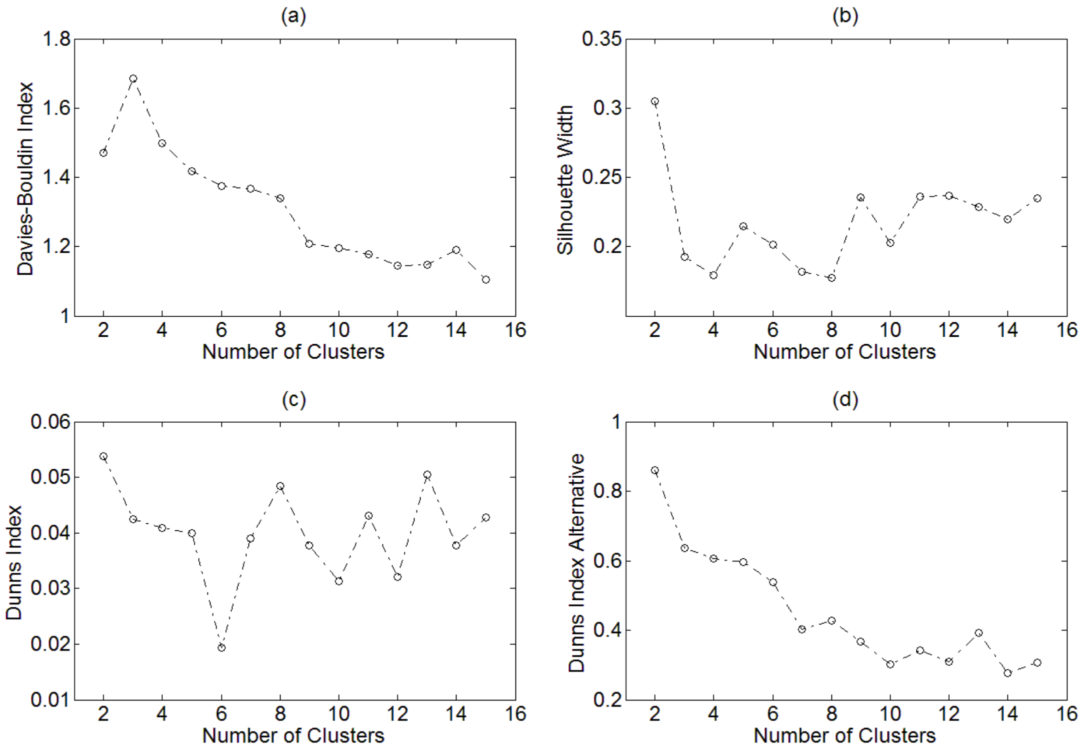

Figure 6-5: Primary Variables Clustering Validity

The clustering validity using only the primary variables (M2) shows significant changes in validity with changes in SOM size (Figure 6-5). As with the results in Figure 6-2 these results show that the DB Index and SW differ. DI yields very erratic results and this can be attributed to the impact of outliers which were not completely removed by Winsorising. To maintain consistency with the previous results the 2x4 SOM was chosen again for further analysis.

Table 6-6: Primary Variables 2x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	20*	0.24	Basic Materials	40	0	4	1	2	1
2	26*	-0.03	Financials	35	1	2	0	0	0
3	17	0.14	Financials	47	1	0	2	4	0
4	11*	0.11	Financials	82	0	1	0	0	0
5	29*	0.19	Consumer Services	24	2	1	1	0	3
6	56*	0.30	Industrials	32	2	2	0	1	1
7	58*	0.28	Consumer Services	26	1	0	2	0	0
8	38*	-0.02	Financials	39	4	1	3	1	4
* Investigated in financial analysis									

The general clustering information for M2 can be seen in Table 6-6. The cluster sizes are relatively constant and two of the clusters achieved negative SWs. Cluster 4 is comprised of mostly financial companies and this can be attributed to a combination of large MC (input 4), low PC (input 2) and V (input 5) (Figure 6-6). Cluster 1 is comprised of companies with significantly higher V (input 5) than the other clusters which is shown by the yellow neuron for input 5. Overall the colour scale is evenly spread for all the inputs with PB (input 1) suffering the most from a cluster of outliers.

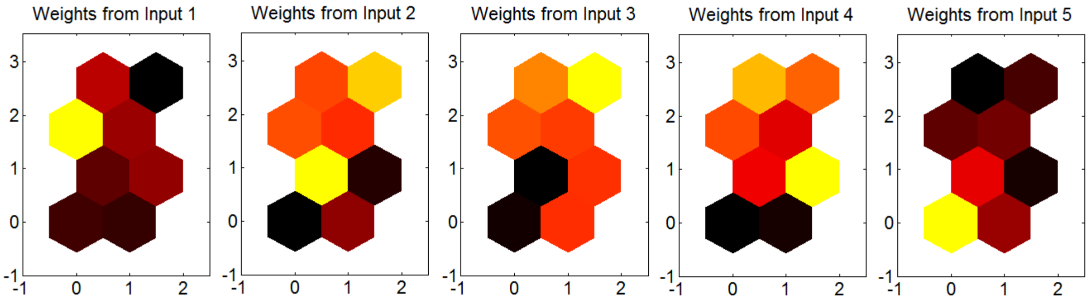

Figure 6-6: Primary Variables 2x4 SOM Weight Planes

Table 6-7: Primary Variables 2x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1	20	1	-0.0356	-0.0414	0.0732	-0.0238	-0.0297	0.0510
		2	-0.0265	0.0003	0.1337	-0.0111	0.0157	0.0692
		3	-0.0286	-0.0432	0.0577	-0.0068	-0.0214	0.0342
		4	-0.0411	-0.0573	0.1784	-0.0184	-0.0345	0.0863
		5	0.0008	-0.0038	0.0330	-0.0083	-0.0130	0.0511
2	26	1	-0.0382	-0.0441	0.0464	-0.0244	-0.0303	0.0358
		2	-0.0677	-0.0409	0.1134	-0.0294	-0.0026	0.0396
		3	-0.0004	-0.0149	0.0441	-0.0042	-0.0188	0.0433
		4	0.0172	0.0011	0.0417	0.0109	-0.0053	0.0318
		5	0.0176	0.0129	0.0367	0.0159	0.0112	0.0224
4**	11	1	-0.0251	-0.0309	0.0849	-0.0132	-0.0190	0.0622
		2	-0.0040	0.0228	0.0897	-0.0005	0.0263	0.0556
		3	0.0200	0.0055	0.0435	0.0191	0.0046	0.0424
		4	0.0122	-0.0040	0.0467	0.0136	-0.0025	0.0396
		5	0.0192	0.0145	0.0325	0.0177	0.0130	0.0280
5	29	1	0.0225	0.0166	0.0678	-0.0177	-0.0236	0.0501
		2	-0.0353	-0.0085	0.1174	-0.0161	0.0107	0.0571
		3	0.0147	0.0001	0.0617	0.0195	0.0050	0.0372
		4	0.0181	0.0020	0.0531	0.0055	-0.0106	0.0375
		5	-0.0035	-0.0082	0.0690	0.0002	-0.0045	0.0404
6	56	1	-0.0091	-0.0150	0.0567	-0.0204	-0.0263	0.0485
		2	-0.0261	0.0007	0.0825	-0.0231	0.0037	0.0539
		3	0.0211	0.0065	0.0343	0.0121	-0.0024	0.0292
		4	0.0085	-0.0076	0.0481	0.0027	-0.0134	0.0366
		5	0.0181	0.0134	0.0254	0.0138	0.0091	0.0265
7	58	1	-0.0004	-0.0063	0.0555	-0.0220	-0.0279	0.0583
		2	-0.0488	-0.0220	0.1140	-0.0070	0.0198	0.0705
		3	0.0170	0.0024	0.0516	0.0218	0.0072	0.0412
		4	0.0202	0.0040	0.0443	0.0126	-0.0035	0.0384
		5	0.0033	-0.0014	0.0437	0.0079	0.0032	0.0260

8	38	1	0.0018	-0.0041	0.0473	-0.0096	-0.0155	0.0428
		2	-0.0338	-0.0070	0.1109	-0.0324	-0.0056	0.0820
		3	0.0130	-0.0016	0.0443	0.0164	0.0019	0.0292
		4	0.0092	-0.0069	0.0352	0.0050	-0.0111	0.0288
		5	0.0048	0.0001	0.0285	-0.0317	-0.0364	0.0870
** Not included in statistical analysis								

The financial results for the clustering completed using only the primary variables shows a trend of the poor first and second year average monthly returns. This is most evident when equal weighting is given to the companies in each portfolio. Considering the equal weightings cluster 2 has achieved the lowest standard deviation. Even though cluster 4 was comprised of predominantly financial companies it was still able to achieve a relatively low standard deviation. Considering the value weighted and equally weighted returns shows that clusters 1 and 2 could be considered the poorest performer, which is also evident in Figure 6-7. The performance of cluster 7 can also be seen to vary significantly in Figure 6-7 due to the clusters small size. None of the clusters achieved notably high returns over any of the years, however cluster 8 achieved very poor fifth year equal weighting results.

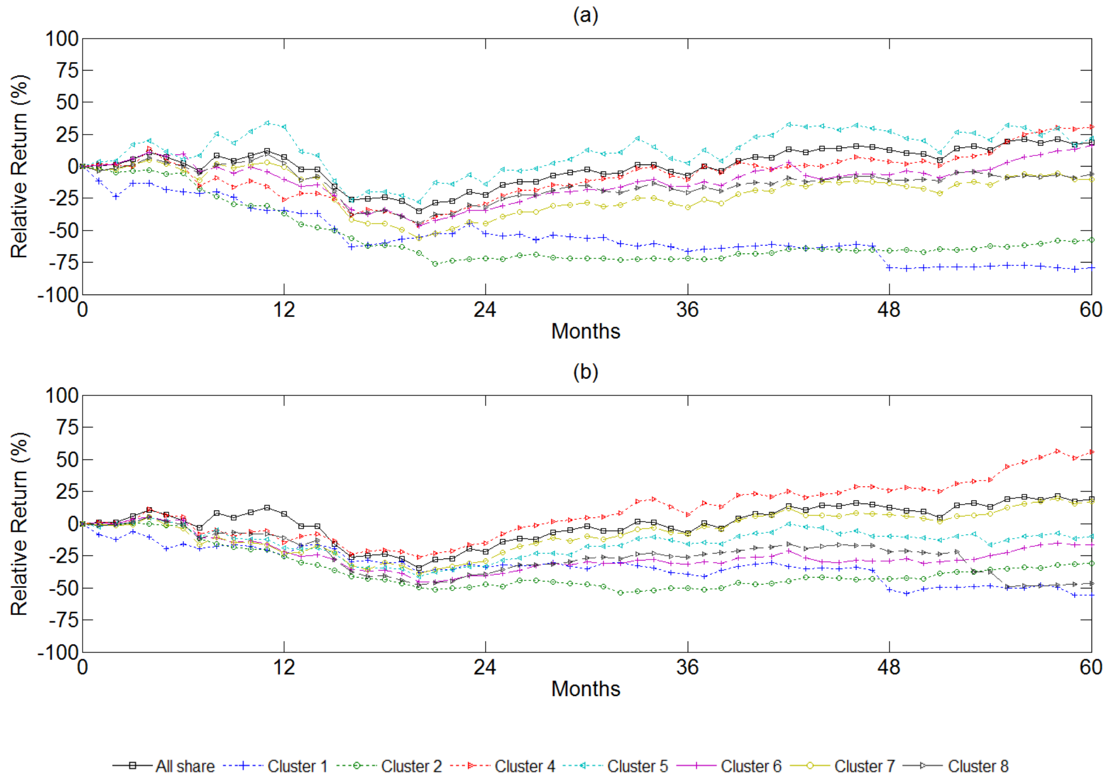

Figure 6-7: Primary Variables 2x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The statistical results shown in Table 6-8 were computed using clusters 1, 2, 5, 6, 7 and 8. As with the results shown in Table 6-5 there is insufficient evidence to support the hypothesis that the clusters achieved significantly different results. By considering the returns shown in Table 6-7 it can be concluded that there is no difference in the performance of the clusters.

A direct comparison between the value and equally weighted results in Table 6-8 shows that the value weighted results were more statistically significant. Cluster 5 was assumed to be the main driver behind this because it had the most different performance over the first 12 months, when using value weighting. It was found that the returns for this cluster were largely driven by BHP Billiton which outweighed its counterparts.

Table 6-8: Primary Variables 2x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.104908	0.132077	0.977726	0.912542
Years 1 – 3	0.296169	0.264155	0.764343	0.426352
Years 1 – 5	0.207067	0.398918	0.512728	0.521047

Although the results in Table 6-9 show that the null hypothesis of similar returns could not be rejected it was decided to investigate the results in more detail. The first year value weighted post hoc Tukey HSD results are shown in Table 6-9 (page 68) where it can be seen that cluster 5's relationship with clusters 1 and 2 was the main driver behind the high statistical significance. Apart from this, clusters 1 and 2 can be seen to have very similar performance as well as clusters 7 and 8 in Table 6-9 (page 68). Similar performance would not normally be considered important however in this case the clusters with similar performance were next to one another in the SOM implying they had similar input vectors.

Table 6-9: Primary Variables 2x4 Year 1 Tukey HSD Results

Portfolio		Mean Difference	Significance	95% Confidence Interval	
I	J	(I-J)		Lower Bound	Upper Bound
1	2	0.0026	1.0000	-0.0677	0.0729
	5	-0.0581	0.1629	-0.1284	0.0122
	6	-0.0265	0.8774	-0.0968	0.0438
	7	-0.0351	0.6866	-0.1054	0.0352
	8	-0.0374	0.6274	-0.1077	0.0329
2	1	-0.0026	1.0000	-0.0729	0.0677
	5	-0.0607	0.1292	-0.1310	0.0096
	6	-0.0291	0.8281	-0.0994	0.0412
	7	-0.0377	0.6174	-0.1081	0.0326
	8	-0.0400	0.5567	-0.1103	0.0303
5	1	0.0581	0.1629	-0.0122	0.1284
	2	0.0607	0.1292	-0.0096	0.1310
	6	0.0316	0.7735	-0.0387	0.1019
	7	0.0230	0.9293	-0.0474	0.0933
	8	0.0207	0.9535	-0.0496	0.0910
6	1	0.0265	0.8774	-0.0438	0.0968
	2	0.0291	0.8281	-0.0412	0.0994
	5	-0.0316	0.7735	-0.1019	0.0387
	7	-0.0086	0.9992	-0.0790	0.0617
	8	-0.0109	0.9975	-0.0812	0.0594
7	1	0.0351	0.6866	-0.0352	0.1054
	2	0.0377	0.6174	-0.0326	0.1081
	5	-0.0230	0.9293	-0.0933	0.0474
	6	0.0086	0.9992	-0.0617	0.0790
	8	-0.0022	1.0000	-0.0726	0.0681
8	1	0.0374	0.6274	-0.0329	0.1077
	2	0.0400	0.5567	-0.0303	0.1103
	5	-0.0207	0.9535	-0.0910	0.0496
	6	0.0109	0.9975	-0.0594	0.0812
	7	0.0022	1.0000	-0.0681	0.0726

6.2 Single Variable Clustering

The single variable clustering yielded mixed results. In general the clustering validity shows a low number of clusters are optimal. From a financial cluster analysis perspective this is not practical and for this reason the 3x4 SOM results were analysed in more detail for DE, PB, PC, PE, QR, RA and RE. In comparison the clustering completed with MC and V achieved the best cluster sizes for financial analysis with 2x4 SOMs.

6.2.1 Debt/ Equity

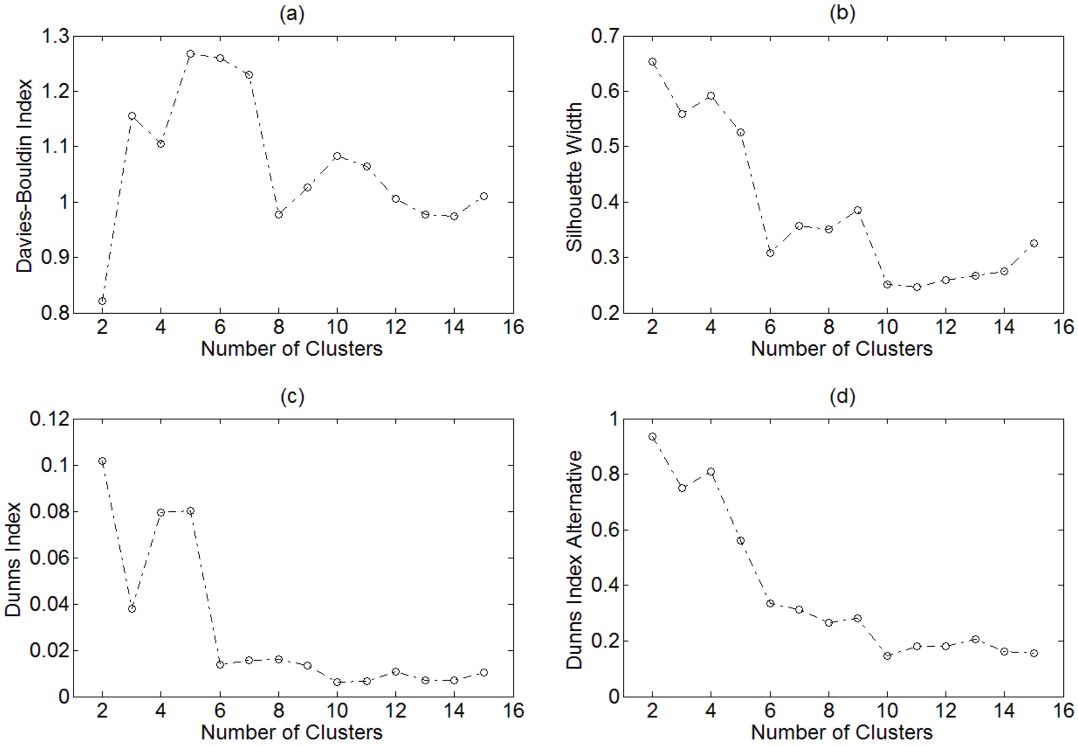

Figure 6-8: Debt/ Equity Clustering Validity

The clustering validity in Figure 6-8 for DE suggests that the ideal number of clusters is two. The cluster sizes related to this test are 36 and 219 and therefore the small SOM does not lend itself to financial interpretation for the ideal number of clusters. A similar trend is noticed with the other small SOMs so by only comparing the larger SOM networks (10 – 15) it can be seen that there is little difference in the clustering validity. The 3x4 achieved the most appropriate cluster sizes for further analysis and has been shown in Table 6-10.

Table 6-10: Debt/ Equity 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1*	13	0.66	Financials	100	0	1	0	0	0
2	11	0.16	Financials	45	1	1	1	2	1
3	5	0.21	Consumer Services	40	2	0	0	0	0
4	2	0.12	Technology	50	1	0	0	0	0
5	8	0.05	Financials	50	0	0	0	1	0
6	4	0.49	Financials	50	0	0	1	0	0
7	9	0.23	Financials	67	1	0	0	1	0
8*	37	-0.01	Industrials	32	2	1	0	2	2
9*	82	0.31	Industrials	26	1	4	3	0	1
10	3	0.25	Technology	33	1	0	0	0	1
11*	71	0.33	Financials	28	2	3	4	1	4
12	10	-0.02	Basic Materials	40	0	1	0	1	0

* Investigated in financial analysis

Three main clusters can be seen in Table 6-10 (neurons 8,9 and 11) and one cluster is comprised of only financial companies (neuron 1). Analysis of Figure 6-9 shows that clusters 8, 9 and 11 are all relatively similar with the smaller clusters being comprised of companies with irregular inputs. The financial cluster (neuron 1) is comprised of companies with high leverage as shown in Figure 6-9. Cluster 4, which is the smallest cluster, had the most erratic inputs with the weight plane for input 3 differing significantly to the other planes. This would not be apparent if the clustering was completed using only one year of input values. Overall the colour range shows that the SOM was able to show small changes in DE and this variable was not distorted by outliers. Clusters 1, 8, 9 and 11 have been investigated in more detail in Table 6-11.

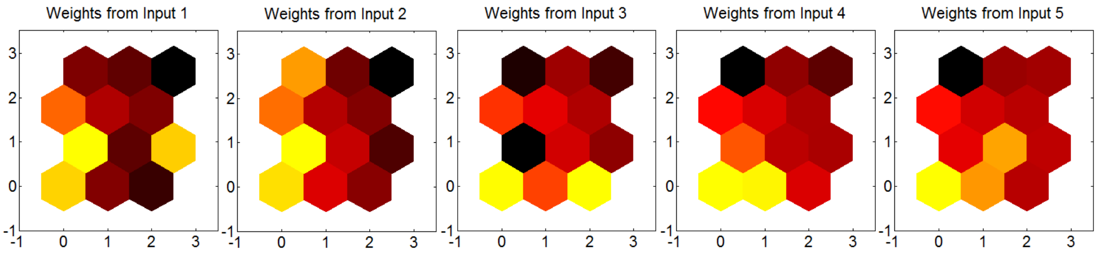

Figure 6-9: Debt/ Equity 3x4 SOM Weight Planes

Table 6-11: Debt/ Equity 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1**	13	1	-0.0322	-0.0381	0.0793	-0.0242	-0.0301	0.0577
		2	0.0061	0.0329	0.0931	0.0066	0.0334	0.0615
		3	0.0174	0.0028	0.0482	0.0219	0.0074	0.0409
		4	0.0100	-0.0061	0.0481	0.0184	0.0023	0.0369
		5	0.0164	0.0117	0.0336	0.0141	0.0094	0.0297
8	37	1	-0.0155	-0.0214	0.0658	-0.0179	-0.0238	0.0567
		2	-0.0054	0.0214	0.0732	-0.0300	-0.0032	0.0627
		3	0.0202	0.0057	0.0323	0.0102	-0.0044	0.0325
		4	0.0152	-0.0009	0.0382	0.0112	-0.0049	0.0400
		5	0.0185	0.0138	0.0261	0.0035	-0.0012	0.0307
9	82	1	0.0124	0.0065	0.0571	-0.0212	-0.0271	0.0438
		2	-0.0488	-0.0220	0.1162	-0.0241	0.0027	0.0609
		3	0.0134	-0.0011	0.0558	0.0155	0.0009	0.0338
		4	0.0169	0.0008	0.0480	0.0058	-0.0104	0.0320
		5	-0.0042	-0.0089	0.0548	0.0042	-0.0005	0.0230
11	71	1	-0.0040	-0.0099	0.0448	-0.0186	-0.0245	0.0378
		2	-0.0470	-0.0202	0.1397	-0.0148	0.0120	0.0476
		3	0.0192	0.0046	0.0247	0.0129	-0.0016	0.0206
		4	0.0150	-0.0011	0.0359	0.0029	-0.0133	0.0327
		5	0.0075	0.0028	0.0447	0.0019	-0.0028	0.0319
** Not included in statistical analysis								

Only four clusters were chosen for financial analysis, of which only three were of adequate size. Cluster 1 was comprised of only financial companies and the portfolio standard deviation for this cluster was greater than the others in the first year, however this is not noticeable in the later years. The first noticeable feature is the mean monthly excess returns for cluster 9 which are relatively near zero, implying that this cluster followed a similar value weighted performance to the JSE All-share Index. This feature is most apparent in the first 12 months in Figure 6-10 (page 72). After this, cluster 9 has poor performance relative to the benchmark for 12 months and then follows the same trend as the benchmark. In general Cluster 8 obtained the best value weighted returns over years three to five, which is apprent in both Table 6-11 and Figure 6-10 (page 72).

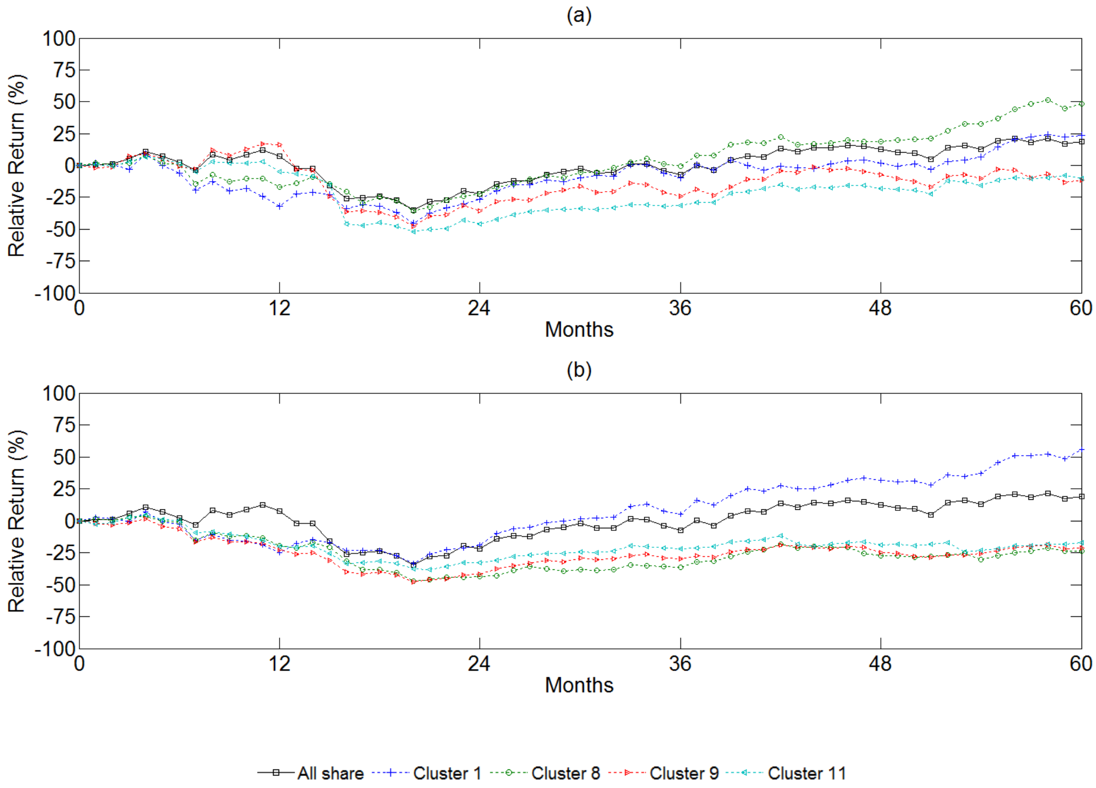

Figure 6-10: Debt/ Equity 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

From the ANOVA and Kruskal-Wallis analysis (Table 6-12) it can be seen that clusters 8,9 and 11 achieved similar result as the tests were unable to reject the null hypothesis. This lack of statistical significance is confrmed visually in Figure 6-10 where all the clusters are seen to follow a similar trend.

Table 6-12: Debt/ Equity 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.484839	0.808591	0.983334	0.94525
Years 1 – 3	0.847135	0.732637	0.878305	0.911849
Years 1 – 5	0.71803	0.429329	0.982826	0.994558

6.2.2 Price/ Book Value

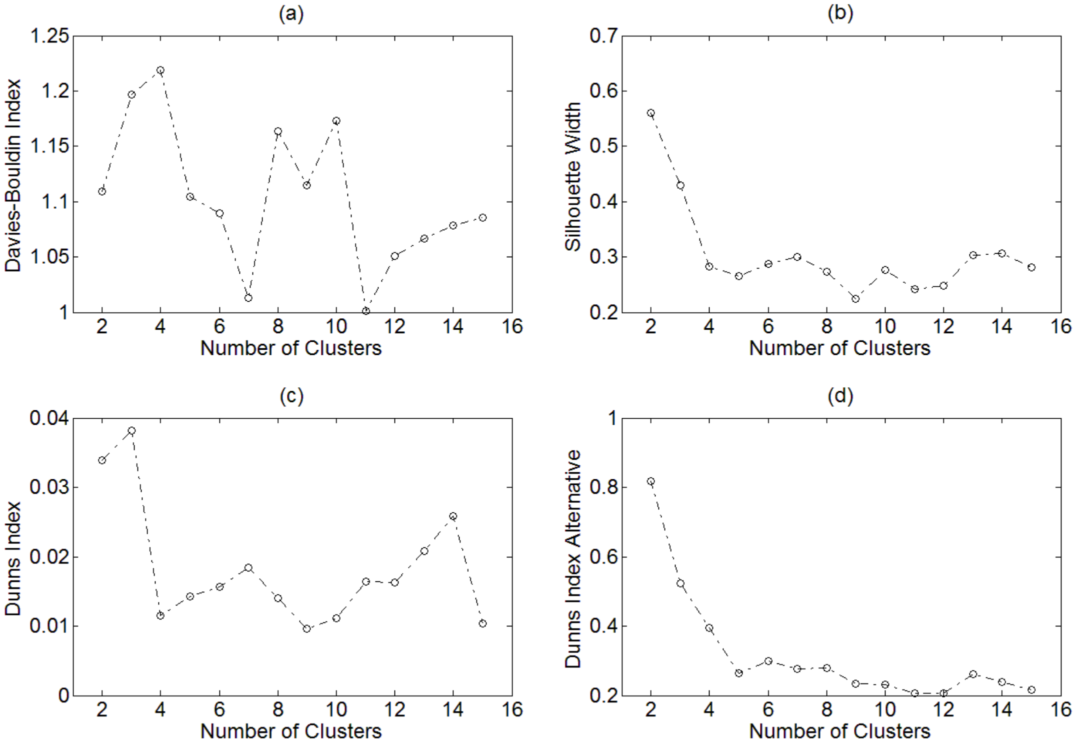

Figure 6-11: Price/ Book Value Clustering Validity

The PB clustering validity reveals mixed results as shown in Figure 6-11. This can be attributed to the presence of outliers within the data making clustering more difficult. Again the smaller SOMs achieved superior validity for the majority of the tests. Considering the larger SOMs, the DB index shows that 10 clusters achieved very poor results. In general 11x1 or 3x4 SOMs have produced the best clustering validity. Again the 3x4 SOM was chosen. Although the 11x1 SOM achieved better validity in several tests it is a one dimensional SOM and this would have compressed the input data more.

Table 6-13: Price/ Book Value 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	7	0.52	Financials	43	1	0	0	1	0
2	5	0.18	Telecommunications	20	0	1	0	1	0
3	6	0.43	Basic Materials	50	0	0	0	2	0
4	13	0.17	Consumer Services	31	2	1	0	0	0
5	41*	0.13	Financials	37	1	0	1	2	0
6	3*	0.48	Consumer Services	100	1	0	0	0	0
7	7	0.00	Technology	29	2	0	0	0	0
8	35*	0.18	Consumer Services	31	0	0	0	0	1
9	83*	0.37	Financials	35	3	3	5	2	3
10	6	-0.04	Technology	33	0	0	1	0	2
11	44*	0.18	Financials	34	1	6	2	0	2
12	5*	0.54	Basic Materials	100	0	0	0	0	1

* Investigated in financial analysis

Four large clusters are apparent in Table 6-13 and the largest cluster (neuron 9) being significantly larger than the others. Overall the SW values for the individual clusters are high and since clusters 6 and 12 were comprised of single industries they have been included in the financial analysis. Clusters 5, 8, 9 and 11 all achieved relatively similar inputs with cluster 11 having slightly lower inputs. The smaller clusters from Table 6-13 are comprised of irregular inputs (Figure 6-12). The overall colour scale in Figure 6-12 is evenly spread however the PB input across the five years varies. The most important clusters from Table 6-13 have been investigated in more detail in Table 6-14.

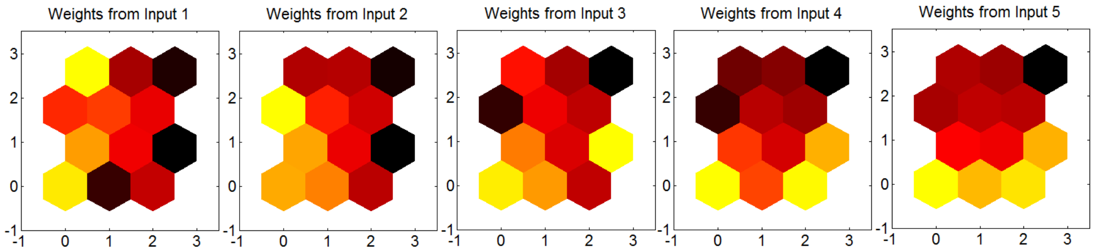

Figure 6-12: Price/ Book Value 3x4 SOM Weight Planes

Table 6-14: Price/ Book Value 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
5	41	1	-0.0082	-0.0141	0.0571	-0.0200	-0.0258	0.0502
		2	-0.0255	0.0013	0.0929	-0.0124	0.0144	0.0456
		3	0.0177	0.0031	0.0344	0.0117	-0.0029	0.0287
		4	0.0133	-0.0029	0.0339	0.0112	-0.0049	0.0336
		5	0.0105	0.0058	0.0382	0.0116	0.0069	0.0196
6**	3	1	-0.0143	-0.0202	0.0472	-0.0103	-0.0162	0.0352
		2	0.0203	0.0471	0.0517	0.0205	0.0473	0.0645
		3	0.0231	0.0085	0.0352	0.0202	0.0057	0.0278
		4	0.0015	-0.0146	0.0577	0.0065	-0.0096	0.0500
		5	0.0114	0.0067	0.0467	0.0129	0.0082	0.0413
8	35	1	0.0168	0.0109	0.0687	-0.0240	-0.0299	0.0609
		2	-0.0506	-0.0238	0.1283	-0.0208	0.0060	0.0694
		3	0.0173	0.0027	0.0660	0.0204	0.0058	0.0385
		4	0.0222	0.0060	0.0525	0.0137	-0.0024	0.0332
		5	-0.0024	-0.0071	0.0651	0.0093	0.0046	0.0220
9	83	1	-0.0211	-0.0270	0.0506	-0.0249	-0.0308	0.0456
		2	-0.0337	-0.0069	0.0915	-0.0215	0.0053	0.0600
		3	0.0129	-0.0016	0.0328	0.0172	0.0027	0.0260
		4	0.0117	-0.0044	0.0352	0.0077	-0.0084	0.0322
		5	0.0082	0.0035	0.0249	0.0094	0.0047	0.0220
11	44	1	-0.0212	-0.0271	0.0455	-0.0183	-0.0242	0.0356
		2	-0.0541	-0.0273	0.0842	-0.0186	0.0082	0.0432
		3	-0.0043	-0.0188	0.0837	-0.0059	-0.0204	0.0187
		4	-0.0060	-0.0221	0.0680	-0.0056	-0.0217	0.0468
		5	0.0045	-0.0002	0.0298	-0.0036	-0.0083	0.0600
12**	5	1	0.0082	0.0023	0.0875	0.0193	0.0134	0.1127
		2	-0.0786	-0.0518	0.2331	-0.1033	-0.0765	0.2065
		3	0.0208	0.0062	0.0785	0.0117	-0.0029	0.1019
		4	-0.0090	-0.0251	0.0553	0.0013	-0.0148	0.0535
		5	-0.0437	-0.0484	0.0767	-0.1341	-0.1388	0.3845
** Not included in statistical analysis								

The financial results in Table 6-14 show that cluster 6 achieved low portfolio standard deviations, even though it only consisted of three companies. This cluster also achieved high returns in the second year, especially considering the poor returns for the market in general.

This is highlighted by the large excess monthly average returns for the second year. Cluster 11 performed very poorly over all five years for both the value and equally weighted results. In fact this cluster only achieved positive mean monthly returns in the fifth year. Cluster 12 was comprised of companies only from the basic materials industry and it did achieve poor portfolio standard deviation for the first three years.

Clusters 6 and 12, which are the two smallest PB clusters, had the greatest difference in performance over the five year period as shown in Figure 6-13. Besides these two clusters none of the returns achieved stand out from the general trend.

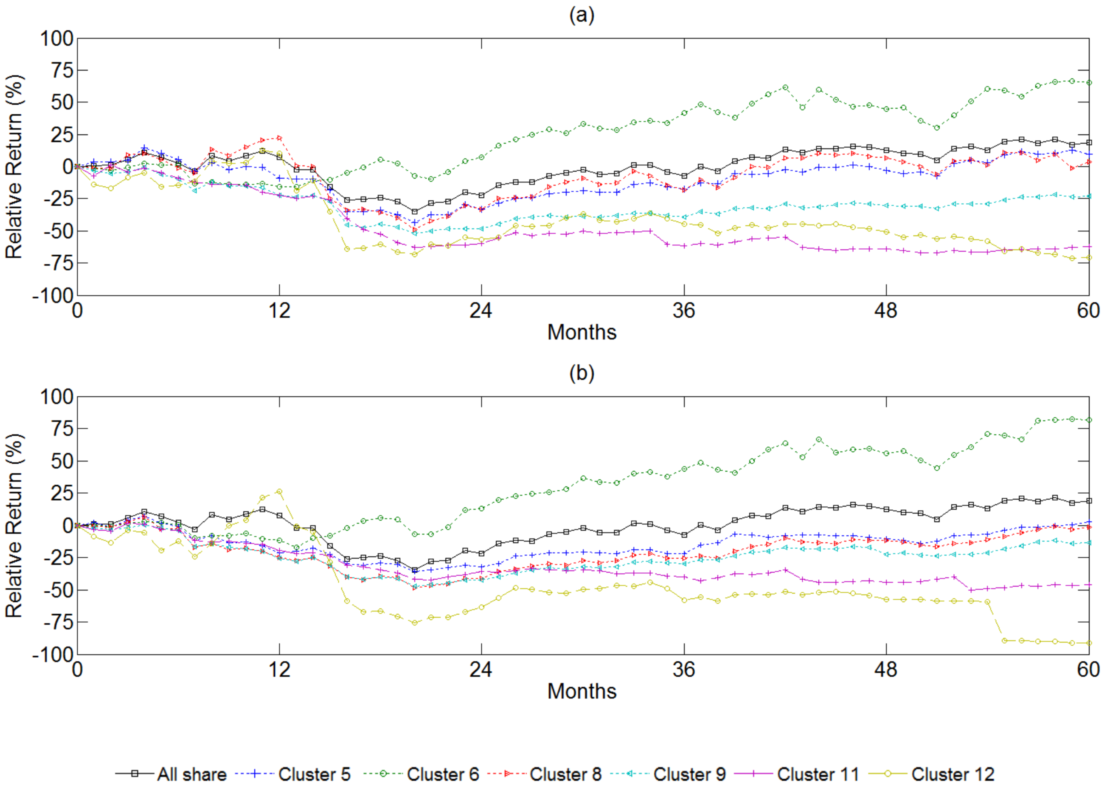

Figure 6-13: Price/ Book Value 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The most significant statistical results in Table 6-15 (page Table 6-15) occurred over the single year and five year periods. Again the value weighted results indicate more likely evidence of the clusters performing differently, however, the significance of these results is still very poor. A more detailed look at the statistical results shows that none of the value weighted clusters achieved significantly different results over any of the test periods and similar results can be noted for the equally weighted tests. Figure 6-13, with the exclusion of clusters 6 and 12 which were not included in the statistical analysis, supports this conclusion that the performance of the clusters was relatively similar.

Table 6-15: Price/ Book Value 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.31558	0.448116	0.985073	0.982655
Years 1 – 3	0.604197	0.650791	0.919716	0.521299
Years 1 – 5	0.432343	0.640406	0.504203	0.294682

6.2.3 Price/ Cash Flow

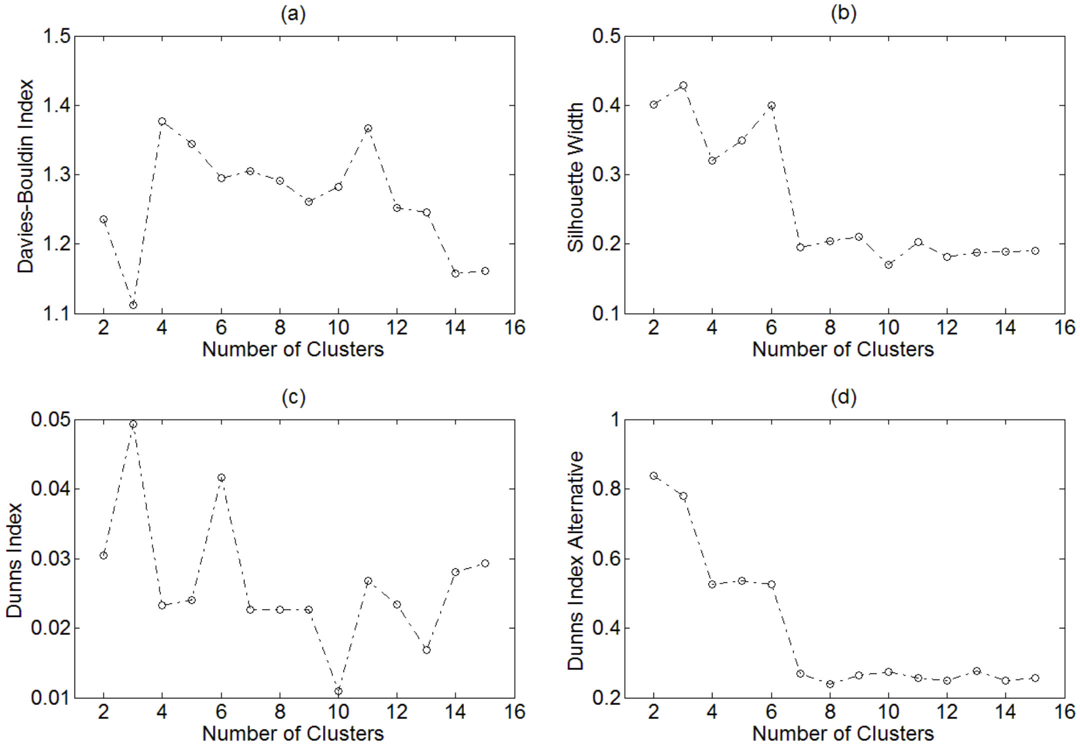

Figure 6-14: Price/ Cash Flow Clustering Validity

As with the PB clustering (Figure 6-11), the PC clustering (Figure 6-14) has yielded large jumps in clustering validity as the SOM size increases. SW, DI and DI alternative suggest that 6 clusters achieved relatively good clustering. The cluster sizes for this test are however not ideal for further analysis and to be consistent with the majority of the tests in this section it was decided to rather analyse the 3x4 SOM in more detail.

Table 6-16: Price/ Cash Flow 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	10	0.09	Financials	50	2	0	0	1	0
2	10	0.15	Financials	60	0	0	0	2	0
3	6	0.07	Financials	50	0	1	1	0	0
4	26*	0.04	Financials	35	0	0	0	1	3
5	69*	0.27	Consumer Services	29	3	1	2	1	1
6	28*	0.11	Basic Materials	29	1	3	0	1	2
7	13	-0.03	Financials	69	1	0	4	1	0
8	70*	0.22	Industrials	34	3	4	2	0	3
9	4	0.12	Basic Materials	50	0	0	0	0	0
10	2*	0.29	Basic Materials	100	0	0	0	0	0
11	11	0.06	Financials	36	1	2	0	0	0
12	6	0.66	Basic Materials	50	0	0	0	1	0

* Investigated in financial analysis

Table 6-16 shows that the PC clustering achieved two very large clusters (neurons 5 and 8) and two slightly smaller clusters (neurons 4 and 6) which have been considered appropriate for financial analysis. Cluster 6 is comprised of companies with lower PC values than the other significant clusters (Figure 6-15). Cluster 10 is comprised of only two companies due to inconsistent PC values (Figure 6-15). Overall the SW values in Table 6-16 are low with only the small cluster at neuron 12 achieving a high SW.

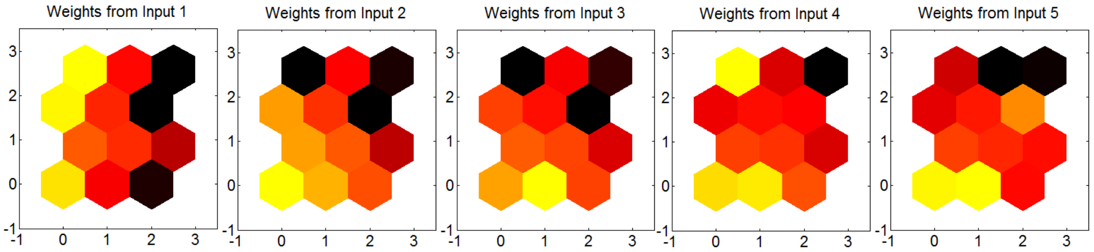

Figure 6-15: Price/ Cash Flow 3x4 SOM Weight Planes

Table 6-17: Price/ Cash Flow 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
4	26	1	0.0043	-0.0016	0.0505	-0.0173	-0.0231	0.0475
		2	-0.0370	-0.0102	0.1466	-0.0127	0.0141	0.0731
		3	0.0224	0.0079	0.0492	0.0205	0.0059	0.0333
		4	0.0058	-0.0103	0.0392	-0.0049	-0.0211	0.0597
		5	0.0105	0.0058	0.0363	-0.0085	-0.0132	0.0763
5	69	1	0.0029	-0.0030	0.0602	-0.0181	-0.0240	0.0515
		2	-0.0466	-0.0198	0.1233	-0.0086	0.0182	0.0567
		3	0.0124	-0.0022	0.0529	0.0193	0.0048	0.0321
		4	0.0200	0.0039	0.0458	0.0142	-0.0020	0.0289
		5	-0.0016	-0.0063	0.0395	0.0044	-0.0003	0.0350
6	28	1	-0.0257	-0.0315	0.0867	-0.0167	-0.0226	0.0495
		2	-0.0050	0.0218	0.0969	-0.0300	-0.0032	0.0509
		3	0.0150	0.0004	0.0421	0.0010	-0.0136	0.0270
		4	0.0058	-0.0103	0.0481	-0.0024	-0.0185	0.0604
		5	0.0155	0.0108	0.0318	-0.0007	-0.0054	0.0404
8	70	1	0.0123	0.0064	0.0598	-0.0197	-0.0256	0.0455
		2	-0.0385	-0.0117	0.0916	-0.0267	0.0001	0.0596
		3	0.0116	-0.0030	0.0495	0.0131	-0.0014	0.0355
		4	0.0213	0.0051	0.0470	0.0079	-0.0082	0.0312
		5	-0.0006	-0.0053	0.0548	0.0051	0.0005	0.0239
10**	2	1	-0.0036	-0.0095	0.1743	0.0052	-0.0007	0.1840
		2	-0.0136	0.0132	0.1818	-0.0078	0.0190	0.1732
		3	-0.0011	-0.0156	0.0621	-0.0223	-0.0369	0.0675
		4	0.0085	-0.0077	0.0768	0.0033	-0.0128	0.0688
		5	-0.0108	-0.0155	0.1011	0.0148	0.0101	0.1080

** Not included in statistical analysis

Only four clusters in Table 6-17 were comprised of a significant number of companies. Cluster 10 consisted of two companies and therefore achieved poor portfolio diversification over the first two years. The performance of cluster 10 in Figure 6-16 (page 81) shows relatively stable returns considering the cluster only contains two companies. For equal weighting, cluster 6 achieved poor returns over all five output years. However, looking at Figure 6-16 shows that this poor performance does not stand out from that of the other clusters.

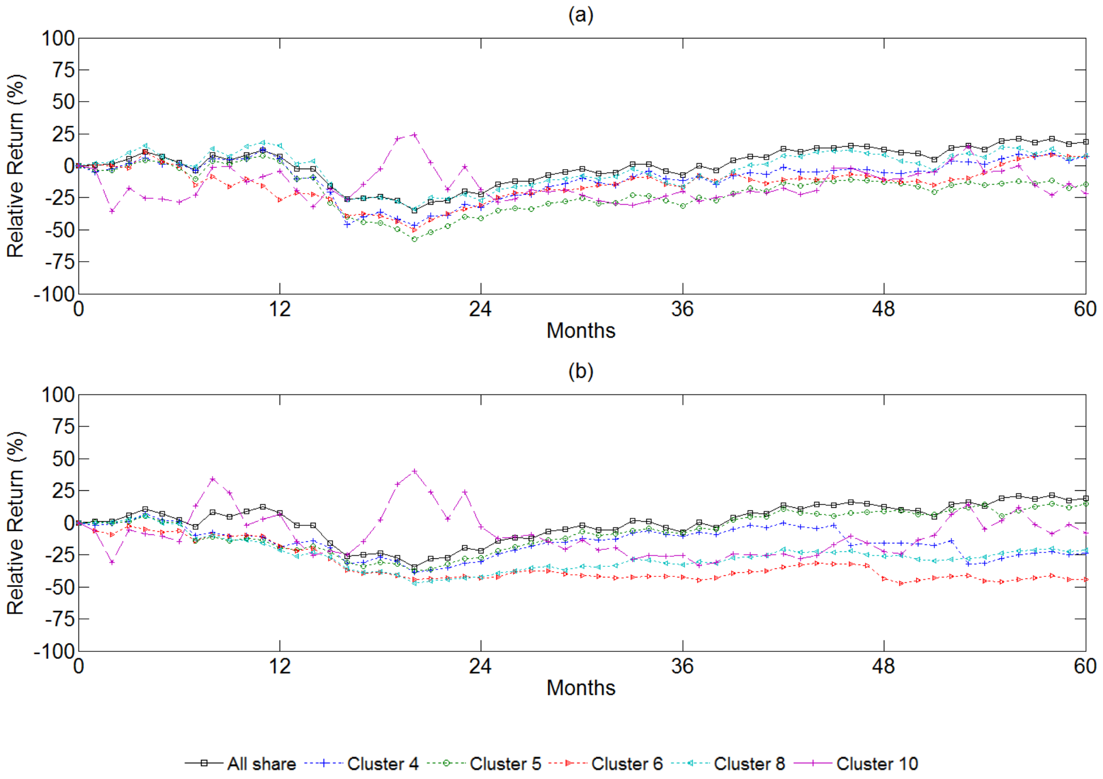

Figure 6-16: Price/ Cash Flow 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The relatively similar returns noted in Figure 6-16 and Table 6-17 are evident in Table 6-18 where it can be seen that the null hypothesis could not be rejected. The non-parametric results for the equally weighted clusters in Table 6-18 over the three and five year periods indicate the greatest likelihood of the clusters having different results. Even though these results are the most indicative of differing cluster performance the evidence strongly suggests that the performance was the same.

Table 6-18: Price/ Cash Flow 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.519774	0.565027	0.998836	0.993571
Years 1 – 3	0.985548	0.928124	0.63778	0.293067
Years 1 – 5	0.987918	0.946698	0.605451	0.313924

6.2.4 Price/ Earnings

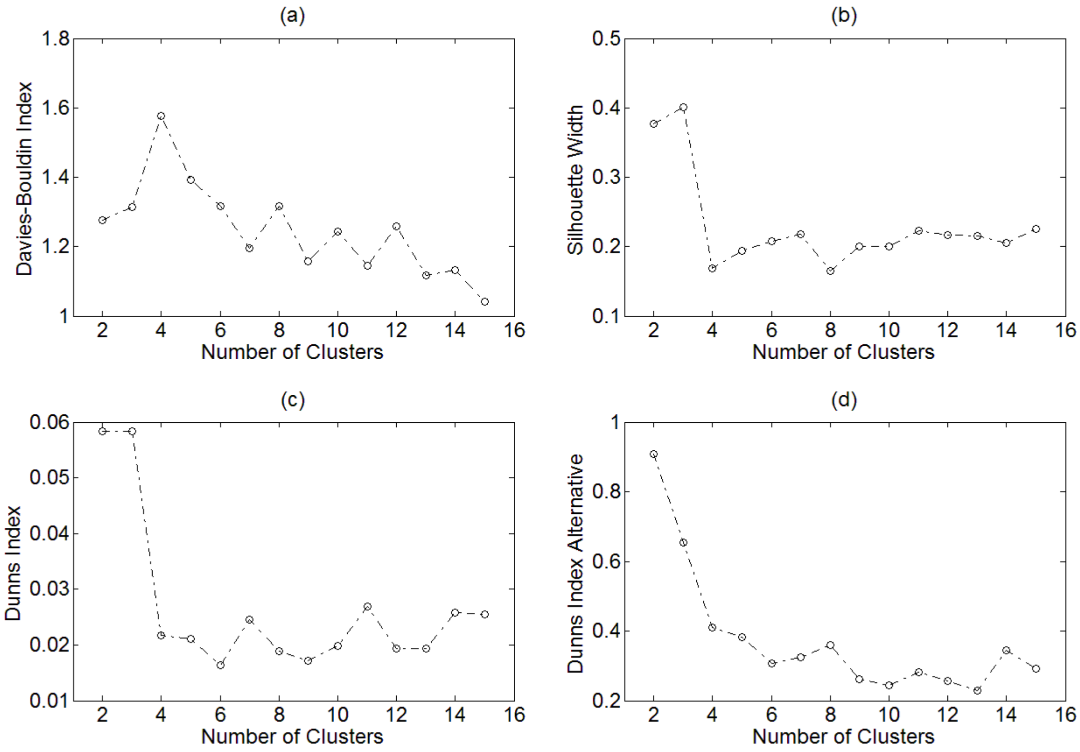

Figure 6-17: Price/ Earnings Clustering Validity

The PE clustering has resulted in the best clustering being achieved with the very small SOMs (Figure 6-17). Considering the cluster sizes from 5 onwards it can be seen that the clustering validity has a general trend of improved performance. Evaluating the largest SOM reveals that as the number of neurons increased, the dominant cluster remained relatively unchanged and the smaller clusters were broken down. This same trend was noted with the other clustering tests and is due to the very compact data for the price ratios. The clusters produced by this variable were consistently uneven, making analysis difficult. It was decided to analyse the 3x4 SOM to be consistent with the previous tests, even though this SOM did not produce ideal clusters for analysis.

Table 6-19: Price/ Earnings 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	13	0.19	Basic Materials	38	0	1	0	1	2
2	13	0.05	Consumer Goods	23	0	2	1	0	1
3	5	-0.05	Financials	60	1	0	1	1	0
4	24*	-0.04	Consumer Services	25	3	1	1	0	1
5	92*	0.47	Industrials	26	3	1	1	0	2
6	10	0.16	Basic Materials	40	0	1	0	3	0
7	8	0.28	Basic Materials	38	0	0	1	0	0
8	53*	0.02	Financials	34	2	2	2	0	1
9	18	0.12	Financials	33	1	3	1	1	2
10	6	0.00	Consumer Services	33	0	0	0	1	0
11	3	0.43	Financials	67	0	0	1	0	0
12	10*	0.16	Basic Materials	80	1	0	0	1	0
* Investigated in financial analysis									

The PE general results are shown in Table 6-19 where it can be seen that this input variable was not able to generate even cluster sizes. This large cluster (neuron 5) achieved a relatively high SW however the majority of the remaining clusters have lower SWs. The second and third largest clusters (neurons 4 and 8 respectively) achieved very poor SWs. In general cluster 4 is comprised of companies with greater PE than cluster 5 (Figure 6-18). Cluster 8 is comprised of companies with lower PE ratios than cluster 5 (Figure 6-18). Cluster 12 consists mostly of basic material companies and this cluster had the lowest PE values across the five years as shown by the dark neurons in Figure 6-18.

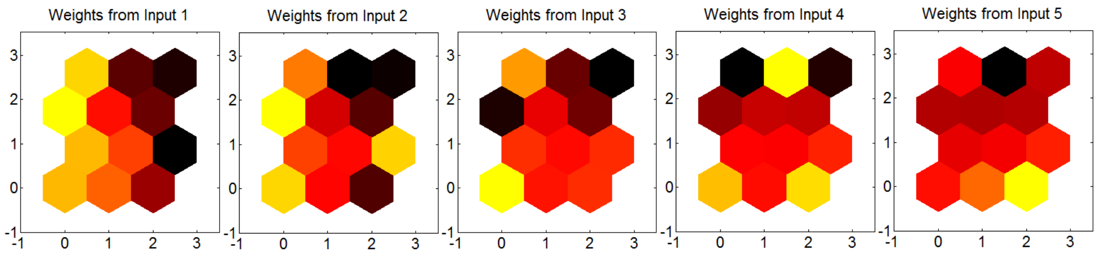

Figure 6-18: Price/ Earnings 3x4 SOM Weight Planes

Table 6-20: Price/ Earnings 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
4	24	1	0.0003	-0.0056	0.0509	-0.0182	-0.0241	0.0344
		2	-0.0228	0.0040	0.1479	-0.0040	0.0228	0.0343
		3	0.0148	0.0003	0.0539	0.0215	0.0070	0.0331
		4	0.0051	-0.0111	0.0524	0.0005	-0.0156	0.0380
		5	0.0028	-0.0019	0.0296	0.0085	0.0038	0.0151
5	92	1	-0.0084	-0.0143	0.0589	-0.0228	-0.0287	0.0551
		2	-0.0382	-0.0114	0.1014	-0.0087	0.0181	0.0540
		3	0.0171	0.0026	0.0483	0.0176	0.0030	0.0314
		4	0.0153	-0.0008	0.0430	0.0098	-0.0063	0.0320
		5	0.0081	0.0034	0.0380	0.0137	0.0091	0.0199
8	53	1	-0.0179	-0.0238	0.0580	-0.0203	-0.0262	0.0414
		2	-0.0217	0.0051	0.0695	-0.0225	0.0043	0.0603
		3	0.0226	0.0081	0.0336	0.0112	-0.0033	0.0280
		4	-0.0045	-0.0207	0.0548	0.0014	-0.0147	0.0392
		5	0.0142	0.0096	0.0196	0.0113	0.0066	0.0235
12**	10	1	-0.0041	-0.0100	0.0549	-0.0094	-0.0152	0.0762
		2	-0.0102	0.0166	0.0815	-0.0184	0.0084	0.1281
		3	0.0227	0.0081	0.0281	0.0036	-0.0109	0.0522
		4	0.0189	0.0028	0.0304	0.0059	-0.0103	0.0638
		5	0.0193	0.0147	0.0370	-0.0023	-0.0070	0.0472

** Not included in statistical analysis

The four clusters shown in Table 6-20 were generated by the PE clustering. All of the clusters achieved poor first year results. Cluster 12, which comprised of companies predominantly from the basic materials sector, did not achieve greater portfolio standard deviations over the five years. The value weighted performance of cluster 12 was however very strong in the final 24 months of the analysis which is more evident in Figure 6-19 (page 85). Apart from cluster 12 there are no significant deviations in financial performance and the clusters chosen for further analysis.

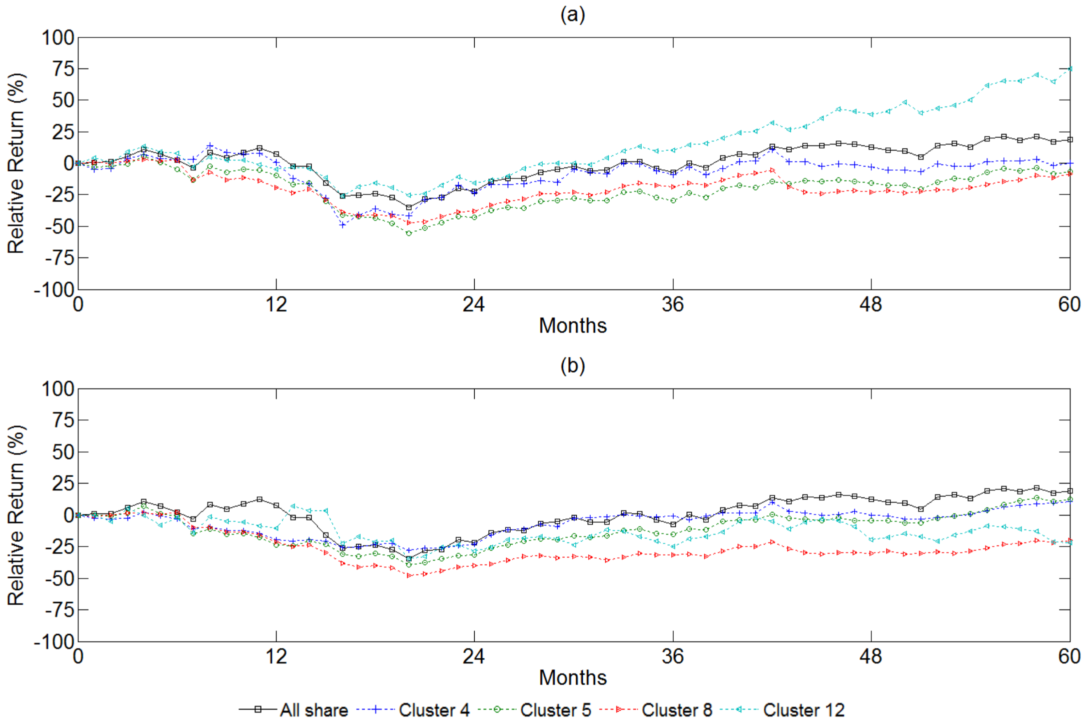

Figure 6-19: Price/ Earnings 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The statistical analysis (Table 6-21) reveals that there is no evidence to support the hypothesis that the clusters achieved different returns, as previously mentioned. Compared to the other clustering tests shown thus far the PE clusters are arguably the most similar with regards to the returns achieved.

Table 6-21: Price/ Earnings 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.731032	0.880842	0.968173	0.904973
Years 1 – 3	0.922865	0.918067	0.619588	0.813328
Years 1 – 5	0.991097	0.890089	0.670656	0.68416

6.2.5 Quick Ratio

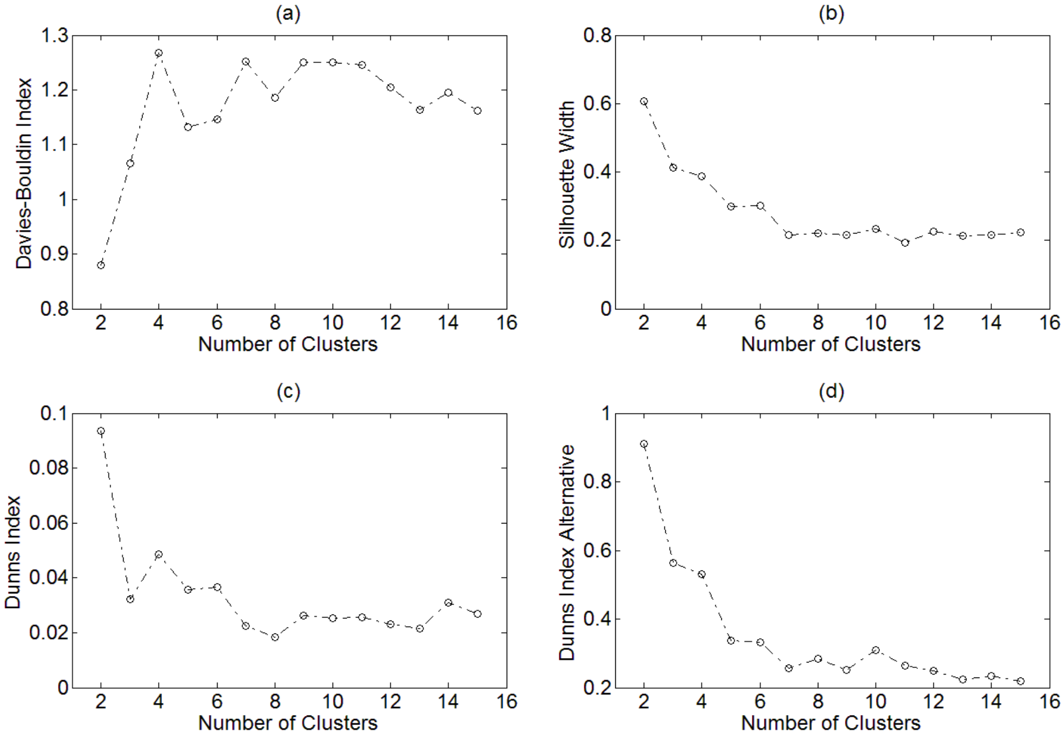

Figure 6-20: Quick Ratio Clustering Validity

The QR clustering achieved a constant trend of decreasing validity with an increase in cluster size (Figure 6-20). As mentioned in the DE clustering the 2x1 SOM has yielded the best validity, however the cluster sizes for this SOM were 40 and 215. The significantly better results achieved by the 2x1 SOM have distorted the scale. From the DB Index and SW it can be seen that the SOMs which produced 12, 13 and 14 clusters were relatively equal with regards to validity. Again it was decided to analyse the 3x4 SOM in more detail.

Table 6-22: Quick Ratio 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1*	7	0.15	Consumer Services	43	1	0	1	0	2
2	11	0.12	Financials	45	0	1	0	1	1
3	11	0.10	Basic Materials	27	0	0	0	2	0
4	13	0.02	Consumer Services	31	0	1	0	0	0
5*	51	0.13	Financials	31	2	4	1	3	0
6*	36	0.37	Financials	36	4	0	0	0	0
7	11	0.39	Financials	64	0	0	3	1	0
8	6	0.20	Financials	50	0	1	0	0	1
9*	63	0.30	Industrials	37	2	2	3	0	3
10	5	0.41	Basic Materials	40	1	1	0	0	0
11	4	0.38	Financials	75	0	1	1	0	0
12*	37	0.16	Industrials	27	1	0	0	1	2
* Investigated in financial analysis									

The results for the 3x4 SOM from the QR clustering are shown in Table 6-22. Analysis of the industry compositions of the clusters shows no conclusive results. The delisting information shows that four out of the seven companies in cluster 1 delisted, which should be investigated in more detail. The cluster sizes are varied however the four largest clusters (5, 6, 9 and 12) are relatively similar in size. Overall these four clusters achieved below average QR inputs (shown in Figure 6-21). Figure 6-21 also shows that the colour map has not been distorted and the neuron colours are well spread across the spectrum.

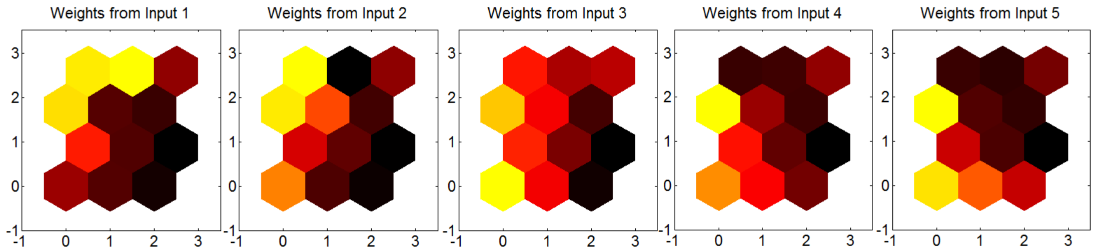

Figure 6-21: Quick Ratio 3x4 SOM Weight Planes

Table 6-23: Quick Ratio 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1**	7	1	-0.0362	-0.0421	0.0765	-0.0256	-0.0315	0.0622
		2	0.0054	0.0322	0.0691	0.0043	0.0311	0.0606
		3	0.0311	0.0166	0.0503	0.0296	0.0150	0.0429
		4	0.0239	0.0078	0.0594	0.0230	0.0068	0.0433
		5	0.0119	0.0072	0.0314	0.0132	0.0085	0.0276
5	51	1	0.0016	-0.0043	0.0605	-0.0237	-0.0296	0.0475
		2	-0.0128	0.0140	0.0654	-0.0155	0.0113	0.0528
		3	0.0035	-0.0110	0.0414	0.0165	0.0020	0.0257
		4	0.0277	0.0116	0.0479	0.0028	-0.0133	0.0409
		5	0.0086	0.0039	0.0282	0.0137	0.0090	0.0176
6	36	1	-0.0057	-0.0116	0.0571	-0.0071	-0.0130	0.0504
		2	-0.0316	-0.0048	0.1189	-0.0141	0.0127	0.0600
		3	0.0236	0.0091	0.0475	0.0112	-0.0034	0.0344
		4	0.0075	-0.0087	0.0375	0.0061	-0.0101	0.0279
		5	0.0142	0.0095	0.0292	0.0055	0.0008	0.0187
9	63	1	0.0086	0.0027	0.0548	-0.0153	-0.0212	0.0463
		2	-0.0480	-0.0212	0.1161	-0.0310	-0.0042	0.0696
		3	0.0146	0.0000	0.0535	0.0124	-0.0022	0.0378
		4	0.0165	0.0004	0.0435	0.0112	-0.0049	0.0325
		5	-0.0051	-0.0097	0.0517	-0.0073	-0.0120	0.0308
12	37	1	-0.0042	-0.0101	0.0628	-0.0216	-0.0274	0.0424
		2	-0.0408	-0.0140	0.1431	-0.0316	-0.0048	0.0510
		3	0.0186	0.0040	0.0398	0.0187	0.0042	0.0359
		4	0.0132	-0.0029	0.0356	0.0079	-0.0082	0.0256
		5	0.0103	0.0056	0.0293	-0.0080	-0.0127	0.0587

** Not included in statistical analysis

The QR clustering financial results in Table 6-23 show that cluster 1 achieved superior returns over years two to five for both the equally weighted and value weighted clusters (Figure 6-22 page 89). Clusters 9 and 12 performed the most poorly of the clusters and this is most evident in the equally weighted results.

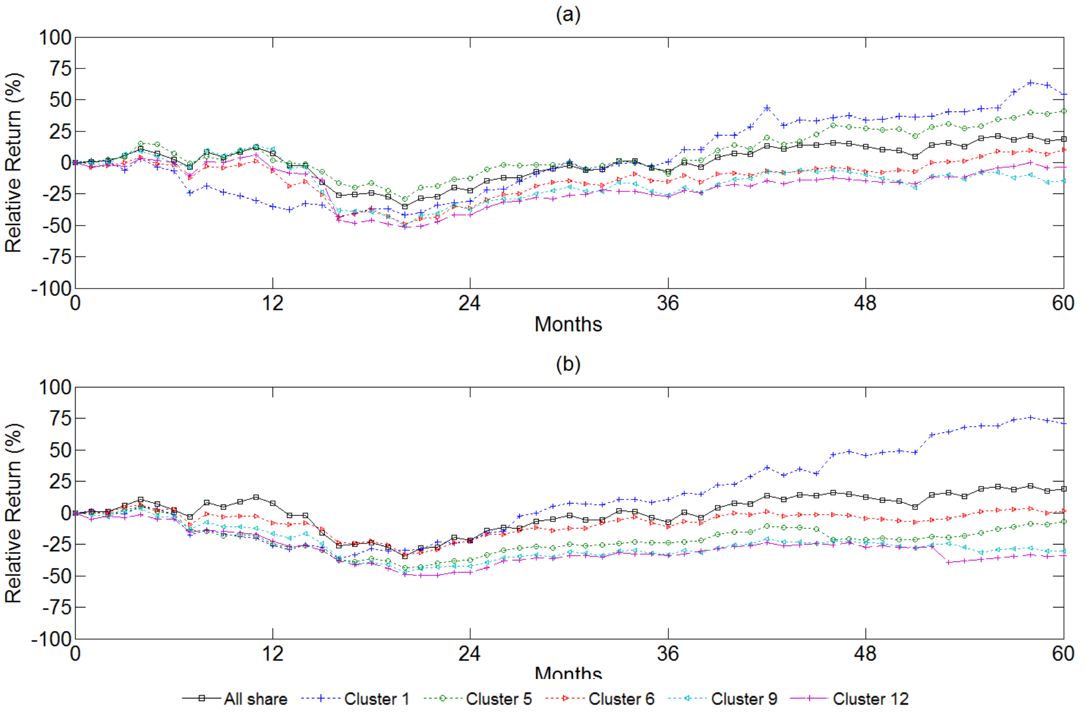

Figure 6-22: Quick Ratio 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

Figure 6-22 shows that overall there is no significant difference in financial performance, excluding cluster 1. Since the statistical analysis only involves the larger clusters it is expected that the results would not yield any statistically significant results and this can be seen in Table 6-24.

Table 6-24: Quick Ratio 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.931718	0.965648	0.822069	0.770223
Years 1 – 3	0.984724	0.948938	0.885212	0.780289
Years 1 – 5	0.913919	0.83181	0.763244	0.753145

6.2.6 Return on Assets

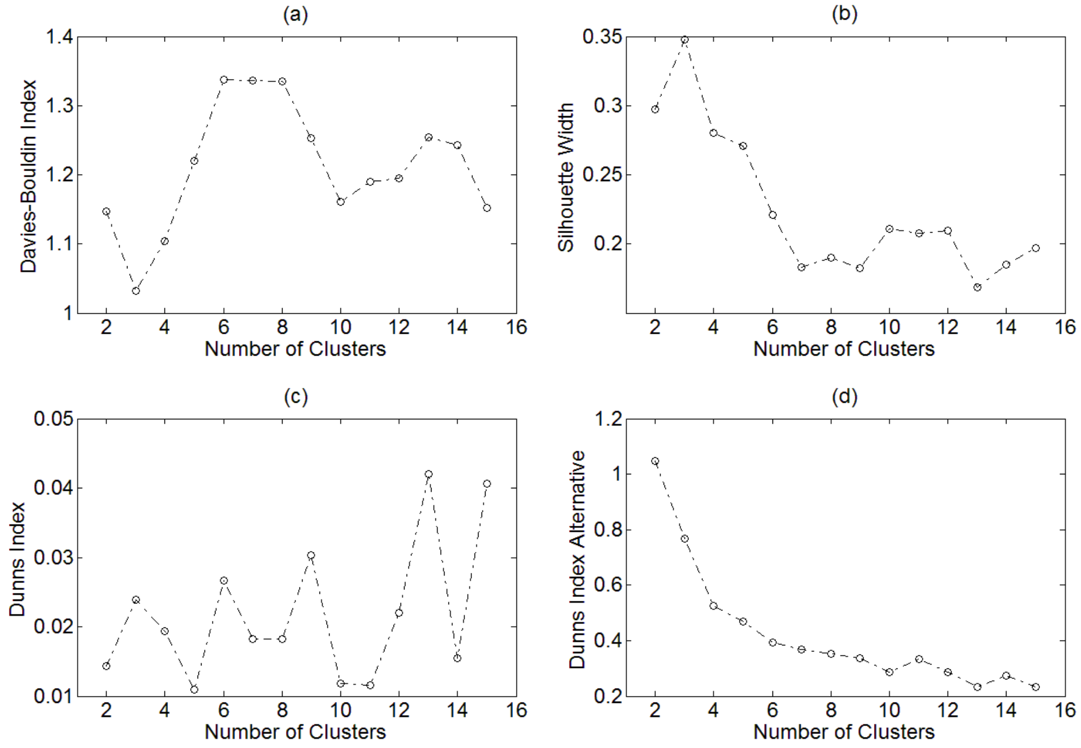

Figure 6-23: Return on Assets Clustering Validity

The clustering for RA (Figure 6-23) shows different results to the other single variable clustering tests. These results show that the 3x1 SOM has outperformed the 2x1 SOM however this small SOM still does not lend itself to further interpretation. As with the previous tests these results show little difference between the results for SOM sizes ranging from 10 to 12 so it was decided to investigate the 3x4 SOM in more detail. This SOM achieved cluster sizes adequate for further analysis.

Table 6-25: Return on Assets 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	11	0.34	Basic Materials	55	0	0	0	0	1
2	8	0.06	Basic Materials	50	1	0	0	0	0
3*	33	0.15	Industrials	30	1	1	1	0	2
4	5	-0.03	Basic Materials	60	0	1	0	0	0
5*	68	0.34	Industrials	28	5	1	0	1	1
6	4	0.15	Consumer Services	25	0	1	1	1	0
7*	1	-	Technology	100	0	0	0	0	0
8*	67	0.23	Financials	45	2	4	1	2	2
9*	32	0.09	Financials	50	1	0	4	3	1
10	5	-0.01	Financials	60	0	0	1	0	0
11	3	0.36	Basic Materials	67	1	0	0	0	0
12	18	0.03	Basic Materials	39	0	3	1	1	2
* Investigated in financial analysis									

The RA clustering achieved both very small and large clusters (Table 6-25). This is the only variable to produce a single firm cluster (neuron 7) for a 3x4 SOM. This firm has been excluded from other clusters because its RA ranged from low to high back to low (Figure 6-24). Of the larger clusters chosen for further analysis (neurons 3, 5, 8 and 9) cluster 3 achieved the greatest RA values and cluster 9 the lowest (Figure 6-24). These clusters and the single firm cluster have been investigated in more detail in Table 6-26.

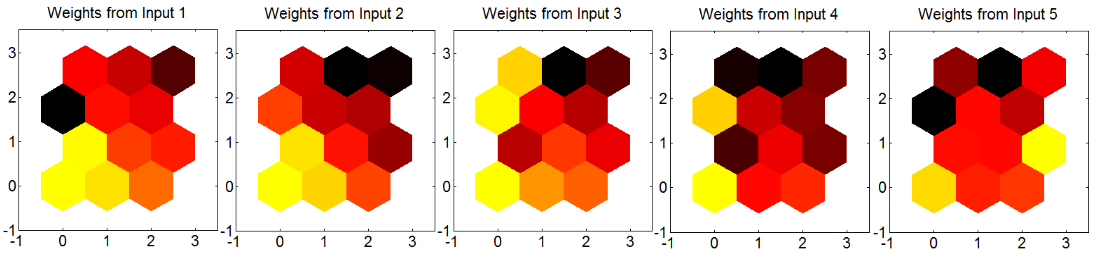

Figure 6-24: Return on Assets 3x4 SOM Weight Planes

Table 6-26: Return on Assets 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
3	33	1	0.0021	-0.0038	0.0716	-0.0171	-0.0229	0.0504
		2	-0.0031	0.0237	0.0578	-0.0109	0.0159	0.0585
		3	0.0119	-0.0026	0.0453	0.0182	0.0037	0.0289
		4	0.0190	0.0029	0.0417	0.0077	-0.0085	0.0424
		5	0.0167	0.0120	0.0288	0.0114	0.0067	0.0163
5	68	1	0.0035	-0.0024	0.0594	-0.0267	-0.0326	0.0513
		2	-0.0495	-0.0227	0.1165	-0.0120	0.0148	0.0537
		3	0.0145	-0.0001	0.0552	0.0183	0.0037	0.0336
		4	0.0196	0.0035	0.0493	0.0101	-0.0060	0.0276
		5	0.0007	-0.0040	0.0456	0.0092	0.0045	0.0248
7**	1	1	0.0664	0.0605	0.2000	0.0664	0.0605	0.2000
		2	-0.0695	-0.0427	0.1478	-0.0695	-0.0427	0.1478
		3	-0.0474	-0.0620	0.1190	-0.0474	-0.0620	0.1190
		4	-0.0152	-0.0313	0.1732	-0.0152	-0.0313	0.1732
		5	0.0261	0.0214	0.0672	0.0261	0.0214	0.0672
8	67	1	-0.0133	-0.0192	0.0510	-0.0180	-0.0239	0.0401
		2	-0.0274	-0.0006	0.1019	-0.0158	0.0110	0.0549
		3	0.0177	0.0031	0.0337	0.0119	-0.0027	0.0284
		4	0.0087	-0.0074	0.0337	0.0006	-0.0155	0.0456
		5	0.0108	0.0061	0.0355	0.0022	-0.0025	0.0420
9	32	1	-0.0312	-0.0371	0.0746	-0.0128	-0.0187	0.0491
		2	-0.0162	0.0106	0.1056	-0.0378	-0.0110	0.0648
		3	0.0207	0.0061	0.0553	0.0091	-0.0055	0.0192
		4	0.0096	-0.0066	0.0468	-0.0038	-0.0199	0.0583
		5	0.0184	0.0138	0.0373	0.0106	0.0059	0.0241

** Not included in statistical analysis

Cluster 7 in Table 6-26 is the only single firm cluster for the RA clustering. This cluster achieved very high first year returns which subsequently dropped significantly before stabilising in year five. As expected this single firm cluster achieved the poorest standard deviation, which is simply the standard deviation of the single firm. Cluster 3 achieved relatively consistent above average performance and is one of the few clusters to have done this. This performance is however only evident in the value weighted results as can be seen in Figure 6-25 (page 93). Cluster 9 achieved poor results in the first two years however

Figure 6-25 shows that these low returns are not significantly different to those achieved by the other clusters. The financial results for clusters 5 and 8 are mixed over the five years.

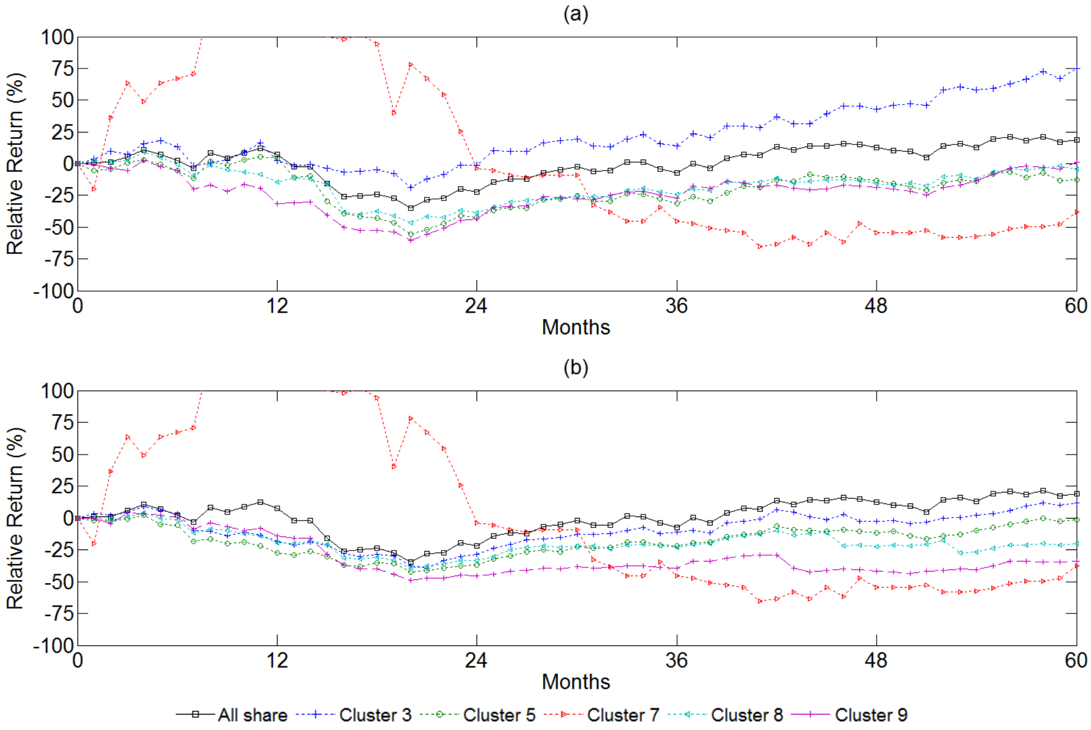

Figure 6-25: Return on Assets 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

Again the statistical analysis shows that the results for the clusters were not significantly different (Table 6-27). Only the value weighted clusters for year one show that there was a possibility of there being a difference in performance. Further analysis of these results shows that no clusters were significantly different, even over this short 12 month period.

Table 6-27: Return on Assets 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.52823	0.507943	0.909897	0.97107
Years 1 – 3	0.845578	0.795427	0.824917	0.775078
Years 1 – 5	0.745445	0.650359	0.707763	0.741686

6.2.7 Return on Equity

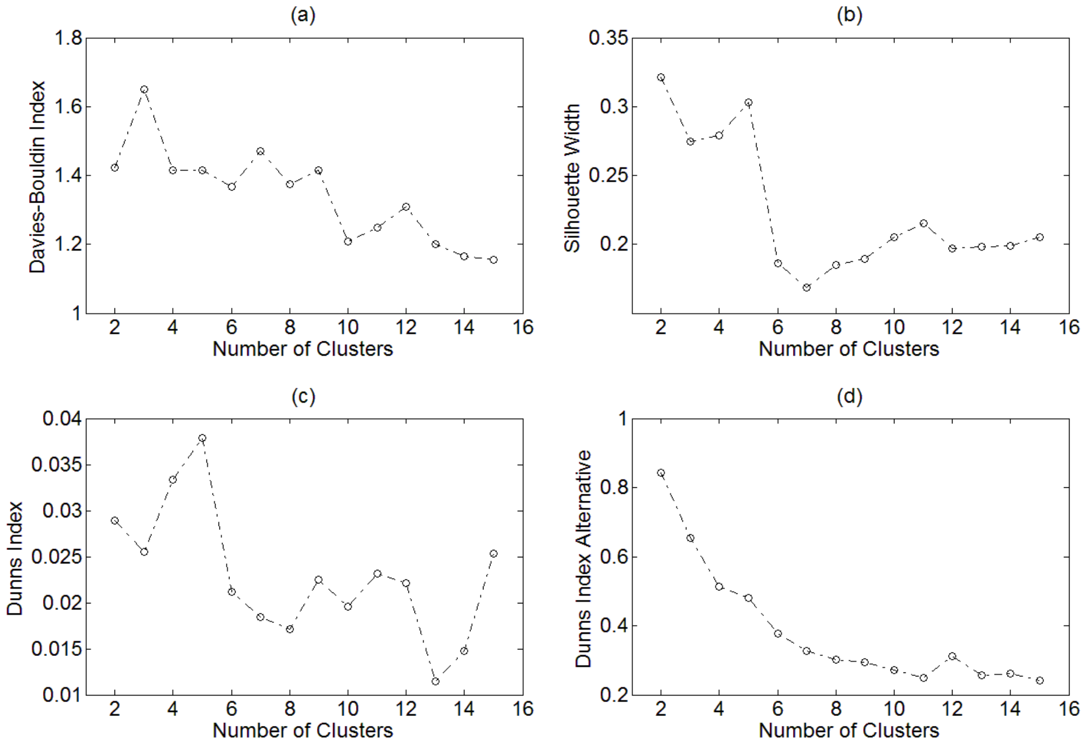

Figure 6-26: Return on Equity Clustering Validity

The validity of the RE clustering (Figure 6-26) reveals results which are consistent with the previous tests. The DB Index (Figure 6-26 (a)) shows that the 2x5 SOM achieved better clustering than similar SOM sizes. This feature is however not apparent in the remaining figures. The 5x1 SOM achieved the best overall clustering however this SOM consists of a single large cluster with 150 companies. Instead it was decided to investigate the 3x4 SOM in more detail.

Table 6-28: Return on Equity 3x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1	10	0.31	Basic Materials	30	1	0	0	0	0
2*	8	-0.01	Basic Materials	38	0	0	2	2	1
3*	3	0.31	Consumer Services	67	2	0	0	0	0
4	15	-0.04	Consumer Services	27	0	1	0	1	0
5*	80	0.29	Industrials	25	2	3	1	1	1
6	5	0.22	Financials	60	0	1	0	0	0
7	11	0.07	Basic Materials	45	1	0	0	0	1
8*	83	0.25	Financials	34	3	3	4	3	4
9	10	-0.05	Financials	60	1	0	1	0	0
10	13	0.01	Basic Materials	31	0	1	0	0	0
11	13	0.00	Basic Materials	38	1	1	1	1	2
12	4	0.34	Oil and Gas	25	0	1	0	0	0

* Investigated in financial analysis

The general information for the clustering completed with RE is shown in Table 6-28. The first notable feature is that cluster 2, which was comprised of only eight companies, had five delistings. Cluster 3 had two of its companies delist in the first year, leaving only one firm. Two clusters were of adequate size to consider for financial analysis (clusters 5 and 8). These clusters achieved very similar average inputs over the five year period (Figure 6-27). In comparison cluster 2 and 3's RE performance decreased going from right to left in Figure 6-27.

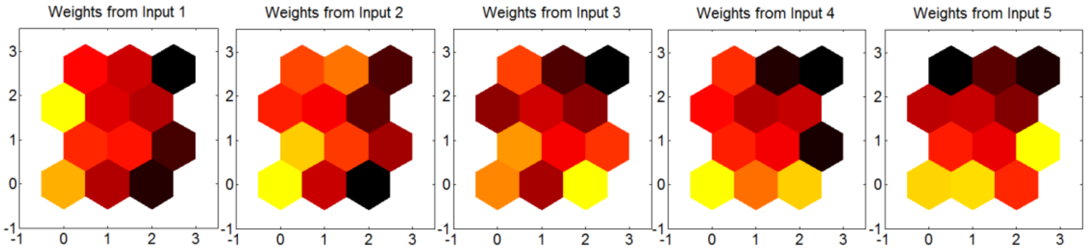

Figure 6-27: Return on Equity 3x4 SOM Weight Planes

Table 6-29: Return on Equity 3x4 SOM Financial Information

NRN	CMP	Year	Value Weighting Mean Monthly	Value Weighting Mean Excess	Value Weighting Std Dev	Equal Weighting Mean Monthly	Equal Weighting Mean Excess	Equal Weighting Std Dev
2**	8	1	-0.0280	-0.0338	0.1031	-0.0325	-0.0384	0.0673
		2	-0.0957	-0.0689	0.2673	-0.0377	-0.0109	0.0783
		3	0.0066	-0.0080	0.0947	-0.0113	-0.0258	0.0597
		4	-0.0055	-0.0216	0.0829	-0.0660	-0.0821	0.2725
		5	-0.0376	-0.0423	0.1246	-0.0490	-0.0537	0.2286
3**	3	1	-0.0129	-0.0188	0.0512	-0.0105	-0.0164	0.0434
		2	0.0202	0.0470	0.0508	0.0202	0.0470	0.0508
		3	0.0248	0.0103	0.0433	0.0248	0.0103	0.0433
		4	-0.0012	-0.0173	0.0640	-0.0012	-0.0173	0.0640
		5	0.0103	0.0056	0.0534	0.0103	0.0056	0.0534
5	80	1	-0.0057	-0.0115	0.0587	-0.0274	-0.0333	0.0536
		2	-0.0387	-0.0119	0.1036	-0.0070	0.0198	0.0546
		3	0.0159	0.0013	0.0477	0.0199	0.0054	0.0336
		4	0.0149	-0.0013	0.0424	0.0079	-0.0082	0.0268
		5	0.0071	0.0024	0.0379	0.0123	0.0076	0.0226
8	83	1	-0.0133	-0.0192	0.0414	-0.0181	-0.0240	0.0373
		2	-0.0381	-0.0113	0.1060	-0.0235	0.0033	0.0536
		3	0.0169	0.0024	0.0322	0.0125	-0.0021	0.0267
		4	0.0098	-0.0064	0.0313	0.0053	-0.0109	0.0333
		5	0.0064	0.0017	0.0401	0.0033	-0.0014	0.0343

** Not included in statistical analysis

The financial results for the RE clustering could only be extended to four clusters, of which only two are of an appropriate size to use in a statistical analysis (as shown in Table 6-29). Cluster 2 had a large number of companies delisting and the financial results of the companies within this cluster are very poor. In addition to this the portfolio standard deviations were poor due to the limited number of companies in this cluster. Cluster 3 achieved a better portfolio standard deviation than cluster 2 even though it was comprised of fewer companies. Figure 6-28 (page 97) shows that the performance of the smaller clusters (clusters 2 and 3) varied the most, however the larger clusters had similar performance with cluster 5 only slightly outperforming cluster 8.

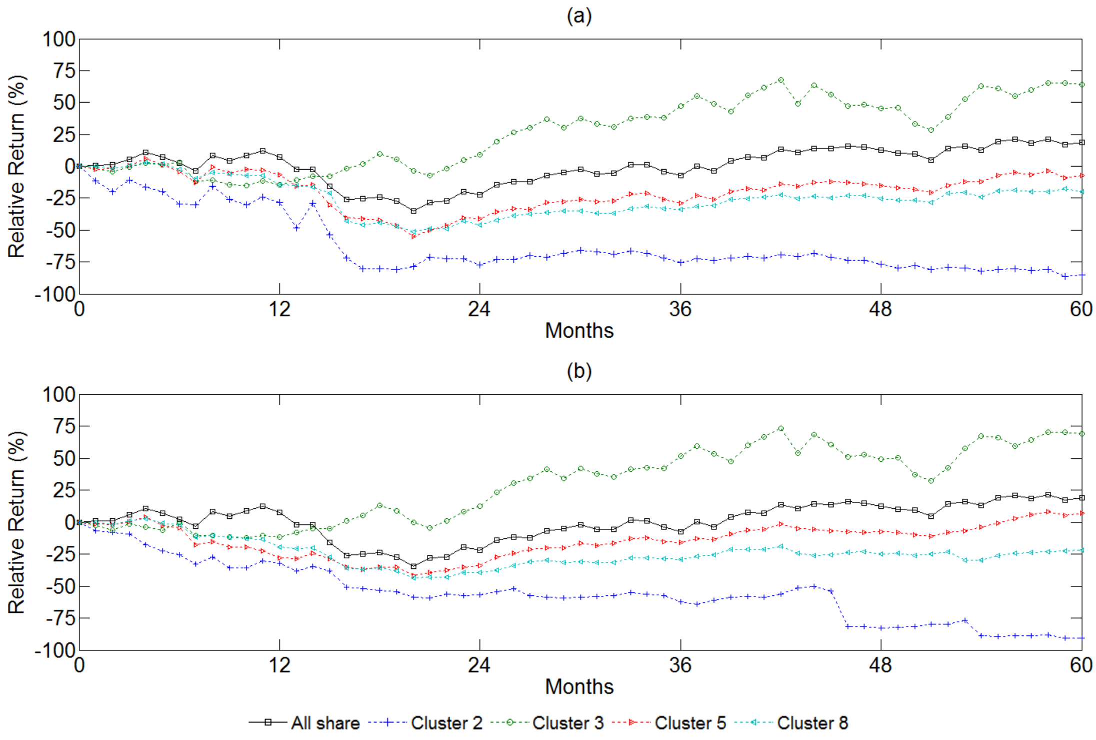

Figure 6-28: Return on Equity 3x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

When considering the value weighted statistical results it can be seen that the returns of these two clusters can be considered very similar. The equally weighted clusters show a greater indication that different returns could have been achieved however it is still not statistically significant.

Table 6-30: Return on Equity 3x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.717066	1	0.625495	0.77283
Years 1 – 3	0.90912	0.946125	0.660393	0.443774
Years 1 – 5	0.835089	0.883162	0.484287	0.352884

6.2.8 Market Capitalisation

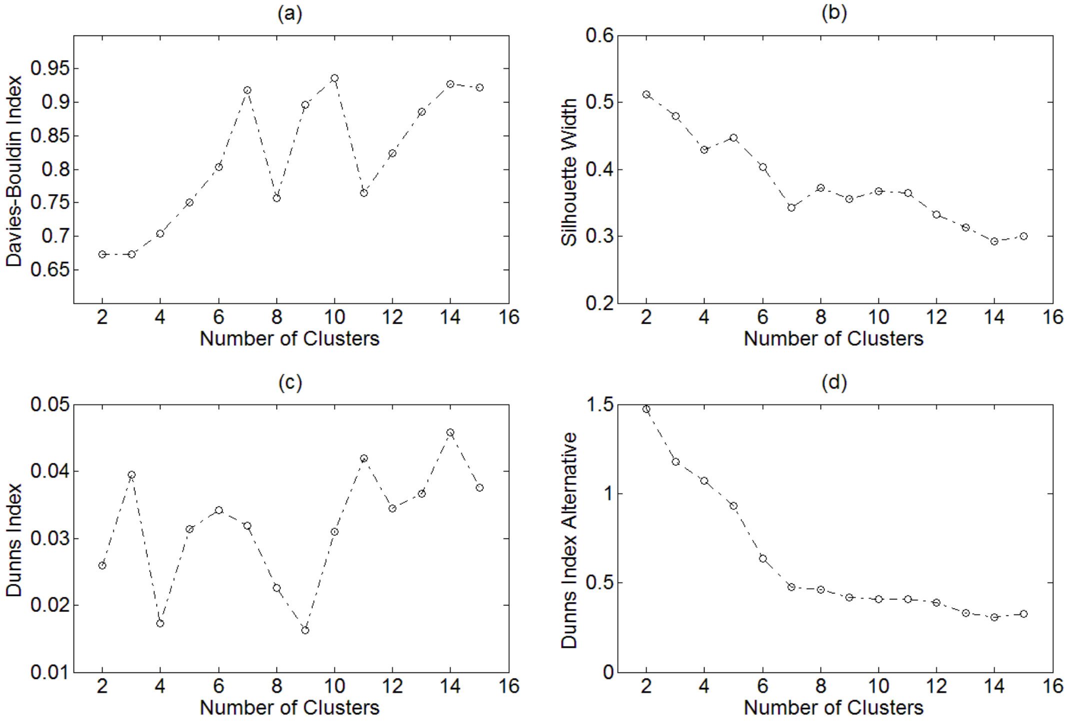

Figure 6-29: Market Capitalisation Clustering Validity

MC is the first variable chosen which does not depend on reported financials. Due to the nature of this data no outliers were present when taking $ln(1 + MC)$ and the overall clustering validity was an improvement over the previous tests. Analysis of all the validity measures in Figure 6-29 reveals mixed results however it can be seen that the 2x4 SOM achieved better validity than similar size clusters. This is most evident with the DB Index and SW (Figure 6-29 (a) and (b)) results and for this reason the 2x4 SOM has been investigated in more detail.

Table 6-31: Market Capitalisation 2x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1*	3	0.58	Financials	67	0	0	0	0	0
2*	29	0.49	Basic Materials	38	0	1	0	0	0
3*	40	0.31	Basic Materials	23	0	3	2	2	1
4*	42	0.56	Consumer Services	29	2	0	1	1	0
5*	39	0.30	Financials	38	5	0	0	1	2
6*	31	0.30	Industrials	39	2	1	0	1	2
7*	39	0.34	Financials	31	1	1	3	2	3
8*	32	0.26	Financials	31	1	5	3	1	1
* Investigated in financial analysis									

The MC clustering achieved the most even cluster sizes of all the tests and this is apparent in Table 6-31. Only one cluster was significantly smaller and this can be attributed to the companies only listing after the first input year. The SW values for these clusters is very high and cluster 2 is comprised of the largest companies as shown by the yellow neuron in Figure 6-30. Clusters 7 and 8 consist of smaller companies, which can be seen by the dark neurons in Figure 6-30. Due to the consistency in the cluster sizes it was decided to take all these clusters forward for financial analysis in Table 6-32.

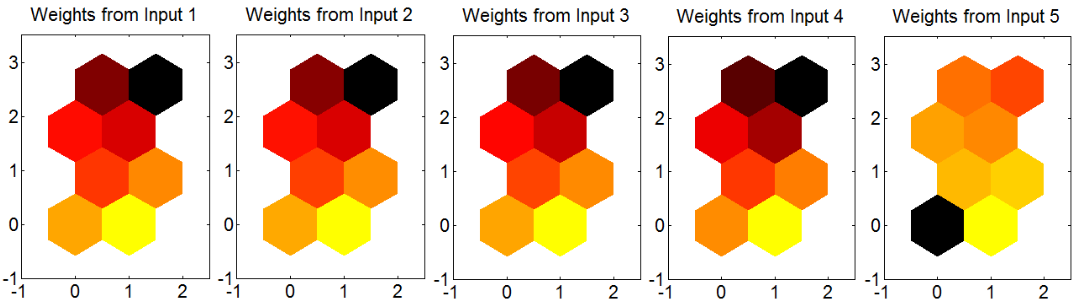

Figure 6-30: Market Capitalisation 2x4 SOM Weight Planes

Table 6-32: Market Capitalisation 2x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1**	3	1	-0.0255	-0.0314	0.0561	-0.0356	-0.0415	0.0527
		2	-0.0580	-0.0312	0.1733	-0.0158	0.0110	0.1048
		3	0.0159	0.0013	0.0739	0.0280	0.0135	0.0639
		4	0.0112	-0.0050	0.0439	0.0235	0.0073	0.0504
		5	-0.0184	-0.0231	0.0447	-0.0080	-0.0127	0.0413
2	29	1	0.0067	0.0008	0.0528	-0.0067	-0.0126	0.0491
		2	-0.0424	-0.0156	0.1144	-0.0285	-0.0017	0.0893
		3	0.0137	-0.0008	0.0482	0.0144	-0.0002	0.0440
		4	0.0152	-0.0009	0.0404	0.0149	-0.0012	0.0371
		5	0.0022	-0.0025	0.0449	0.0045	-0.0002	0.0375
3	40	1	-0.0165	-0.0224	0.0564	-0.0138	-0.0197	0.0422
		2	-0.0207	0.0061	0.0935	-0.0148	0.0120	0.0797
		3	0.0215	0.0070	0.0342	0.0211	0.0065	0.0331
		4	0.0058	-0.0103	0.0428	0.0048	-0.0113	0.0315
		5	0.0190	0.0143	0.0218	0.0126	0.0079	0.0156
4	42	1	-0.0167	-0.0226	0.0615	-0.0181	-0.0240	0.0592
		2	-0.0045	0.0223	0.0710	0.0005	0.0273	0.0571
		3	0.0226	0.0080	0.0425	0.0220	0.0075	0.0385
		4	0.0144	-0.0017	0.0436	0.0099	-0.0062	0.0356
		5	0.0139	0.0092	0.0302	0.0108	0.0061	0.0276
5	39	1	-0.0234	-0.0293	0.0654	-0.0294	-0.0353	0.0534
		2	-0.0349	-0.0081	0.0999	-0.0185	0.0083	0.0706
		3	0.0278	0.0132	0.0378	0.0209	0.0064	0.0304
		4	0.0149	-0.0012	0.0289	-0.0046	-0.0208	0.0476
		5	-0.0246	-0.0293	0.1343	-0.0092	-0.0139	0.0541
6	31	1	-0.0255	-0.0314	0.0499	-0.0189	-0.0248	0.0484
		2	-0.0310	-0.0042	0.0712	-0.0241	0.0027	0.0586
		3	0.0045	-0.0100	0.0379	0.0201	0.0056	0.0333
		4	-0.0532	-0.0694	0.1489	-0.0193	-0.0354	0.0639
		5	0.0111	0.0065	0.0237	0.0080	0.0033	0.0273
7	39	1	-0.0168	-0.0227	0.0367	-0.0211	-0.0270	0.0328
		2	-0.0581	-0.0313	0.1098	-0.0337	-0.0069	0.0530
		3	0.0025	-0.0121	0.0260	-0.0024	-0.0170	0.0308
		4	0.0230	0.0069	0.0296	0.0168	0.0007	0.0288
		5	0.0065	0.0018	0.0378	-0.0130	-0.0177	0.0671

8	32	1	-0.0552	-0.0611	0.0585	-0.0236	-0.0295	0.0517
		2	-0.0568	-0.0300	0.0629	-0.0211	0.0057	0.0322
		3	-0.0445	-0.0591	0.0645	-0.0143	-0.0289	0.0314
		4	-0.0094	-0.0255	0.0460	-0.0032	-0.0193	0.0337
		5	0.0157	0.0110	0.0676	0.0125	0.0078	0.0281

** Not included in statistical analysis

The financial results for all the clusters generated from the MC are shown in Table 6-32. For these financial results it is expected that the value weighted and equally weighted returns are similar due to the clustering being dependent on MC. The results for the two different weightings are similar with no significant differences, as shown in Figure 6-31. Since the clusters range from large to small, with an increase in cluster number, it was expected that a trend in either increasing or decreasing performance would become apparent if there was a relationship between the firm size and performance. This is not the case and the performance of the clusters does not follow any trend related to the cluster numbers.

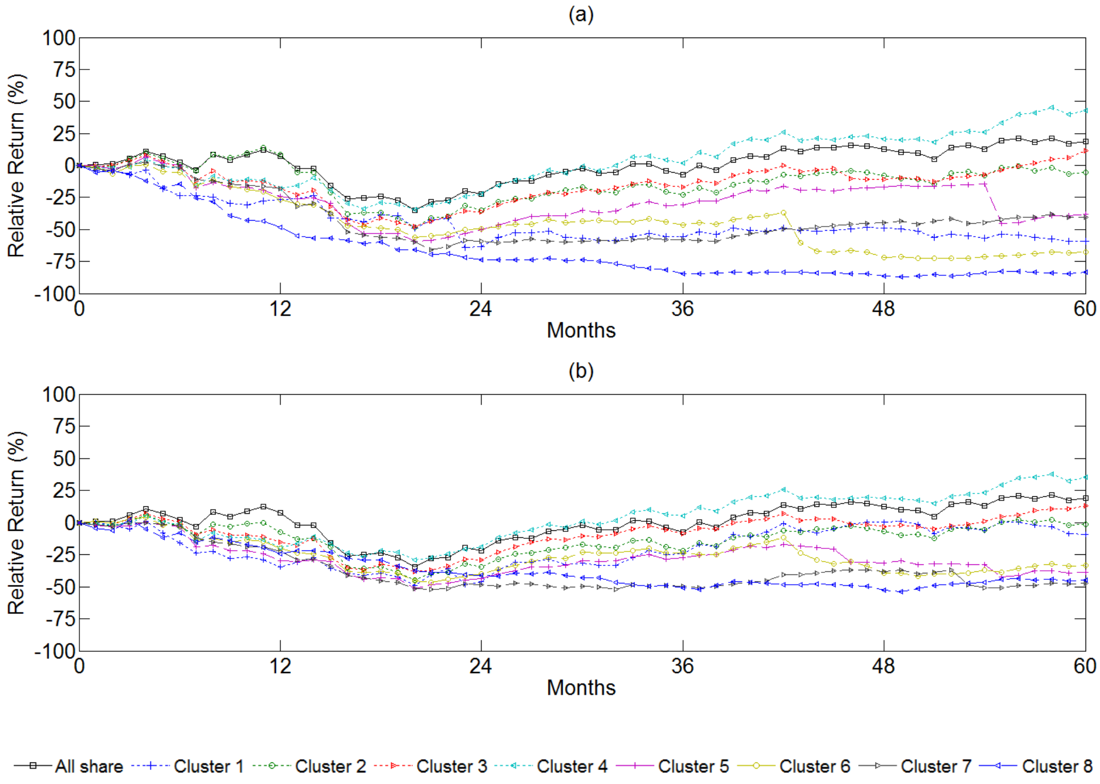

Figure 6-31: Market Capitalisation 2x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

Results from the statistical analysis (Table 6-33) show that it can be concluded that there were different results over the three and five year periods. Over the five year period the ANOVA test was not significant within a 5% level. However there appears to be strong evidence to support the hypothesis that the returns achieved were different over the three year period. It is important to note that this is not evident with the equally weighted results, as shown in Table 6-33.

Table 6-33: Market Capitalisation 2x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.256006	0.349016	0.952611	0.919157
Years 1 – 3	0.025114	0.004805	0.558449	0.125307
Years 1 – 5	0.056844	0.013072	0.416887	0.16882

To support the statistical results found in Table 6-33, for the value weighted portfolios over years 1-3, a Tukey HSD post hoc analysis was completed. The results of this test can be found in Table 6-34 where the most significant difference in cluster performance occurred between clusters 4 and 8. Apart from these two clusters there are no other statistically significant results for α = 0.05. Cluster 8 does have significance values near 0.05 with several of the other clusters, so it has been assumed that cluster 8 is critical in highlighting a market capitalisation trend. Referring to Figure 6-30 (page 99) reveals that this cluster was comprised of the smallest companies. Analysis of the companies in this cluster shows that there wasn't a single firm with a significantly higher market capitalisation hence cluster 8's unique performance cannot entirely be attributed to a single poor performing firm.

Table 6-34: Market Capitalisation 2x4 Years 1-3 Tukey HSD Results

Portfolio		Mean Difference (I-J)	Significance	95% Confidence Interval	
I	J			Lower Bound	Upper Bound
2	3	-0.0021	1.0000	-0.0495	0.0452
	4	-0.0078	0.9990	-0.0552	0.0396
	5	0.0028	1.0000	-0.0445	0.0502
	6	0.0100	0.9959	-0.0374	0.0574
	7	0.0168	0.9401	-0.0305	0.0642
	8	0.0448	0.0767	-0.0025	0.0922

3	2	0.0021	1.0000	-0.0452	0.0495
	4	-0.0057	0.9998	-0.0530	0.0417
	5	0.0050	0.9999	-0.0424	0.0523
	6	0.0121	0.9883	-0.0352	0.0595
	7	0.0189	0.8978	-0.0284	0.0663
	8	0.0470	0.0536	-0.0004	0.0943
4	2	0.0078	0.9990	-0.0396	0.0552
	3	0.0057	0.9998	-0.0417	0.0530
	5	0.0106	0.9942	-0.0367	0.0580
	6	0.0178	0.9225	-0.0296	0.0652
	7	0.0246	0.7171	-0.0227	0.0720
	8	0.0526	0.0186***	0.0053	0.1000
5	2	-0.0028	1.0000	-0.0502	0.0445
	3	-0.0050	0.9999	-0.0523	0.0424
	4	-0.0106	0.9942	-0.0580	0.0367
	6	0.0072	0.9994	-0.0402	0.0545
	7	0.0140	0.9755	-0.0334	0.0614
	8	0.0420	0.1195	-0.0053	0.0894
6	2	-0.0100	0.9959	-0.0574	0.0374
	3	-0.0121	0.9883	-0.0595	0.0352
	4	-0.0178	0.9225	-0.0652	0.0296
	5	-0.0072	0.9994	-0.0545	0.0402
	7	0.0068	0.9995	-0.0405	0.0542
	8	0.0348	0.3062	-0.0125	0.0822
7	2	-0.0168	0.9401	-0.0642	0.0305
	3	-0.0189	0.8978	-0.0663	0.0284
	4	-0.0246	0.7171	-0.0720	0.0227
	5	-0.0140	0.9755	-0.0614	0.0334
	6	-0.0068	0.9995	-0.0542	0.0405
	8	0.0280	0.5770	-0.0193	0.0754
8	2	-0.0448	0.0767	-0.0922	0.0025
	3	-0.0470	0.0536	-0.0943	0.0004
	4	-0.0526	0.0186***	-0.1000	-0.0053
	5	-0.0420	0.1195	-0.0894	0.0053
	6	-0.0348	0.3062	-0.0822	0.0125
	7	-0.0280	0.5770	-0.0754	0.0193
*** Statistically significant for α = 0.05					

6.2.9 Volatility

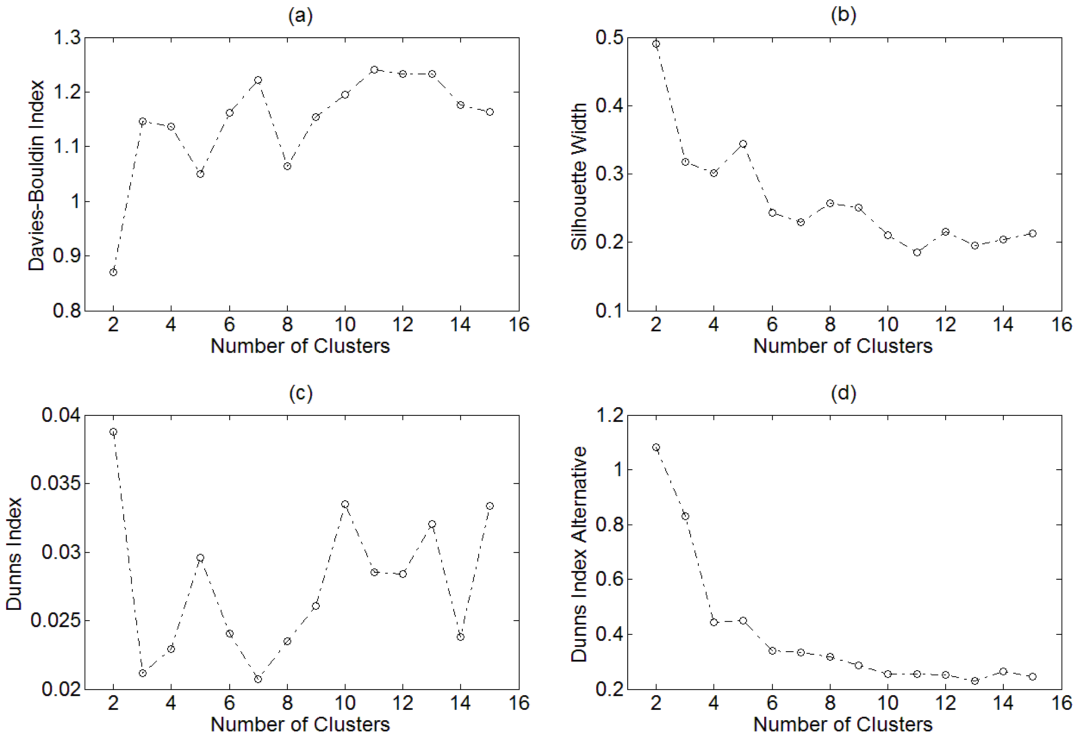

Figure 6-32: Volatility Clustering Validity

As with the MC clustering the V clustering also does not depend on reported financials. This variable is dependent on price movements and the clustering was completed using $ln(1 + v)$. Again no outliers were present and this variable was still able to achieve relatively even cluster sizes. The 2x4 SOM appears to have achieved the more compact clusters than similar SOM sizes (Figure 6-32). The 2x4 SOM also produced clusters which could be analysed in more detail and was therefore chosen for financial analysis.

Table 6-35: Volatility 2x4 SOM General Information

NRN	CMP	SW	Industry Composition		Delisted (Per Year)				
			Industry	Cluster (%)	1	2	3	4	5
1*	46	0.24	Industrials	28	1	2	0	1	4
2	3	0.38	Financials	67	1	0	0	0	0
3*	25	0.29	Basic Materials	32	1	0	4	2	2
4*	73	0.29	Financials	32	3	2	1	1	1
5*	21	0.18	Financials	29	0	4	1	0	1
6*	73	0.23	Financials	26	4	3	2	3	1
7	6	0.29	Financials	50	0	0	1	0	0
8	8	0.27	Financials	63	1	0	0	1	0
* Investigated in financial analysis									

The V clustering general results are shown in Table 6-35. Although this variable was not greatly affected by outliers, the cluster sizes are still not as even as the MC clusters. Overall the SWs are consistent for all the clusters. The industry composition and delisting data does not show any significant anomalies which may be considered for further analysis. For this reason only the four largest clusters (1, 3, 4, 5 and 6) have been included in the financial analysis. The plane weights in Figure 6-33 show that the small clusters are comprised of companies with inconsistent weights with many of the neurons changing from brown to yellow.

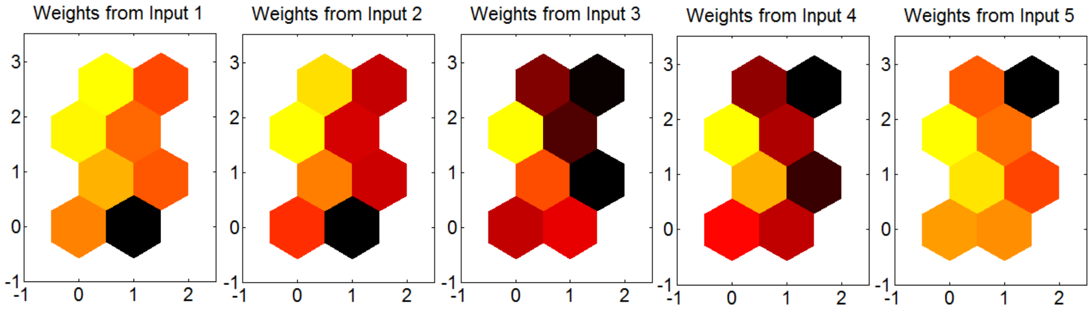

Figure 6-33: Volatility 2x4 SOM Weight Planes

Table 6-36: Volatility 2x4 SOM Financial Information

NRN	CMP	Year	Value Weighting			Equal Weighting		
			Mean Monthly	Mean Excess	Std Dev	Mean Monthly	Mean Excess	Std Dev
1	46	1	-0.0100	-0.0159	0.0662	-0.0180	-0.0239	0.0430
		2	-0.0216	0.0052	0.1089	-0.0292	-0.0024	0.0650
		3	0.0131	-0.0015	0.0380	0.0166	0.0020	0.0280
		4	0.0026	-0.0135	0.0456	-0.0003	-0.0164	0.0473
		5	-0.0197	-0.0244	0.0931	-0.0045	-0.0092	0.0536
3	25	1	-0.0246	-0.0305	0.0715	-0.0202	-0.0261	0.0397
		2	-0.0431	-0.0163	0.0827	-0.0321	-0.0053	0.0583
		3	-0.0012	-0.0157	0.0396	-0.0094	-0.0240	0.0441
		4	0.0157	-0.0004	0.0271	0.0126	-0.0036	0.0295
		5	0.0165	0.0118	0.0393	0.0083	0.0036	0.0489
4	73	1	-0.0187	-0.0246	0.0578	-0.0169	-0.0228	0.0476
		2	-0.0059	0.0209	0.0645	0.0012	0.0280	0.0459
		3	0.0210	0.0065	0.0414	0.0217	0.0072	0.0294
		4	0.0124	-0.0037	0.0373	0.0136	-0.0025	0.0271
		5	0.0205	0.0159	0.0273	0.0086	0.0039	0.0330
5	21	1	-0.0488	-0.0546	0.0840	-0.0225	-0.0284	0.0626
		2	-0.0458	-0.0190	0.1185	-0.0341	-0.0073	0.0280
		3	-0.0507	-0.0653	0.0763	-0.0094	-0.0239	0.0255
		4	-0.1000	-0.1161	0.3579	-0.0166	-0.0327	0.0906
		5	0.0013	-0.0034	0.0570	-0.0057	-0.0104	0.0495
6	73	1	0.0125	0.0066	0.0584	-0.0184	-0.0243	0.0490
		2	-0.0486	-0.0218	0.1280	-0.0278	-0.0010	0.0836
		3	0.0130	-0.0016	0.0526	0.0188	0.0042	0.0389
		4	0.0164	0.0002	0.0456	-0.0009	-0.0170	0.0465
		5	-0.0032	-0.0079	0.0489	0.0016	-0.0031	0.0224
** Not included in statistical analysis								

The five clusters chosen for financial analysis from the V clustering are shown in Table 6-36 where is can be seen that cluster 5 performed poorly considering both the equally weighted and value weighted returns. The performance of cluster 4 was found to be above that of the benchmark over years two and three, making it the only cluster out of the volatility tests to do so. These features can be seen again in Figure 6-34 (page 107) along with the relatively average performance of clusters 1 and 6.

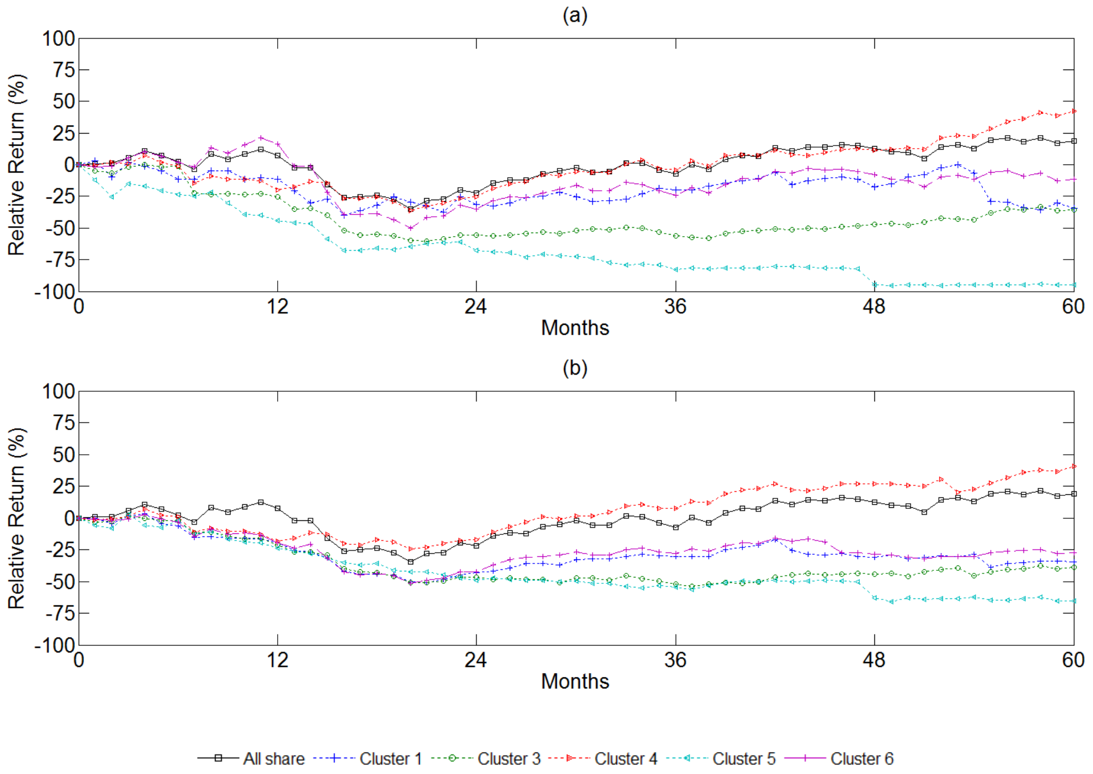

Figure 6-34: Volatility 2x4 SOM Financial Performance Comparison (a) Value Weighting (b) Equally Weighting

The statistical analysis was completed using all the clusters from Table 6-36 and the results can be seen in Table 6-37. For the statistical analysis there is a notable difference between the Kruskal Wallis and ANOVA results. For the 6 statistical tests there are 3 cases where one of the tests is able to reject the null hypothesis while the other is not. Due to the difference in results it was decided to incorporate visual analysis of Figure 6-34 to determine which weighting and time frames would be most appropriate for further analysis. Using this approach it was decided to investigate both the 1-3 and 1-5 year time frames from the value weighted portfolios in more detail.

Table 6-37: Volatility 2x4 SOM Statistical Results

Period	Value Weighting		Equal Weighting	
	ANOVA Significance	Kruskal Wallis Significance	ANOVA Significance	Kruskal Wallis Significance
Year 1	0.290457	0.332622	0.998844	0.960542
Years 1 – 3	0.068745	0.097231	0.239352	0.034302
Years 1 – 5	0.023359	0.105974	0.144957	0.042915

For the 1-3 year time frame the variance of the clusters was found to be equal and the Tukey post-hoc results have been presented in Table 6-38. Here it can be seen that the greatest difference arises from clusters 4 and 5.

Table 6-38: Volatility 2x4 Years 1-3 Tukey HSD Results

Portfolio		Mean Difference	Significance	95% Confidence Interval	
I	J	(I-J)		Lower Bound	Upper Bound
1	3	0.01679	0.88855	-0.03347	0.06705
	4	-0.00496	0.99879	-0.05522	0.04530
	5	0.04227	0.14404	-0.00799	0.09253
	6	0.00156	0.99999	-0.04870	0.05181
3	1	-0.01679	0.88855	-0.06705	0.03347
	4	-0.02175	0.75544	-0.07201	0.02851
	5	0.02548	0.63024	-0.02478	0.07574
	6	-0.01524	0.91921	-0.06549	0.03502
4	1	0.00496	0.99879	-0.04530	0.05522
	3	0.02175	0.75544	-0.02851	0.07201
	5	0.04723	0.07657	-0.00303	0.09749
	6	0.00652	0.99648	-0.04374	0.05678
5	1	-0.04227	0.14404	-0.09253	0.00799
	3	-0.02548	0.63024	-0.07574	0.02478
	4	-0.04723	0.07657	-0.09749	0.00303
	6	-0.04071	0.17256	-0.09097	0.00955
6	1	-0.00156	0.99999	-0.05181	0.04870
	3	0.01524	0.91921	-0.03502	0.06549
	4	-0.00652	0.99648	-0.05678	0.04374
	5	0.04071	0.17256	-0.00955	0.09097

For the five year time frame the clusters were found to have different variance so instead the Games Howell test was used (Table 6-39 page 109). This has been done according to the recommendation by Lomax [140] for n > 50. Again this test reveals that the greatest difference occurs between clusters 4 and 5. However, unlike the ANOVA results these are not statistically significant.

Table 6-39: Volatility 2x4 Years 1-5 Games Howell Results

Portfolio		Mean Difference	Significance	95% Confidence Interval	
I	J	(I-J)		Lower Bound	Upper Bound
1	3	0.0002	1.0000	-0.0337	0.0341
	4	-0.0130	0.7875	-0.0447	0.0188
	5	0.0417	0.4390	-0.0267	0.1101
	6	-0.0051	0.9957	-0.0428	0.0326
3	1	-0.0002	1.0000	-0.0341	0.0337
	4	-0.0132	0.6691	-0.0406	0.0142
	5	0.0415	0.4161	-0.0252	0.1081
	6	-0.0053	0.9926	-0.0396	0.0289
4	1	0.0130	0.7875	-0.0188	0.0447
	3	0.0132	0.6691	-0.0142	0.0406
	5	0.0547	0.1467	-0.0110	0.1203
	6	0.0079	0.9600	-0.0242	0.0400
5	1	-0.0417	0.4390	-0.1101	0.0267
	3	-0.0415	0.4161	-0.1081	0.0252
	4	-0.0547	0.1467	-0.1203	0.0110
	6	-0.0468	0.3228	-0.1153	0.0217
6	1	0.0051	0.9957	-0.0326	0.0428
	3	0.0053	0.9926	-0.0289	0.0396
	4	-0.0079	0.9600	-0.0400	0.0242
	5	0.0468	0.3228	-0.0217	0.1153

7. DISCUSSION

A detailed discussion, with regards to the results shown in Section 6, has been presented below. The discussion first looks at the data mining approach and its application in this research (Section 7.1), thereafter the different clustering methodologies have been discussed in Sections 7.2 and 7.3. Finally a discussion with regards to EMH and this study is completed in Section 7.4.

7.1 Data Mining and the Self-Organising Map

Since the data analysis was intended to be done as a data mining task it is important to review the steps taken and determine whether the desired data mining approach was achieved.

By referring back to the definition of data mining (Section 2.5) it can be seen that there are a few key aspects to data mining. First, the process must reveal potentially useful information and second it must be an automatic (or a semi-automatic) process. The remaining sections of this discussion will investigate whether the results produced have revealed any potentially useful information. With regards to the process being automated, it can be said that the majority of the steps taken (by designing an OOP system and performing numerous automated tests) have resulted in the overall process requiring little human intervention, except when analysing the final results.

Further analysis of the CRISP-DM method (Section 2.5.1) shows that six steps need to be followed in data mining. These six steps have been followed and although the results presented may appear to have been completed in a single attempt it should be noted that only the final results have been included. To reach these results numerous iterations at many of the stages were necessary, especially for the data understanding, data preparation and model building.

With regards to the SOM, it was necessary to use aggressive data pre-processing techniques to handle outliers in the data. Although previous research [4], [6], [7] has not been done in this manner it is important to note that this was required due to the presence of outliers. The impact of outliers is not unique to this study and Sian and Kelvin [7] found that outliers affected the visual analysis of components planes when using 470 stocks from the S&P 500. It is likely that the study presented here is comprised of more outliers for two reasons: the stock market is less developed and the chosen time frame was extremely volatile. In particular the less developed stock market (with many small firms) is considered to be an issue because these smaller companies are not necessarily information efficient and can therefore have reported financials which lie outside of the normal range. Although the study could have been completed with less firms, which were more developed, it was decided that the inclusion of smaller firms, with irregular behaviour, would be beneficial. Evidence of the importance of the inclusion of smaller companies was found by Banz [52].

Wang [6] produced uneven cluster sizes when clustering with turnover and price/ earnings as input variables. It was found that a similar outcome was produced in this research if the data was simply normalised between -1 and +1, without any additional pre-processing. By producing portfolios of significantly different sizes the analyses which could be completed after the clustering would be limited. Since the research presented here placed more emphasis on financial and statistical analysis than previous studies ([4], [6], [7]) it was necessary to conduct additional pre-processing to handle outliers within the data.

Winsorising was chosen for the pre-processing as it would not discretise the data and would limit the information loss to companies with irregular financial variables. In addition to this it would enable the SOM clustering process to produce more evenly sized clusters than if it had not been applied. To limit the impact that the pre-processing would have on the data it was decided to only apply Winsorising to the variables when necessary. For MC and V the natural log of the values was instead taken and this bypassed the need for Winsorising.

This approach to pre-processing was found to be only partially successful and it is recommended that alternative pre-processing techniques be investigated. The main difficulty with the data was that the majority of the values would be grouped together, within a small range, thereby making differentiation between these points difficult. As a result the clustering will inherently group these similar companies together with outliers distorting the distance metrics. Since the most effective results were found with MC and V clustering (Sections 6.2.8 and 6.2.9 respectively), which did not undergo Winsorising, it is likely that more meaningful results could be found if a more effective method of data normalisation is applied.

Alternatively, fuzzy clustering may be more suited to the clustering process however different validity indices would be necessary to evaluate the performance. By using fuzzy clustering the visual and topographic advantages of the SOM would no longer be available. If however the clusters achieved are able to provide better results, the use of fuzzy clustering would be recommended.

7.2 Multiple Variable Clustering

Overall, the clustering completed using multiple financial variables was less effective than the clustering completed using single variables, when considering the clustering validity (Figure 6-2 and Figure 6-5). This poorer clustering validity can be attributed to the greater variation in inputs for multiple variable clustering. When using a single variable it is less likely that each dimension will differ significantly i.e. a firm with a low PE is likely to have a low PE in the following year. When using different variables it is less likely that the input vectors will have the same level of consistency i.e. a firm with a low PE will not necessarily have a low PB, hence increasing the range of possibilities for input vectors.

Although the clustering was less effective from a validity perspective it is still important to consider how the clustering process applies to companies from a financial portfolio point of view. In this regard the clustering completed with multiple variables has the distinct advantage of accounting for different aspects of a firm's performance. This was found to play a significant role in the clustering process and single clusters were found to have a defining factor which was caused by one of the input variables. One example would be cluster 1 from M1-08 being defined by the QR input (Figure 6-3).

Neither of the multiple variable clustering tests was successful at achieving abnormal returns. The clustering completed using the main variables (PB, PC, PE, MC and V) was more successful than the clustering using all 10 variables. Although these results were still not statistically significant the results achieved still provide insight into the clustering process. First it can be concluded that the addition of variables does not necessarily add any significant benefits to the clustering. It is likely that there is no need for additional variables because the main variables contain all the necessary information and it is simply repeated by including additional variables.

In addition to the above, the topographic nature of the SOM can be noted in the M2-08 test by the similar performance of clusters 1 and 2 which are next to one another on the SOM network. The input weights for clusters 1 and 2 (Figure 6-6) show that these two clusters differed for several inputs; however their market capitalisation weights (input 4) were similar. Since these clusters achieved very low market capitalisation values it is possible that small firm effect is the underlying feature in this test. This is further validated by the strong MC results found in Section 6.2.8. This impact of the small companies on the clustering has been discussed in more detail in Section 7.3.8.

Previous financial clustering studies ([4], [6], [7]) have placed little emphasis on the choice of financial variables. Considering the above it would be beneficial to look at different combinations of variables for clustering. Although a regression analysis will quickly reveal which variables are directly related, it is likely that more subtle relationships could be revealed through SOM clustering.

Overall the multiple variable clustering was not able to produce statistically significant results and the most significant results have been attributed to the market capitalisation input. There is a need to investigate clustering with different variables in future studies in order to confirm whether or not there are benefits to having multiple variables.

7.3 Single Variable Clustering and Financial Ratios

In order to gain an understanding of the predictive nature of single variables it was decided to include clustering completed using one variable at a time. To take advantage of the SOM it was decided to use five years of inputs rather than one. By doing so the SOMs ability to compress high dimensional data into a lower dimension was leveraged.

7.3.1 Debt/ Equity

The tests with DE achieved relatively good clustering validity, especially with the 2x1 SOM. Analysis of the input data from the two clusters in DE-02 shows that there was a relatively clear split in the DE data between high and low values. This meant that the clustering could be extremely effective with a very small SOM. Increases in the SOM size forced this split to be broken down more, resulting in less effective clustering.

Of interest is that cluster 1 achieved the greatest Silhouette Width in Table 6-10 and is comprised of purely financial companies. This can be attributed to the high DE ratio found in the financial companies. Clusters 8 and 9 have similar DE values with the companies in cluster 9 in general being more stable whereas the companies in cluster 8 had more irregular inputs. These inputs would have been the driving factor behind the poor Silhouette Width achieved by cluster 8 in Table 6-10.

In terms of the returns achieved there was nothing worth noting for the DE clustering. As a clustering variable DE would not be considered for further work as it was largely influenced by industry and did not show any indication of being a proxy for future returns.

7.3.2 Price/ Book Value

The PB clustering achieved relatively good validity results when considering the Davies-Boudin Index and Silhouette Width (Figure 6-11). With PB being a commonly reported ratio it is less likely to obtain many outliers and instead have a more evenly distributed range of values. In addition to this it is less likely that the five year range would have significantly different results between years. Taking this into consideration it is easier to cluster the data, thereby also increasing the validity of the clusters. This hypothesis is further supported by the even colour distribution in Figure 6-12.

The smaller clusters of interest in the PB clustering (Table 6-13) are clusters 3 (only consumer services) and 12 (only basic materials). For cluster 12 the input vectors were all low in value, which proved to be a defining factor for these companies. Cluster 3 however does not show any notable trend and since the cluster is very small it has been considered an anomaly rather than a trend. With these two clusters being small it becomes difficult to draw conclusions with regards to their performance. By only considering the larger clusters it can be seen that PB cannot be considered a proxy for future returns.

From a clustering point of view PB is suitable and this has mostly been attributed the large amount of attention it receives. As a proxy for returns no evidence was found to promote further analysis or to contradict EMH.

7.3.3 Price/ Cash Flow

The PC clustering was expected to be one of the less effective in terms of validity due to the wide range of values present in the data. This was however not the case when evaluating the overall validity (Figure 6-14) and instead the clustering validity was better than expected.

When taking the individual clusters from the 3x4 SOM (Table 6-16) the poorer clustering validity becomes evident. In addition to this the fluctuations between the years for the input vectors is evident in Figure 6-15 where the colour of the weight planes for each neuron differ significantly between each input/ year. This variation is due to the inherent unstable nature of cash flow.

The clusters produced by the SOM using PC were not able to achieve abnormal returns and even cluster 10, with only 2 companies, did not achieve returns which differed significantly. The statistical analysis further supports the hypothesis that the PC clustering was not able to produce abnormal returns and based on these points it can be concluded that the PC clustering was ineffective.

7.3.4 Price/ Earnings

The PE clustering achieved very poor validity in Figure 6-17 and Table 6-19 which was not expected since PE is so commonly reported.

Analysis of the input vectors showed that the PE ratios would often differ considerably from year to year. Furthermore, the Winsorising increased the effect because many of the variables were able to fluctuate from very low to very high within the refined range. Without the Winsorising the clustering would not have been possible since the SOM would continuously group the majority of the companies together. Compared to the other clusters this appears to be most apparent with PE, making it less suited to clustering than most of the other variables. This poor clustering resulted in little differentiation between the large clusters and the returns achieved being very similar.

For the PE tests it can therefore be concluded that the poor clustering has played the most significant role in not achieving abnormal returns. In order to overcome this problem a different approach to the pre-processing will be required in future work so as to enable the clustering algorithm to differentiate between these similar input vectors.

7.3.5 Quick Ratio

Overall the clustering validity for QR was relatively good (Figure 6-20). The raw data presented few outliers and as a result the data was not greatly affected by Winsorising and could be easily clustered. An evaluation of the clusters formed by the 3x4 SOM shows that there were often examples of clusters which similar trends in QR and these trends appeared to be less random that those found in the PE and PC tests. Considering the Silhouette Widths found in Table 6-22 it can be seen that the individual clusters were often above 0.30 as a result of these consistent trends being present. In addition to the clustering validity the QR clustering did not simply group companies from the same sector together.

As a proxy for future returns the QR clustering revealed no significant trends. Considering only the larger clusters there was little deviation between the clusters and this was reiterated in the statistical results.

Considering these points it can be concluded that QR was effective from a clustering perspective but did not yield any significant financial results.

7.3.6 Return on Assets

The clustering completed using RA was very poor when considering the validity indices in Figure 6-23. Considering the cluster sizes in Table 6-25 it can be seen that using RA as an input variable resulted in inconsistent cluster sizes. This was also the only clustering test analysed which produced a single firm cluster. The reason for the single firm cluster can be seen in Figure 6-24 where cluster 7 is found to have large fluctuations across the weight planes. Since the results presented in detail refer to the 3x4 SOM, which can be considered relatively large, it is not surprising that a single firm could be separated. An investigation into the smaller SOMs revealed that the single firm cluster was first formed with the 2x4 SOM making it the only cluster with less than 10 companies. Since this occurred with such a small SOM the input vector for this cluster can be considered highly erratic.

The returns from the RA clustering show greater deviations than many of the other variables. Since cluster 7 only contained a single firm it was expected to have a greater deviation from the All-Share Index and this is evident in Figure 6-25. Apart from the cluster 7's unique performance it was found that the larger clusters also achieved varied results. These differences could however not be confirmed with the statistical analysis and therefore it is not possible to assume that the performance was abnormal.

For the RA test it can be said that the clustering was not effective when only considering validity. The RA clustering was however effective at placing outlier companies in unique clusters which could have practical applications for financial analysis.

7.3.7 Return on Equity

The validity results for the RE clustering in Figure 6-26 and Table 6-28 were poor. In addition to this the cluster sizes from the RE clustering were the least suited for financial analysis. Instead of producing several clusters which were large enough for further analysis the RE clustering produced two very large clusters. An analysis of the smaller SOMs leading up to the 3x4 SOM reveals that there was a single large cluster which split in the 2x3 SOM to make two smaller clusters. Although the Winsorising was used to improve the cluster sizes it was not successful with the RE clustering as the SOM only produced two clusters which could be considered for portfolio generation.

Since these two portfolios contained the majority of the companies from the analysis, with similar RE values, it was assumed that their performance would be similar in nature. This is evident in Figure 6-28 where clusters 3 and 5 had similar performance over all five years.

The RE clustering was not successful as it suffered from outliers and even with the Winsorising the SOM could still not produce an adequate number of clusters. As a result of this poor clustering the financial and statistical analyses were limited and no conclusions regarding abnormal returns could be made.

7.3.8 Market Capitalisation

The MC clustering was the most successful of all the clustering tests. Since the data was evenly distributed it didn't require Winsorising and instead the data could be normalised by taking the natural log. With the even distribution of values the SOM was able to successfully produce evenly sized clusters with relatively good validity (Figure 6-29).

Since the MC clustering was capable of producing even cluster sizes it was possible to use a 2x4 SOM. Table 6-31 shows that the cluster sizes were constant besides for cluster 1, which contained large companies which listed after the first input year. In addition to the large clusters the Silhouette Width of each cluster was above 0.25 and compared to the other tests this is the best set of results achieved. Furthermore the clusters were not comprised of companies from the same sectors but instead presented a range of sectors per cluster.

The use of five years as an input for MC does not seem to have been beneficial for the clustering. This is because MC doesn't change drastically over time and if one year of inputs had been used a similar result could have been expected, with the exception of cluster 1.

The most significant output from the MC clustering was the statistical analysis which revealed a statistically significant difference in the returns (Table 6-33). This result has been attributed to cluster 8's poor performance which implies that for the time frame taken the small cap companies have performed worse than their larger counterparts. Although many studies have found a negative relationship between firm size and returns [52], [53], [54] there has also been some evidence to support the trend found in this study [57]. Considering that the time frame chosen for this research was during poor market performance it is not surprising that the larger companies have outperformed their smaller counterparts. Investors would be more hesitant to invest in smaller companies due to uncertainty making it difficult for them to grow. Although there are contradictions between the various studies it should be noted that in order to identify these firm size trends it is important to include the smallest companies. Without the inclusion of cluster 8 no trend would have been noted. This same point was highlighted by Moor and Sercu [58] and Banz [52].

Overall the MC tests were very successful considering both clustering and financial aspects. The tests highlighted the importance of including all companies in the analysis as well as the benefit of having evenly distributed input data.

7.3.9 Volatility

The volatility raw data required normalisation using the natural log as described in Section 5.1.1. As a result no Winsorising was necessary, however unlike the MC clustering, the validity for the V tests were not superior to the other tests (Figure 6-32). The Silhouette Widths achieved by the individual clusters in Table 6-35 are however consistent with the lowest value achieved being 0.18. The V clustering produced an adequate number of large clusters for financial analysis, without industry being an underlying defining factor. The colour range of the weight planes in Figure 6-33 shows that the input vectors were well distributed, taking full advantage of the input number range.

The returns achieved by the five largest clusters did differ by more than the results in some of the previous tests. Further investigation into these returns reveals that the differences are however not statistically significant. The main point of interest from the results is that cluster 5 achieved poor returns (Figure 6-34) and it had the greatest volatility of the analysed clusters (the yellow neurons in Figure 6-33). As discussed in Section 7.3.8 this study occurred over a time frame during which there was overall poor financial performance in the market. Considering this it is unlikely that investors would be willing to invest in highly volatile companies and would rather invest in companies with stable performance.

Overall the V tests proved successful when considering the clustering validity. Future work would provide more insight into the possible relationships between this variable and future returns.

7.4 Efficient Market Hypothesis

Some additional points, which were not directly considered in Sections 7.1 to 7.3, have been expanded upon in this section. This section is intended to take the points which have already been discussed and see how they are related to EMH as well as the possible implications of these results for real world applications.

Of all the tests completed only MC clustering produced results which were statistically significant. Although none of the other tests could provide any evidence to contradict EMH it is important to remember that the analyses were only conducted on clusters which were large enough to be considered for portfolios. The inclusion of all the very small clusters would have been likely to produce more abnormal returns; however they were excluded because the interpretation of the results would have been more random in nature. This is because very small clusters would be likely to have statistical different results, even if selected at random. So in order to not obtain misleading results small clusters were excluded.

Since the clustering groups together similar companies it is likely that the large clusters simply represent a market average. Further evidence of this is present in the smaller SOMs, in particular the 2x1 SOMs. In these SOMs there would be one very large cluster and one small break away cluster. As the SOM sizes were increased the large cluster would get broken down into a few larger clusters. Since the analyses were focused on the larger clusters, and they predominantly came from the same original cluster, they were less likely to have major differences. So rather than producing clusters with significantly different results the SOM appears to have removed outliers from the industry average. The reason for the clustering producing results of this nature is because the majority of the companies would achieve similar values for the input variables and even with the Winsorising the impact of outliers was still significant. The MC clustering however was different in that the distribution of the input variables was relatively even (when taking a log scale) and hence MC-02 produced two clusters of sizes 145 and 110.

The second important point to note is the time frame chosen for the analysis. In order to confirm that EMH is no longer valid it would be necessary to complete the analysis over a significantly longer time frame. The purpose of this research was to determine if it would be possible to challenge EMH over a short time period so that further work could be completed over a suitable time frame. By limiting this research to a short time frame it is possible that long term trends were not considered. Since the clusters were only formed once it was vital that a trend appeared in the first three years after formation because the likelihood of a trend becoming apparent five years after formation is unlikely.

Another concern with the data used is the fact that companies have different year ends, thus creating a system whereby the input variables are not taken at the same time. Due to the limited size of the JSE it would still be advisable to use the different year ends, however to overcome this problem it may be beneficial to use dynamic time warping [141]. This technique is often employed in time series analyses because two time series may have similar patterns, but are displaced with respect to one another along the time axis. The current SOM model measures the Euclidian distance between each point assuming that they are aligned in time. By taking the traditional SOM algorithm distance measurement and applying dynamic time warping it will be possible to compare the distances and determine the winning neuron. This research relied on the use of the MATLAB SOM algorithm; however it would be necessary to program a new SOM toolbox which incorporates dynamic time warping. It may also be necessary to interpolate numerous input values between each annual input for this method to be effective.

Although the methodology employed did not produce clusters which consistently outperformed the benchmark, the results could still be applied to real world applications. When investigating a particular firm the clustering process could be completed using various financial ratios. The clusters which are associated with this firm could then be analysed to reveal which firms could be related with the main one. In addition to this if an investor wished to analyse a firm then it may be important to note which neuron its cluster occurs at. If the cluster is at a neuron with very large neighbourhood distances or few firms then its behaviour would be regarded as irregular.

8. CONCLUSIONS

The conclusions for the study follow from the discussion in Section 7 and have been broken down into the relevant topics below:

1. *Data mining and the self-organising map:*

The data mining approach was successful in terms of automating a large portion of the analysis. Overall the input data was found to be difficult to manage due to the presence of outliers. The Winsorising methodology was successful at improving the quality of the clustering, from a financial perspective, however there is potential for improvement. Considering the results of this study, in parallel to previous studies, it can be concluded that the SOM has limited capabilities for handling financial data.

2. *Multiple variable clustering:*

Overall the multiple variable clustering was not able to produce statistically significant results and the most significant results have been attributed to the market capitalisation input. The addition of more inputs was found to be unbeneficial as the majority of the information required for clustering appears to be contained within major variables (PE, PB, PC, MC and V).

3. *Single variable clustering:*

Apart from the market capitalisation tests, no significant results were found. Some variables were found to cluster firms based on industry or remove extreme outliers. Using multiple years as inputs yielded limited benefits, apart from removing abnormal behaviour. The following gives more detailed conclusions on the individual tests:

- *Debt/ equity:* The DE clustering was not found to be a suitable proxy for future returns as the clustering was mostly impacted by the industry of the firm.
- *Price/ book value:* The PB clustering was found to be an effective variable for clustering. It was concluded that this is because it is a relatively stable variable which is frequently reported thereby limiting the number of outliers. As a proxy for future returns this variable was not found to be effective and it was not able to contradict EMH.

- *Price/ cash flow:* The PC clustering had limited validity because one of the underlying variables in this ratio is cash, which is relatively volatile. In addition to this PC was not found to be a suitable proxy for future returns.
- *Price/ earnings:* The clusters formed from the P/E tests showed little inter-cluster variation and they were not able to achieve abnormal returns. A different approach to the pre-processing will be required in future work so as to enable the clustering algorithm to differentiate between these similar input vectors.
- *Quick ratio:* Although the QR tests were effective from a clustering perspective, they were not able to yield abnormal financial results. It is possible that QR clustering could still have value as a proxy for high risk groups.
- *Return on assets:* The RA clustering was not effective when only considering validity. The RA clustering was however effective at placing outlier companies in unique clusters which could practical applications for financial analysis.
- *Return on equity:* The RE tests suffered considerably from outliers and were the least effective from a clustering perspective. This limited the financial analysis and as a result no financial conclusions can be drawn for this variable.
- *Market capitalisation:* The MC clustering was the most effective of all the tests. The inclusion of the smallest firms was found to be vital to produce the statistically significantly different performance between clusters. It was found that the largest firms performed the best, which can be attributed to the time frame chosen for the analysis. During this time general stock performance was poor so investors would have favoured larger firms, thereby improving their returns.
- *Volatility:* The volatility clustering was the second most effective from a clustering perspective. Although not statistically significant, the financial results for the more stable stocks outperformed their counterparts. As with the market capitalisation tests, this was attributed to the time frame chosen for the analysis.

4. The efficient market hypothesis:

Overall there were no contradictions to EMH found in the analysis. Market capitalisation was the only variable which was able to yield abnormal returns, however it would be necessary to extend the study over a longer time frame to conclude whether EMH is possibly invalid. Since the abnormal MC returns were attributed to the time frame chosen, when investors were seeking low risk stocks, it implies that over a longer time frame EMH would still be valid. Apart from EMH, the clustering did prove to potentially have some useful outputs which would be useful from a practical point of view.

9. FUTURE WORK

Following on from the conclusions drawn the following recommendations have been made:

- *New pre-processing methods:* One of the major limitations found in this study was the presence of outliers in the input data. Although the Winsorising approach did enable substantially improved clustering it still had limitations. A full study into the impact of different pre-processing methods would be necessary to decide which method is best suited to SOM clustering.
- *Fuzzy clustering:* This study focused on using SOMs for clustering but an investigation into the use of fuzzy methods for the purpose of clustering could potentially yield insightful results. One possible approach would be to use the fuzzy weighting assigned to each vector to define how it is weighted in the portfolio. It would therefore be possible to make portfolios balanced according to a new weighting. This would also enable portfolios to be made of a range of companies, whereas SOMs and K-means are constrained by having to allocate firms to single clusters.
- *Longer analysis time frame:* In order to find evidence which can contradict EMH it will be necessary to carry out the tests over a longer time frame. The purpose of this study was to provide insight into what direction one might take in this approach. It can therefore be concluded that should a study be continued over a longer time frame then market capitalisation would be an essential variable to include in the analysis.
- *Ideal multi variable clustering combinations*: Apart from the single variable clustering this study also looked at using multiple variables for the purpose of clustering. The results indicated that the addition of new variables had limited value, but it would be beneficial to understand, in more detail, which variables should be combined. It would be recommended to start with the PE, PB, PC, MC, V combination as a base and work from that reference point.
- *More developed stock markets*: This study focused on the JSE, however with such a small stock market the quality of data becomes more of an issue. By taking a larger, more developed stock market, it could be possible to use firms with less outliers. By doing so the problems associated with pre-processing would be partially overcome and the results could be significantly more meaningful. Apart from this, most of the reference material refers to more developed stock markets, thereby making it easier to compare results to previous studies.
- *Dynamic time warping:* Due to the various firms having different year ends the input vectors are technically misaligned. To overcome this one could use dynamic time

warping to measure the distance between input vectors and nodes. This could possibly take into account micro trends with the stock exchange and enable all firms to be evaluated on a common base.

- *Practical implications*: As it was noted, there could be possible practical implications from the work. Although the clustering does not seem to be able to achieve abnormal returns it does not mean that the results have no value. A study which involves industry to understand how these results could be applied to real life applications could yield interesting results.

REFERENCES

1. *A clustering method to identify representative financial ratios.* **Wang, Y and Lee, H.** 4 (1087-1097), s.l. : Elsevier, Information Sciences, 2008, Vol. 178.

2. *Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks.* **Enke, D, Grauer, M and Mehdiyev, N.** 201-206, s.l. : Elsevier, Procedia Computer Science, 2011, Vol. 6.

3. *Analysis of data clusters obtained by self-organizing methods.* **Gafiychuk, V V, Datsko, B Y and Izmaylova, J.** (547-555), s.l. : Elsevier, Physics A, 2004, Vol. 341.

4. *Clustering Indian stock market data for portfolio management.* **Nanda, S R, Mahanty, B and Tiwari, M K.** 8793-8798, s.l. : Elsevier, Expert Systems with Applications, 2010, Vol. 37.

5. *Mining stock category association and cluster on Taiwan stock market.* **Liao, S, Ho, H and Lin, W.** 1-2 (19-29), s.l. : Elsevier, Expert Systems with Applications, 2008, Vol. 35.

6. *Stock Selection Based on Data Clustering Method.* **Wang, R.** s.l. : Seventh International Conference on Computational Intelligence and Security, 2011.

7. **Kelvin, H and Sian, S.** *Using Self-Organizing Maps (SOM) to Cluster.* 2006.

8. **Marques, B and Silva, N.** *Feature clustering with self-organizing mas and an application to financial time series for portfolio selection.* 2010.

9. *Using SOM and PCA for analysing and interpreting data from a P-removal SBR.* **Aguado, D, et al.** 6 (919-930), s.l. : Elsevier, Engineering Applications of Artificial Intelligence, 2008, Vol. 21.

10. *Comparing the peformance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping.* **Budayan, C, Dikmen, I and Birgonul, M T.** 9 (11772-11781), s.l. : Elsevier, Expert Systems with Applications, 2009, Vol. 36.

11. *A New SOM-Based Method For Profile Generation: Theory And An Application In Direct Marketing.* **Seret, A, et al.** 1, Leuven, Belgium : Elsevier, European Journal of Operational Research, 2012, Vol. 220.

12. *SOM of SOMs.* **Furukawa, T.** 4, Kitakyushu, Japan : Elsevier, Neural Networks, 2009, Vol. 22.

13. **Haykin, S.** *Neural Networks A Comprehensive Foundation.* Delhi, India : Pearson Education Inc., 1999. 81-7808-300-0.

14. **Reilly, F K and Brown, K C.** *Investment Analysis and Portfolio Management 7e.* Mason, Ohio : Thomson Learning, 2003.

15. **Gibson, C H.** *Financial Reporting & Analysis 11e.* Ohio : South-Western Cengage Learning, 2009. 0-324-66083-9.

16. **Gallati, R.** 15.433 Investments. [Online] 2003. [Cited: 03 02 2013.] http://ocw.mit.edu/courses/sloan-school-of-management/15-433-investments-spring-2003/lecture-notes/154331introduction.pdf.

17. *Investment perfomance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis.* **Basu, S.** 3, s.l. : The Journal of Finance, 1977, Vol. 32.

18. **Zakrajsek, E.** *Financial Time Series and Their Characteristics.* s.l. : University of Ljubljana Faculty of Mathematics & Physics, 2009.

19. **Tsay, R S.** *Analysis of Financial Time Series 3e.* New Jersey : John Wiley & Sons Inc, 2010. 978-0-470-41435-4.

20. **Defusco, R A, et al.** *Quantitative Investment Analysis 2nd Edition.* New Jersey : John Wiley & Sons Inc., 2007. 13 978-0-470-05220-4.

21. **Hawawini, G and Keim, D B.** *The Cross Section of Common Stock Returns: A Review of the Evidence and Some New Findings.* 1998.

22. **Jago, F.** *A South African look at value vs. growth investing, extrapolation, and risk.* Johannesburg : University of the Witwatersrand School of Economics and Business Sciences, 2011.

23. **Murphy, J J.** *Technical Analysis of the Financial Markets.* New York : New York Institute of Finance, 1999. 0-7352-0066-1.

24. **Stevens, L.** *Essential Technical Analysis.* New York : John Wiley & Sons, Inc., 2002. 0-471-15279.

25. **Anderson, D R, Sweeney, D J and Williams, T A.** *Statistics for Business and Economics.* Ohio : Thomson Higher Education, 2009. 0-324-65837-0.

26. **Mendelsohn, L B.** *Trend Forecasting With Technical Analysis.* s.l. : Market Technologies Corporation, 2000. 1-883272-91-2.

27. **Bernstein, J.** *The Complete Day Trader Vol. 1.* New York : McGraw Hill Inc, 1995. 0-07-009251-6.

28. **Kordos, M and Cwiok, A.** *A New Approach to Neural Network Based Stock Trading Strategy.* Bielsko-Biala : University of Bielsko-Biala, Department of Mathematics and Computer Science.

29. **Appel, G.** *Technical Analysis.* New Jersey : Financial Times Prentice Hill, 2005. 0-13-147902-4.

30. **Duarte, J.** *Market Timing for Dummies.* Indianapolis : Wiley Publishing Inc, 2009. 978-0-470-38975-1.

31. *Efficient capital markets; A review of theory and emperical work.* **Fama, E F.** 2 (383-417), s.l. : The Journal of Finance, 1970, Vol. 25.

32. *The Efficient Market Hypothesis and Its Critics.* **Malkiel, B G.** 1 (59-82), s.l. : Journal of Economic Perspectives, 2003, Vol. 17.

33. **Sewell, M.** *History of the Efficient Market Hypothesis.* s.l. : UCL Department of Computer Science, 2011.

34. **Aronson, D.** *Evidence-Based Technical Analysis: Applying the Scientific Method and Statistical Inference of Trading Signals.* New Jersey : John Wiley & Sons, 2011. 978-0-470-00874-4.

35. **Lee, C F, Lee, A C and Lee, J C.** *Financial Analysis, Planning and Forecasting: Theory and Application.* Singapore : World Scientific Publishing Co Pte Ltd, 2009. 978-981-270-608-9.

36. **Lo, A W.** Efficient Markets Hypothesis. [book auth.] S. Durlauf L. Blume. *The New Palgave: A Dictionary of Economics 2e.* New York : Palgrave McMillan, 2007.

37. *Risk and Return: CAPM and CCAPM.* **Chen, M H.** 1, s.l. : North Holland, The Quartely Review of Economics and Finance, 2003, Vols. 43 (369-393).

38. *CAPM for estimating the cost of equity capital: Interpreting the emprical evidence.* **Da, Z, Guo, R J and Jagannathan, R.** 1, s.l. : Elsevier, Journal of Financial Economics, 2012, Vols. 103 (204-220).

39. *The Cross-Section of Expected Stock Returns.* **Fama, E F and French, K R.** 2, s.l. : The Journal of Finance, 1992, Vol. XLVII.

40. *Common risk factors in the returns on stocks and bonds.* **Fama, E F and French, K R.** Chicago : Journal of Financial Economics, 1993, Vols. 33 (3-56).

41. **Brigham, E F and Daves, P R.** *Intermediate Financial Management.* s.l. : Cengage Learning, 2009. 9780324594713.

42. *Value versus growth stocks in Singapore.* **Yen, J Y, Sun, Q and Yan, Y.** 1 (19-34), s.l. : Journal of Multinational Financial Management, 2004, Vol. 14.

43. *Value versus growth: The International Evidence.* **Fama, E F and French, K R.** 6, s.l. : The Journal of Finance, 1998, Vol. 3.

44. *Contrarian Investment, Extrapolation, and Risk.* **Lakonishok, J, Shleifer, A and Vishny, R W.** 5, s.l. : The Journal of Finance, 1994, Vol. 44.

45. *On the computation of returns in tests of the stock market overreaction hypothesis.* **Dissanaike, G.** 6 (1083-1094), s.l. : Journal of Banking and Finance, 1994, Vol. 18.

46. *The Overreaction Hypothesis and the UK Stockmarket.* **Clare, A and Thomas, S.** 7, s.l. : Journal of Business Finance & Accounting, 1995, Vol. 22.

47. *Does the Stock Market Overreact.* **De Bondt, F M and Thaler, R.** 3, s.l. : The Journal of Finance, 1985, Vol. 40.

48. *The evolution of the January effect.* **Moller, N and Zilca, A.** 3 (447-457), s.l. : Journal of Banking & Finance, 2080, Vol. 32.

49. *Further Evidence on Investor Overreaction and Stock Market Seasonality.* **De Bondt, F M and Thaler, R.** 3 (557-581), s.l. : The Journal of Finance, 1986, Vol. 42.

50. *Size, Seasonality, and Stock Market Overreaction.* **Zarowin, P.** (113-125), s.l. : Journal of Finance, 1990, Vol. 25.

51. *Tests of the Contrarian Investment Strategy Evidence from the French and German stock markets.* **Mun, J C, Vasconcellos, G M and Kish, R.** 3 (215-234), s.l. : International Review of Financial Analysis, 1999, Vol. 8.

52. *The relationship between return and market value of common stocks.* **Banz, R W.** s.l. : Journal of Financial Economics, 1981, Vols. 9 (3-18).

53. *The firm size effect on stock returns in a developing stock market.* **Wong, K A.** Singapore : Economics Letters, 1989, Vols. 30 (61-65).

54. *The size effect in the Mexican stock market.* **Herrera, M J and Lockwood, L J.** 4, s.l. : Elsevier, Journal of Banking and Finance, 1993, Vols. 18 (621-632).

55. *Firm life expectancy and the heterogeneity of the book-to-market effect.* **Chen, H.** 2 (402-403), s.l. : Journal of Financial Economics, 2011, Vol. 100.

56. *Three analyses of the firm size premium.* **Horowitz, J L, Loughran, T and E, Savin N.** (143-153), s.l. : Journal of Emperical Finance, 2000, Vol. 7.

57. *New Paradigm or Same Old Hype in Equity Investing.* **Chan, L K C, Karceski, J and Lakonishok, J.** (23-36), s.l. : Financial Analysts Journal, 2000, Vol. 56.

58. *The smallest firm effect: An international study.* **De Moor, L and Sercu, P.** s.l. : Journal of International Money and Finance, 2012, Vol. (In Press).

59. *A liquidity-augmented capital asset pricing model.* **Liu, W.** s.l. : Elsevier, Journal of Financial Economics, 2006, Vols. 82 (631-671).

60. *Asset pricing with liquidity risk.* **Pedersen, V V and Acharya, L H.** s.l. : Elsevier, Journal of Financial Economics, 2004, Vols. 77 (375-410).

61. *Liquidity and asset pricing: Evidence from the Hong Kong stock market.* **Tam, K S K and Lam, L H K.** s.l. : Elsevier, Journal of Banking & Finance, 2011, Vols. 35 (2217-2230).

62. *Size, value and liquidity. Do They Really Matter on an Emerging Stock Market.* **Voronkova, J and Lischewski, S.** s.l. : Elsevier, Emerging Markets Review, 2012, Vols. 13 (8-25).

63. *Liquidity and stock returns in Japan: New evidence.* **Chang, Y Y, Faff, R and Hwang, C Y.** s.l. : elsevier, Pacific-Basin Finance Journal, 2010, Vols. 18 (90-115).

64. **LaBarge, K P and Hamilton, D J.** *Lower dividend yields today: Lower stock returns tomorrow?* s.l. : Vanguard, 2011.

65. *Payout policy in the 21st century.* **Brav, A, et al.** s.l. : Elsevier, Journal of Financial Economics, 2005, Vols. 77 (483-527).

66. *Disappearing dividends: changing firm characteristics or lower propensity to pay?* **Fama, E F and French, K R.** 1, s.l. : Journal of Financial Economics, 2001, Vols. 60 (3-43).

67. *Are dividends disappearing? Dividend concentration and the consolidation of earnings.* **DeAngelo, H, DeAngelo, L and Skinner, D J.** 3 (425-456), s.l. : Journal of Finance Economics, 2004, Vol. 72.

68. *Are fewer firms paying more dividends? The international evidence.* **Ferris, S P, Sen, N and Yui, H P.** 4 (333-362), s.l. : Journal of Multinational Financial Management, 2006, Vol. 16.

69. *Dividend policy, growth, and valuation of shares.* **Miller, M and Modiglina, F.** (411-433), s.l. : Journal of Business, 1961, Vol. 34.

70. *Do dividends matter more in declining markets?* **Fuller, K P and Goldstein, M A.** 3 (457-473), s.l. : Journal of Corporate Finance, 2011, Vol. 17.

71. *The irrelevance of the MM dividend irrelevance theorem.* **DeAngelo, H and DeAngelo, L.** 2 (293-315), s.l. : Journal of Financial Economics, 2006, Vol. 79.

72. *Price-Earnings Ratios.* **Nicholson, F.** s.l. : Financial Analysts Journal, 1960, Vols. (43-50).

73. *The Relationship Between Earning's Yield, Market value and the Returns for NYSE Common Stocks: Further Evidence.* **Basu, S.** s.l. : Journal of Financial Economics, 1983, Vols. 12 (129-156).

74. *A Misspecification of Capital Asset Pricing: Emperical Anomalies Based on Earngins Yields and Market Values.* **Reinganum, M.** s.l. : Journal of Financical Economics, 1981, Vols. 9 (19-46).

75. *The international price-earnings ratio phenomenon.* **Bildersee, J S, Cheh, J J and Lee, C.** 3, s.l. : Japan and the World Economy, 1990, Vols. 2 (263-282).

76. *The short term predictive ability of earnings-price ratios: The recent evidence (1994-2003).* **Giannetti, A.** 1 (26-39), s.l. : The Quartely Review of Economics and Finance, 2007, Vol. 47.

77. *An empirical analysis of analysts' cash flow forecasts.* **DeFond, M L and Hung, M.** 1 (73-100), s.l. : Journal of Accounting and Economics, 2003, Vol. 35.

78. *An examination of attitudes involving cash flow accounting: Implications for the content of cash flow statements.* **McEnroe, J E.** 1 (1-22), s.l. : The International Journal of Accounting, 1996, Vol. 32.

79. **Palepu, K G, Healy, P M and Bernard, V L.** *Business Analysis and Valuation: Using Financial Statements 2e.* s.l. : Southwestern Pub Co, 1999. 978-0324015652.

80. *The cross-section of realized stock returns.* **Fama, E F and French, K R.** (427-465), s.l. : Journal of Finance, 1992, Vol. 47.

81. *Contrarian investment extrapolation, and risk.* **Lakonishok, J, Shleifer, A and Vishny, R W.** (1541-1578), s.l. : Journal of Finance, 1994.

82. *The time-series relations among expected return, risk, and book-to-market.* **Lewellen, J.** 1 (5-43), s.l. : Journal of Financial Economics, 1999, Vol. 54.

83. *Market Reactions to Tangible and Intangible Information.* **Daniel, K and Titman, S.** 4 (1605-1643), s.l. : Journal of Finance, 2006, Vol. 61.

84. *Institutional investors, intangible information, and the book-to-market effect.* **Jiang, H.** 1 (98-126), s.l. : Journal of Financial Economics, 2010, Vol. 96.

85. *Firm life expectancy and the heterogeneity of the book-to-market effect.* **Chen, H.** 2 (402-423), s.l. : Journal of Financial Economics, 2011, Vol. 100.

86. *Leverage change, debt overhang, and stock prices.* **Cai, J and Zhang, Z.** 3 (391-402), s.l. : Journal of Corporate Finance, 2011, Vol. 17.

87. *Financial accounting return on investment and financial leverage.* **Louma, G A and Spiller, E Jr A.** 2 (131-138), s.l. : Journal of Accounting Education, 2002, Vol. 20.

88. *Leverage and corporate performance: International evidence.* **Gonzalez, V M.** (169-184), s.l. : Review of Economics & Finance, 2013, Vol. 25.

89. *Leverage, investment, and firm growth.* **Lang, L, Ofek, E and Stulz, R M.** 1 (3-29), s.l. : Journal of Financial Economics, 1996, Vol. 40.

90. *Determinants of capital structure: Theory vs. practice.* **Kjellman, A and Hansen, S.** 2 (91-102), s.l. : Scandanavian Journal of Management, 1995, Vol. 11.

91. *Capital structure policies in Europe: Survey evidence.* **Brounen, D, de Jong, A and Koedijk, K.** 5 (1409-1442), s.l. : Journal of Banking & Finance, 2006, Vol. 30.

92. **JSE.** Johannesburg Stock Exchange. *JSE.* [Online] 2010. [Cited: 03 03 2013.] http://www.jse.co.za/Libraries/How_To_List_a_Company_MainBoard_Dual_Listings_Fact_Sheet/JSE_Dual_Listings_Fact_Sheet_04_2010.sflb.ashx.

93. **FTSE.** FTSE/JSE Top 40 Index Fact Sheet. [Online] 2012. [Cited: 03 02 2013.] http://www.ftse.com/Indices/FTSE_JSE_Africa_Index_Series/Downloads/J200.pdf.

94. **FTSE**. FTSE/JSE All-Share Index Fact Sheet. [Online] 2012. [Cited: 03 02 2013.] http://www.ftse.com/Indices/FTSE_JSE_Africa_Index_Series/Downloads/J203.pdf.

95. **FTSE**. FTSE/JSE RAFI All-Share Index. [Online] 2012. [Cited: 03 02 2013.] http://www.ftse.com/Indices/FTSE_JSE_Africa_Index_Series/Downloads/J263.pdf.

96. **FTSE**. FTSE/JSE Dividend+ Index. [Online] 2012. [Cited: 02 03 2013.] http://www.ftse.com/Indices/FTSE_JSE_Africa_Index_Series/Downloads/J259.pdf.

97. **Finance, Price Waterhouse Coopers Corporate.** *Signs of the Times: Valuation Methodology Survey 2009/2010 5th Edition.* s.l. : PWC, 2011.

98. **Zaïane, O R.** CMPUT590 Principles of Knowledge Discovery In Databases. [Online] 1999. [Cited: 03 03 2013.] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf.

99. **Tan, P, Steinbach, M and Kumar, V.** *Introduction to Data Mining.* s.l. : Pearson Education Limited, 2005. 0321321367.

100. **Witten, I H, Frank, E and Hall, M A.** *Data Mining - Practical Machine Learning Tools and Techniques 3e.* Burlington, MA : Morgan Kaufmann Publishers, 2011. 978-0-12-374856-0.

101. **Olson, D L and Delen, D.** *Advanced Data Mining Techniques.* Berlin : Springer, 2008. 978-3-540-76916-3.

102. **Kantardzic, M.** *Data Mining: Concepts, Models, Methods and Algorithms.* New Jersey : John Wiley & Sons Inc., 2011. 978-1-118-02914-5.

103. *Clustering if the Self-Organizing Map.* **Vesanto, J and Alhoniemi, E.** 3, s.l. : IEEE Transactions On Neural Networks, 2000, Vol. 11.

104. **Patel, N.** 15.062 Data Mining. [Online] 2003. [Cited: 03 03 2013.] http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/index.htm.

105. *Mining association rules from imprecise ordinal data.* **Chen, Y and Weng, C.** (460-474), s.l. : Elsevier, Fuzzy Sets and Systems, 2007, Vol. 159.

106. *A genetic algorithm with gene rearrangement for k-means clustering.* **Chang, D, Zhang, X and Zheng, C.** 7 (1210-1222), s.l. : Elsevier, Pattern Recognition, 2009, Vol. 42.

107. *Application of Multi-SOM Clustering Approach To Macrophage Gene.* **Ghouila, A, et al.** 328-336, s.l. : Elsevier, Infiection, Genetics and Evolution, 2008, Vol. 9.

108. *Determination of Cluster Number In Clustering Microarray Data.* **Shen, J, et al.** 1172-1185, s.l. : Elsevier, Applied Mathematics and Computation, 2005, Vol. 169.

109. **Fillenbaum, A and Rapport, S.** *An experimental study of semantic structures, Multidimensional Scaling Vol. II.* New York : Seminar Press, 1972. Vol. 2.

110. *Data Mining: A Preprocessing Engine.* **Shalabi, L A, Shaaban, Z and Kasasbeh, B.** 9 (735-739), s.l. : Journal of Computer Science, 2005, Vol. 2.

111. **Radivojac, P.** I400 Data Mining. [Online] 2005. [Cited: 03 03 2013.] http://www.informatics.indiana.edu/predrag/classes/2005springi400/lecture_notes_4_1.pdf.

112. **Gopalan, N P and Sivaselvan, B.** *Data Mining Techniques and Trends.* New Dehli : PHI Learning Private Limited, 2009. 978-81-203-3812-8.

113. *Evaluating and comparison of clustering algorithms in analyzing cell gene expression data.* **Chen, G, Jaradat, S A and Banerjee, N.** (241-262), s.l. : Statitica Sinica, 2002, Vol. 12.

114. **Rahnenfuhrer, J and Markowetz, F.** Exploratory Data Analysis: Clustering gene expression data. *Computational Diagnistics Group.* [Online] 2006. [Cited: 03 03 2013.] http://compdiag.molgen.mpg.de/ngfn/docs/2006/nov/cluster-exercises.pdf.

115. *Nonparametric Genetic Clustering: Comparison of Validity Indices.* **Bandyopadhyay, S and Maulik, U.** 1, s.l. : Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, Vols. 31 (120-125).

116. *A Cluster Separation Measure.* **Davies, D L and Bouldin, D W.** 2, s.l. : Pattern Analysis and Machine Intelligence, 1979, Vols. PAMI-1 (224-227).

117. *Validity index for crisp and fuzzy clusters.* **Pakhira, M K, Bandyopadhyay, S and Maulik, U.** 3 (487-501), s.l. : The Journal of the Pattern Recognition Society, 2004, Vol. 37.

118. *Cluster Validation with Generalized Dunn's Index.* **Bezdek, J C and Pal, N R.** s.l. : Artificial Neural Networks and Expert Systems, 1995.

119. *The use of data mining and neural networks for forecasting stock market returns.* **Enke, D and Thawornwong, S.** 4 (927-940), s.l. : Elsevier, Expert Systems with Applications, 2005, Vol. 29.

120. *Financial Forecasting through Unsupervised Clustering and Neural Networks.* **Pavlidis, N G, et al.** s.l. : University of Patras Artificial Intelligence Research Center (UPAIRC), 2006.

121. **Mcnelis, P D.** *Neural Networks in Finance: Prediction Edge in the Market.* California : Elsevier Inc, 2005. 0-12-485967.

122. *A Novel Self-Organizing Map (SOM) Neural Network for Discrete Groups of Data Clustering.* **Ghaseminezhad, M H and Karami, A.** 4, Rasht, Iran : Elsevier, Applied Soft Computing, 2010, Vol. 11.

123. *Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density.* **Chow, T W S and Wu, S.** (175-188), s.l. : Pergamon, Pattern Recognition, 2004, Vol. 37.

124. *Comparing SOM Neural Network with Fuzzy c-Means, K-Means and Traditional Hierarchical Clustering Algorithms.* **Mingoti, S A and Lima, J O.**

125. **Ultsch, A and Siemon, H.** *Exploratory Data Analysis: Using Kohonen Networks On Transputers.* s.l. : University of Dortmund, 1989. Technical Report 329.

126. *Application of Visual Clustering Properties of Self Organising Map In Machine-Part Cell Formation.* **Chattopadhyay, M, Dan, P K and Mazumdar, S.** 600-610, s.l. : Elsevier, Applied Soft Computing, 2012, Vol. 12.

127. *CAVE-SOM: Immense Visual Data Mining Using 3D Self-Organizing Maps.* **Wijayaserkara, D, Linda, O and Manic, M.** 2471-2478, s.l. : International Joint Conference on Neural Networks, 2011.

128. *A cross-national market segmentation of online game industry using SOM.* **Lee, S C, et al.** (559-570), s.l. : Elsevier, Expert Systems with Applications, 2004, Vol. 27.

129. **Beale, M H, Hagan, M T and Demuth, H B.** *Neural Network Toolbox User's Guide.* Natick : MathWorks, 2011.

130. **MATLAB.** SOM Toolbox 2.0 downloads. *Laboratory of Computer Science and Information Science.* [Online] Laboratory of Computer Science and Information Science Adaptive Infomatics Research Centre, 23 03 2005. [Cited: 13 04 2012.] http://www.cis.hut.fi/projects/somtoolbox/download/.

131. *SOM-Based Data Visualization Methods.* **Vesanto, J.** 2 (111-126), Helsinki : Elsevier, Intelligent Data Analysis, 1999, Vol. 3.

132. **Gaffney, J V.** *Value versus Glamour Investing: A South African Case.* s.l. : University of Pretoria, Gordon Institure of Business Science, 2009.

133. *Firm Size, Book-to-Market ratio, and Security Returns: A Holdout Sample of Financial Firms.* **Barber, B M and Lyon, J D.** 2 (875-883), s.l. : The Journal of Finance, 1997, Vol. 52.

134. **Salkind, N J.** *Encyclopedia of Research Design, Volume 3.* California : SAGE Publications, 2010. 978-1-4129-6127-1.

135. **Panik, M J.** *Advanced Statistics from an Elementary Point of View.* s.l. : Elsevier Academic Press, 2005. 13: 978-0-12-088494-0.

136. **Peck, R, Olsen, C and Devore, J.** *Introduction to Statistics & Data Analysis 4e.* Boston : Brooks / Cole, 2012. 13: 978-0-8400-5490-6.

137. *A further analysis of small firm stock returns.* **Patel, J B.** 7 (653-659), s.l. : Managerial Finance, 2012, Vol. 38.

138. *Evidence of a value-growth phenomenon on the Johannesburg Stock Exchange.* **Graham, M and Uliana, E.** s.l. : Investment Analysts Journal, 2001, Vol. 53.

139. **Boros, M.** Methodology and Statistics: Cluster Analysis. *University of Groningen.* [Online] 2011. [Cited: 03 02 2013.] http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/Boros-Clustering-2011-May-24.pdf.

140. **Lomax, R G.** *An Introduction to Statistical Concepts.* s.l. : Lawrence Erlbaum Associates Inc, 2007. 978-0-8058-5739-9.

141. **Ratanamahatana, C and Keogh, E.** *Making Time-Seris Classification More Accurate Using Learned Constraints.* s.l. : SAIM International Conference On Data Mining, 2004.

142. **Kumar, D and Nagesh, R.** *Multicriterion Analysis in Engineering and Management.* New Dehli : PHI Learning, 2010. 8120339762.

143. **Ramosall, J.** Matlab Central. [Online] [Cited: 23 07 2012.] http://www.mathworks.com/matlabcentral/fileexchange/27859-dunns-index/content/dunns.m.

BIBLIOGRAPHY

Anson, M J P, et al. *CAIA Level 1: An Introduction to Core Topics in Alternative Investments.* s.l. : John Wiley & Sons, 2012. 978-0470447024.

Carlin, B P. *Bayesian Methods for Data Analysis*, Boca Raton: CRC Press, 2009, 9781584886976.

Damadodaran, A. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*, New York: John Wiley & Sons, 1996, 0471112135.

Dowra, T. *MECN 4006: Data Mining Clustering for Stock Selection.* s.l. : University of the Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, 2011.

Elton, E J. *Modern Portfolio Theory and Investment Analysis*, New York: J. Wiley & Sons, 2003, 9780471238546.

Heaton, J. *Introduction to Neural Networks with Java 2e.* s.l. : Heaton Research, 2008. 1-60439-008-5.

Jolliffe, I T. *Principal Component Analysis 2e.* New York : Sringer, 2002. 0-387-95442-2.

Mirkin, B G. *Clustering for Data Mining*, Boca Raton, FL: Chapman & Hall/CRC, 2005, 1584885343.

Morningstar. Standard Deviation and Sharpe Ratio (Morningstar Methodology Paper). *Morningstar.* [Online] 2005. [Cited: 03 03, 2013.] http://corporate.morningstar.com/bf/documents/MethodologyDocuments/MethodologyPapers/StandardDeviationSharpeRatio_Definition.pdf.

National Treasury. *2009 National Budget Chapter 5 Asset and Liability Management.* s.l. : National Treasury of the Republic of South Africa, 2009.

National Treasurey. *2011 National Budget Chapter 6 Asset and Liability Management.* s.l. : National Treasury of the Republic of South Africa, 2011.

Reilly, F K, Norton, E A. *Investments*, Mason, Ohio: Thomson South-Western, 2006, 0324323840.

Refaat, M. *Data Preparation for Data Mining Using SAS.* s.l. : Elsevier Inc., 2007. 978-0-12-373577-5.

Ross, S A, Westerfield, R W, Jordan, B D. *Corporate Finance Fundamentals*, Boston, Mass.; London: McGraw-Hill Irwin, 2008, 9780071285636.

Rupert, D. *Statistics and Data Analysis for Financial Engineering*, New York: Springer, 2011, 9781441977861.

Smith, G. *Essential Statistics, Regression, and Econometrics*, Amsterdam; Boston: Academic Press, 2012, 9780123822215.

Standard Bank Online Share Trading. *Basic Investment Course - Unit 1: Introduction to Share Investment.* Johannesburg : Standard Bank, 2010.

Sullivan, D G. *Data Mining IV: Preparing the Data.* s.l. : Boston University, Computer Science 105, 2012.

Tuffery, S. *Data Mining and Statistics for Decision Making*, Chichester, West Sussex; Hoboken, NJ. : Wiley, 2011, 9780470688298.

Woodward, W A, Gray, H L, Elliott, A C. *Applied Time Series Analysis*, Boca Raton: Chapman & Hall/CRC, 2012, 9781439818374.

Witten, I H, Frank, E, Hall, M A. *Data Mining - Practical Machine Learning Tools and Techniques 3e.* Burlington, MA : Morgan Kaufmann Publishers, 2011. 978-0-12-374856-0.

Zieffler, A, Harring, J R, Long, J D. *Comparing Groups: Randomization and Bootstrap Methods Using R*, Hoboken, N.J.: Wiley, 2011, 9780470621691.

APPENDIX A

A complete list of all the companies used for the clustering may be found in Table A-1, where tickers for each of the companies can be found.

Table A-1: Ticker List of Companies Used In Clustering Analysis

ABL	ACL	ACP	ADH	ADI	ADR	AEG	AFB
AFE	AFR	AFX	AGI	AGL	ALT	AMA	AME
AMS	ANA	ANG	AOO	APK	APN	ARD	ARI
ARL	ART	ASA	ASR	ATN	ATS	AVI	AWT
BAT	BAU	BAW	BCF	BDM	BEE	BEL	BIL
BJM	BNT	BRC	BRT	BSB	BSR	BTG	BVT
CAE	CAP	CAT	CCL	CDZ	CFR	CKS	CLE
CLH	CLS	CMA	CMH	CML	CND	CNL	CNX
COM	CPI	CPL	CRG	CRM	CRW	CSB	CSO
CUL	CVI	CVN	CVS	DAW	DCT	DDT	DGC
DLV	DMR	DON	DRD	DST	DSY	DTA	DTC
EHS	ELE	ELH	ELR	EMG	ENV	EOH	ERM
EXL	EXX	FBR	FPT	FRT	FSR	FVT	GDH
GDO	GFI	GGM	GIJ	GLL	GMB	GND	GRF
GRT	HAR	HCI	HDC	HWA	HWN	HYP	IFR
ILA	ILV	IMP	INL	INP	IPL	ITE	ITR
IVT	JCD	JDG	JNC	JSC	KAP	KGM	KIR
KLG	KNG	LBH	LGL	LNF	LON	MAF	MAS
MCU	MDC	MFL	MIP	MMG	MMI	MOB	MPC
MRF	MSM	MST	MTA	MTL	MTN	MTX	MUR
MVL	NAI	NCS	NED	NHM	NPK	NPN	NTC
NWL	OCE	OCT	OML	OMN	PAL	PAM	PAP
PCN	PET	PGR	PHM	PIK	PMA	PMM	PMV
PNC	PPC	PPE	PPR	PSG	PWK	RAH	RBW
RDF	REM	RES	RLO	RMB	RNG	RTO	SAB
SAC	SAL	SAP	SBK	SBV	SCL	SDH	SER
SFN	SHF	SHP	SIM	SJL	SKJ	SLM	SMR
SNT	SNU	SNV	SOL	SOV	SPA	SPG	SPO
SQE	STA	SUI	SUR	SYC	TBS	TBX	TDH
TFG	TIW	TKG	TMT	TON	TPC	TRE	TRT
TRU	TSH	TSX	UCS	VIL	VLE	VTL	WBO
WES	WHL	WLO	WNH	YRK	ZCI	ZSA	

APPENDIX B

The following sections have been divided into their respective calculation groups. The calculations also follow the order presented in Methodology (Section 5.1). The majority of the results have been written using three significant figures but greater accuracy was used in the calculations. As a result small differences may arise if repeating the calculations with the values shown.

B-1 Davies Bouldin Index

To simply the explanation of the cluster validity analysis it has been decided to use a small sample dataset. By doing so it becomes possible to show the relationship between all clusters and the vectors within them. The calculations presented in this section follow the same logic as that presented by Kumar and Nagesh [142].

In Figure B-1 an illustrative representation of the inputs vectors used in the sample calculation can be seen. Each of the vectors consist of five dimensions and they have been grouped into three clusters. The input vectors have been shown with a ● symbol and the centre of the clusters are represented by a ⊗.

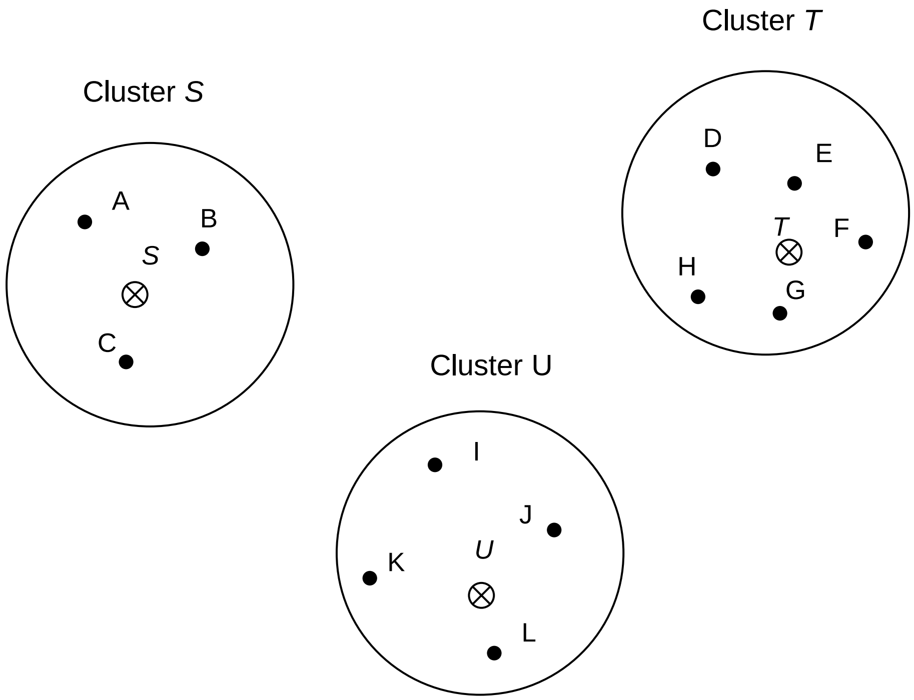

Figure B-1: Sample Calculation Clusters (Illustrative)

The values for the 12 input vectors can be seen in Table B-1 along with the centre of each of the clusters. It is important to note that in the actual calculations for this research, normalised values were used to ensure that equal weighting was given to the different variables and to reduce the impact of outliers (discussed in Section 7.1).

Table B-1: Sample Calculation Input Vectors and Clusters

Cluster	Vector	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
S	*x*	*2.08*	*-1.70*	*-3.14*	*0.0433*	*2.89*
	A	3.54	-2.53	-2.88	1.54	2.66
	B	-0.0512	-1.15	-3.96	-1.38	3.85
	C	2.76	-1.44	-2.57	-0.0300	2.17
T	*y*	*0.512*	*-3.402*	*3.468*	*0.112*	*-3.10*
	D	1.89	-3.81	2.56	3.80	-3.04
	E	-2.35	-2.37	4.44	2.38	-4.11
	F	2.62	-3.60	4.05	0.791	-0.792
	G	1.09	-3.15	2.29	-4.41	-3.71
	H	-0.69	-4.08	4.00	-2.00	-3.87
U	*z*	*1.09*	*3.02*	*-1.79*	*0.363*	*-0.77*
	I	-0.802	3.60	-3.57	0.792	-0.159
	J	2.43	2.07	1.58	-1.13	1.41
	K	4.29	3.35	-1.88	-2.90	-2.37
	L	-1.55	3.06	-3.30	4.69	-1.97

For the Davies Bouldin Index the first component of the calculation determined the distance of each vector from its cluster centre. The distance metric chosen for this was the Euclidian distance because the same metric was used in the SOM clustering process. Using the standard equation for Euclidian distance gives the following:

$$d(X_i, A_i) = \left\{ \sum_{k=1}^{N} (x_{ki} - a_{ki})^2 \right\}^{1/2}$$

Using Cluster *S* as a sample cluster for the calculation gives the following:

$$d(A, S) = \{(3.54 - 2.08)^2 + (-2.53 + 1.70)^2 + \cdots + (2.66 - 2.89)^2\}^{1/2} = 2.27$$

$$d(B, S) = \{(-0.0512 - 2.08)^2 + (-1.15 + 1.70)^2 + \cdots + (3.85 - 2.89)^2\}^{1/2} = 2.91$$

$$d(C, S) = \{(2.76 - 2.08)^2 + (-1.44 + 1.70)^2 + \cdots + (2.17 - 2.89)^2\}^{1/2} = 1.17$$

The above was then repeated with the remaining vectors to provide the absolute distance of each vector from its respective cluster centre. The similarity within each cluster was then determined using Equation 20 from Section 2.5.5. The value for q was chosen to be 1 to place less emphasis on outliers.

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} \left| X_j - A_i \right|^q \right)^{1/q}$$

Substituting the values obtained for cluster x (from the previous calculation) yielded the following:

$$S_s = \frac{1}{3}[2.27 + 2.91 + 1.17] = 2.12$$

$$S_T = \frac{1}{5}[9.29 + 11.5 + 8.35 + 9.80 + 10.7] = 3.76$$

$$S_U = \frac{1}{4}[6.82 + 6.34 + 8.27 + 9.01] = 4.41$$

The value of M_{ij} can then be calculated using Equation 21 from Section 2.5.5 with a value of $p = 2$ (Euclidian distance).

$$M_{ij} = \left(\sum_{k=1}^{N} \left| a_{ki} - a_{kj} \right|^p \right)^{1/p}$$

For the three cluster example provided three permutations were required:

$$M_{ST} = (|2.083 - 0.512|^2 + |-1.71 + 3.40|^2 + \cdots + |2.89 + 3.10|^2)^{1/2} = 9.22$$

$$M_{SU} = (|2.083 - 1.09|^2 + |-1.71 - 3.02|^2 + \cdots + |2.89 + 0.772|^2)^{1/2} = 6.22$$

$$M_{UT} = (|0.512 - 1.09|^2 + |-3.40 - 3.02|^2 + \cdots + |-3.10 + 0.772|^2)^{1/2} = 8.65$$

With the above steps completed it was then possible to determine the respective R_{ij} values (Equation 22, Section 2.5.5).

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

$$R_{ST} = \frac{2.12 + 3.76}{9.22} = 0.638$$

$$R_{SU} = \frac{2.12 + 4.41}{6.22} = 1.05$$

$$R_{TU} = \frac{3.76 + 4.41}{8.65} = 0.945$$

Then taking the maximum values for the three clusters:

$$R_i = \max(R_{ij})$$

$$R_S = \max(0.638, 1.05) = 1.05$$

$$R_T = \max(0.638, 0.945) = 0.945$$

$$R_U = \max(1.05, 0.945) = 1.05$$

The final stage in the Davies Bouldin Index calculation then required the average of the above values (Equation 24, Section 2.5.5) as shown below:

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$$

$$\bar{R} = \frac{1}{3}(1.05 + 0.945 + 1.05)$$

$$\bar{R} = 1.01$$

B-2 Silhouette Width

Although the implementation of the Silhouette Width was already incorporated into the MATLAB environment a sample calculation has been included since it provides insight into how this metric may be used.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The data presented in Section B-3 has been provided again in Table B-2 as it refers to the Silhouette Width calculations. As with Davies Bouldin calculations the Silhouette Width calculations were also completed using normalised values in the actual calculation.

Table B-2: Sample Calculation Data for Silhouette Width

Cluster	Vector	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
S	*x*	*2.08*	*-1.70*	*-3.14*	*0.0433*	*2.89*
	A	3.54	-2.53	-2.88	1.54	2.66
	B	-0.0512	-1.15	-3.96	-1.38	3.85
	C	2.76	-1.44	-2.57	-0.0300	2.17
T	*y*	*0.512*	*-3.402*	*3.468*	*0.112*	*-3.10*
	D	1.89	-3.81	2.56	3.80	-3.04
	E	-2.35	-2.37	4.44	2.38	-4.11
	F	2.62	-3.60	4.05	0.791	-0.792
	G	1.09	-3.15	2.29	-4.41	-3.71
	H	-0.69	-4.08	4.00	-2.00	-3.87
U	*z*	*1.09*	*3.02*	*-1.79*	*0.363*	*-0.77*
	I	-0.802	3.60	-3.57	0.792	-0.159
	J	2.43	2.07	1.58	-1.13	1.41
	K	4.29	3.35	-1.88	-2.90	-2.37
	L	-1.55	3.06	-3.30	4.69	-1.97

Initially the distance of each vector from other vectors within the same cluster was determined as follows (where X_{ki} is the vector of interest and X_{kj} is another vector from the same cluster).

$$D(X_i, X_j) = \left\{\sum_{k=1}^{N} \left(X_{ki} - X_{kj}\right)^2\right\}^{1/2}$$

Using cluster *S* with vectors A, B and C results in the calculations below:

$$d(A,B) = \{(3.54 + 0.0512)^2 + (-2.53 + 1.15)^2 + \cdots + (2.66 - 3.85)^2\}^{1/2} = 5.09$$

$$d(A,C) = \{(3.54 - 2.76)^2 + (-2.53 + 1.44)^2 + \cdots + (2.66 - 2.17)^2\}^{1/2} = 2.14$$

$$d(B,C) = \{(-0.0512 - 2.76)^2 + (-1.15 + 1.44)^2 + \cdots + (3.85 - 2.17)^2\}^{1/2} = 3.82$$

With all the distance values for each cluster it was then possible to take the mean of the distances to determine $a(i)$, as shown below.

$$a(i) = \frac{1}{T-1}\sum_{j=1}^{T-1}\left|X_i - X_j\right|$$

Using the distances calculated above the values of $a(A)$, $a(B)$ and $a(C)$ can be determined as follows:

$$a(A) = \frac{1}{2}[5.09 + 2.14] = 3.62$$

$$a(B) = \frac{1}{2}[5.09 + 3.82] = 4.45$$

$$a(C) = \frac{1}{2}[2.14 + 3.82] = 2.98$$

The next step in calculation of the Silhouette Width required the calculation of $b(i)$. This variable has been shown in the equation along with the respective calculations for the vectors in cluster *S*. For vectors A, B and C it was found that cluster *U* was the closest and therefore the distances shown are to vectors I, J, K and L.

$$b(i) = \frac{1}{T_n}\sum_{j=1}^{T_n}\left|X_{im} - X_{jn}\right|$$

$$b(A) = \frac{1}{4}[8.09 + 7.14 + 9.01 + 9.42] = 8.41$$

$$b(B) = \frac{1}{4}[6.64 + 7.30 + 9.19 + 9.55] = 8.17$$

$$b(C) = \frac{1}{4}[6.72 + 5.61 + 7.39 + 8.88] = 7.15$$

Finally to calculate $s(i)$ (the Silhouette Width for each vector) the values of $a(i)$ and $b(i)$ from above can be used.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(A) = \frac{8.41 - 3.62}{8.41} = 0.570$$

$$s(B) = \frac{8.17 - 4.45}{8.17} = 0.455$$

$$s(C) = \frac{7.15 - 2.98}{7.15} = 0.583$$

The average of these values then yields the Silhouette width for cluster *S*.

$$s(i) = \frac{1}{N_i}\left(\sum_{x \in i} s(x)\right)$$

$$s(S) = \frac{1}{3}\left(\sum_{x \in s} s(x)\right)$$

$$s(S) = \frac{1}{3}(0.570 + 0.455 + 0.583)$$

$$s(S) = 0.536$$

Finally the overall Silhouette Width was calculated taking the weighted average of cluster Silhouette Widths as mentioned in Section 5.1.3.

B-3 Dunn's Index

The Dunn's Index calculations presented in this section have been completed using the same sample data set as Section B-3, which can be seen in Table B-3 below. The MATLAB algorithms required for this calculation come from work completed by Ramosall [143]. As mentioned previously the calculations for the research used normalised values and the values below are for illustrative purposes.

Table B-3: Sample Calculation Input Vectors and Clusters for Dunn's Index

Cluster	Vector	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
S	*x*	*2.08*	*-1.70*	*-3.14*	*0.0433*	*2.89*
	A	3.54	-2.53	-2.88	1.54	2.66
	B	-0.0512	-1.15	-3.96	-1.38	3.85
	C	2.76	-1.44	-2.57	-0.0300	2.17
T	*y*	*0.512*	*-3.402*	*3.468*	*0.112*	*-3.10*
	D	1.89	-3.81	2.56	3.80	-3.04
	E	-2.35	-2.37	4.44	2.38	-4.11
	F	2.62	-3.60	4.05	0.791	-0.792
	G	1.09	-3.15	2.29	-4.41	-3.71
	H	-0.69	-4.08	4.00	-2.00	-3.87
U	*z*	*1.09*	*3.02*	*-1.79*	*0.363*	*-0.77*
	I	-0.802	3.60	-3.57	0.792	-0.159
	J	2.43	2.07	1.58	-1.13	1.41
	K	4.29	3.35	-1.88	-2.90	-2.37
	L	-1.55	3.06	-3.30	4.69	-1.97

The Dunn's Index equation (Equation 27, Section 2.5.5) shows the two basic components required for the calculation are cluster diameter ($\Delta(X_k)$) and distance between clusters ($\delta(X_i, X_j)$).

$$DI = \min_{1 \le i \le c} \left\{ \min_{1 \le j \le c, j \ne i} \left\{ \frac{\delta(X_i, X_j)}{\max\limits_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\}$$

As discussed in Section 2.5.5 there are several definitions for these two variables and for this reason two versions of the Dunn's Index were calculated. The first sample calculation refers to the standard equations for diameter and distance between clusters. The second sample calculation presented in this section uses alternative definitions for diameter and distance as recommended by Bezdek and Pal [118].

B-3-1 Dunn’s Index

The basic Dunn’s Index calculation takes the diameter as the maximum distance between two points in a cluster. Using the Euclidian distance between two vectors in a cluster results in the following equation:

$$d\left(X_i, X_j\right) = \left\{\sum_{k=1}^{N}\left(X_{ki} - X_{kj}\right)^2\right\}^{1/2}$$

Applying this to the vectors in cluster *S* from Table B-3 gives the following results:

$$d(A,B) = \{(3.54 + 0.0512)^2 + (-2.53 + 1.15)^2 + \cdots + (2.66 - 3.85)^2\}^{1/2} = 5.09$$

$$d(A,C) = \{(3.54 - 2.76)^2 + (-2.53 + 1.44)^2 + \cdots + (2.66 - 2.17)^2\}^{1/2} = 2.14$$

$$d(B,C) = \{(-0.0512 - 2.76)^2 + (-1.15 + 1.44)^2 + \cdots + (3.85 - 2.17)^2\}^{1/2} = 3.82$$

With the above step repeated for all the different combinations it is then possible to calculate the diameter for cluster *S*.

$$\Delta(S) = max\{d(A,B), d(A,C), d(B,C)\}$$

$$\Delta(S) = max\{5.09, 2.14, 3.82\}$$

$$\Delta(S) = 5.09$$

The distance between clusters can then be calculated using the same distance formula with vectors belonging to different clusters. The sample calculation below takes vector A from cluster *S* and compares it to cluster *U*.

$$d\left(X_i, X_j\right) = \left\{\sum_{k=1}^{N}\left(X_{ki} - X_{kj}\right)^2\right\}^{1/2}$$

$$d(A,I) = \{(3.54 + 0.802)^2 + (-2.53 - 3.60)^2 + \cdots + (2.66 + 0.159)^2\}^{1/2} = 8.09$$

$$d(A,J) = \{(3.54 - 2.43)^2 + (-2.53 - 2.07)^2 + \cdots + (2.66 - 1.41)^2\}^{1/2} = 7.14$$

$$d(A,K) = \{(3.54 - 4.29)^2 + (-2.53 - 3.35)^2 + \cdots + (2.66 + 2.37)^2\}^{1/2} = 9.01$$

$$d(A,L) = \{(3.54 + 1.55)^2 + (-2.53 - 3.06)^2 + \cdots + (2.66 + 1.97)^2\}^{1/2} = 9.42$$

Repeating the above process for vectors B and C enables the distance between clusters *S* and *U* to be calculated.

$$\delta(S,U) = min\{d(A,I), d(A,J), d(A,K), \dots, d(C,J), d(C,K), d(C,L)\}$$

$$\delta(S,U) = min\{8.09, 7.14, 9.01 \dots 5.61, 7.39, 8.88\}$$

$$\delta(S,U) = 5.61$$

With the calculations for cluster diameter, $\Delta(i)$, and inter-cluster distances, $\delta(i,j)$, repeated for clusters *T* and *U* it is possible to derive Table B-4 (shown below).

Table B-4: Dunn's Index Inter-cluster distances $\delta(i,j)$ and cluster diameters $\Delta(i)$

Cluster	*S*	*T*	*U*	$\Delta(i)$
S	0	7.61	5.61	5.09
T	7.61	0	6.84	8.31
U	5.61	6.84	0	9.69

Using Table B-4 it is then possible to compute the Dunn's Index for the sample dataset. First the maximum cluster diameter can be calculated (shown below).

$$\max_{1 \le k \le c}\{\Delta(X_k)\} = \max\{\Delta(S), \Delta(T), \Delta(U)\}$$

$$\max_{1 \le k \le c}\{\Delta(X_k)\} = \max\{5.09, 8.31, 9.69\}$$

$$\max_{1 \le k \le c}\{\Delta(X_k)\} = 9.69$$

With the maximum diameter calculated the individual Dunn's Index values for each cluster can be calculated as follows:

$$\min_{1 \le j \le c, j \ne i}\left\{\frac{\delta(X_i, X_j)}{\max_{1 \le k \le c}\{\Delta(X_k)\}}\right\}$$

$$\min_{j=S}\left\{\frac{\delta(X_i, X_j)}{\max_{1 \le k \le c}\{\Delta(X_k)\}}\right\} = \min\left\{\left(\frac{7.61}{9.69}\right), \left(\frac{5.61}{9.69}\right)\right\} = 0.578$$

$$\min_{j=T}\left\{\frac{\delta(X_i, X_j)}{\max_{1 \le k \le c}\{\Delta(X_k)\}}\right\} = \min\left\{\left(\frac{7.61}{9.69}\right), \left(\frac{6.84}{9.69}\right)\right\} = 0.706$$

$$\min_{j=U}\left\{\frac{\delta(X_i,X_j)}{\max\limits_{1\le k\le c}\{\Delta(X_k)\}}\right\} = \min\left\{\left(\frac{5.61}{9.69}\right),\left(\frac{6.84}{9.69}\right)\right\} = 0.578$$

Using these values the Dunn's Index for the set of clusters can be determined by taking the minimum value.

$$DI = \min_{1\le i\le c}\left\{\min_{1\le j\le c, j\ne i}\left\{\frac{\delta(X_i,X_j)}{\max\limits_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\}$$

$$DI = \min\{0.578, 0.706, 0.578\}$$

$$DI = 0.578$$

B-3-2 Alternative Dunn's Index

The Alternative Dunn's Index follows the same logic as the method presented above with the only difference being the cluster diameter, $\Delta(i)$, and inter-cluster, $\delta(i,j)$, distance metrics.

The cluster diameter can be defined with several different metrics, however for the purpose of this study the cluster diameter was defined as shown below (Equation 35, Section 5.1.3).

$$\Delta(i) = 2\left(\frac{\sum_{x\in i} d(x, v_i)}{|i|}\right), \text{where } v_i = \frac{1}{|i|}\sum_{x\in i} x$$

Using the sample dataset shown in Table B-3 and completing the calculation for cluster S results in the following:

$$\Delta(S) = 2\left(\frac{\sum_{x\in S} d(x, v_S)}{|S|}\right), \text{where } v_S = \frac{1}{|S|}\sum_{x\in S} x$$

$$\Delta(S) = 2\left(\frac{d(A,S) + d(B,S) + d(C,S)}{|S|}\right)$$

The above equation shows that it is necessary to calculate the distance of each vector from the cluster centre. The distances of vectors A, B and C from cluster centre *S* can be seen below:

$$d(A,S) = \{(3.54 - 2.08)^2 + (-2.53 + 1.70)^2 + \cdots + (2.66 - 2.89)^2\}^{1/2} = 2.27$$

$$d(B,S) = \{(-0.0512 - 2.08)^2 + (-1.15 + 1.70)^2 + \cdots + (3.85 - 2.89)^2\}^{1/2} = 2.91$$

$$d(C,S) = \{(2.76 - 2.08)^2 + (-1.44 + 1.70)^2 + \cdots + (2.17 - 2.89)^2\}^{1/2} = 1.17$$

Now continuing with the previous set of equations it is possible to calculate the diameter for cluster *S*.

$$\Delta(S) = 2\left(\frac{d(A,S) + d(B,S) + d(C,S)}{|S|}\right)$$

$$\Delta(S) = 2\left(\frac{2.27 + \ 2.91 + \ 1.17}{3}\right)$$

$$\Delta(S) = 4.24$$

Repeating the distance calculation from Section B-3-1 yields the following distances between vector A and vectors in cluster *U*.

$$d(A,I) = \{(3.54 + 0.802)^2 + (-2.53 - 3.60)^2 + \cdots + (2.66 + 0.159)^2\}^{1/2} = 8.09$$

$$d(A,J) = \{(3.54 - 2.43)^2 + (-2.53 - 2.07)^2 + \cdots + (2.66 - 1.41)^2\}^{1/2} = 7.14$$

$$d(A,K) = \{(3.54 - 4.29)^2 + (-2.53 - 3.35)^2 + \cdots + (2.66 + 2.37)^2\}^{1/2} = 9.01$$

$$d(A,L) = \{(3.54 + 1.55)^2 + (-2.53 - 3.06)^2 + \cdots + (2.66 + 1.97)^2\}^{1/2} = 9.42$$

With this process repeated for vectors B and C and the vectors from cluster *U* enables the inter-cluster distance between clusters *S* and *U* to be calculated as follows:

$$\delta(S,T) = \delta_{avg}(S,T) = \frac{1}{|S||T|}\sum_{x \in S, y \in T} d(x,y)$$

$$\delta(S,T) = \frac{1}{|S||T|}\big(d(A,I) + d(A,J) + d(A,K) + \cdots + d(C,J) + d(C,K) + d(C,L)\big)$$

$$\delta(S,T) = \frac{1}{|3||4|}(8.09 + \ 7.14 + \ 9.01 + \cdots + 5.61 + \ 7.39 + \ 8.88)$$

$$\delta(S,T) = 10.1$$

With the calculations for cluster diameter, $\Delta(i)$, and inter-cluster distances, $\delta(i,j)$, repeated for clusters *T* and *U* it is possible to derive Table B-5 (shown below).

Table B-5: Alternative Dunn's Index Inter-cluster distances δ(i,j) and cluster diameters ∆(i)

Cluster	*S*	*T*	*U*	$\Delta(i)$
S	0	10.0	7.91	4.24
T	10.1	0	10.4	7.52
U	7.91	10.4	0	8.81

Since the only difference between the two Dunn's Index calculations is the diameter and inter-cluster metrics the final steps have not been presented again.

B-4 Financial Analysis

For the purpose of the financial sample calculation it was decided to use cluster 7 from the PE clustering test. This cluster was chosen because it had an appropriate number of companies, enabling the calculation to show detail without too many repetitive steps. The companies in this cluster were:

1. Datatec Limited (DTC) - Technology
2. Conduit Capital Limited (CND) – Financials
3. Enterprise Risk Management (ERM) - Technology
4. Merafe Resources Limited (MRF) – Basic materials
5. Petmin Limited (PET) – Basic materials
6. Primeserv Group Limited (PMV) - Industrials
7. Tradehold Limited (TDH) - Financials
8. ZCI Limited (ZCI) – Basic materials

As shown in Equation 1 Section 2.1.1 the return for a share requires the share prices as well as the value of dividends paid over the chosen period.

Using Datatec Limited (DTC) for the purpose of the sample calculation and using the closing price difference between 29 June 2007 and 31 July 2007 results in the following:

$$R = \frac{P_T + D_T - P_0}{P_0}$$

$$R = \frac{4251 + 0 - 4155}{4155}$$

$$R = 0.0231 = 2.31\%$$

Then taking the natural log of the above for continuous compounding (Equation 2 Section 2.1.1)

$$r = \ln(1 + R)$$

$$r = \ln(1 + 0.0231)$$

$$r = 0.0228 = 2.28\%$$

The previous calculation was then completed for each firm in the cluster (portfolio) and the total return was determined. In the case of weighted returns the market capitalisation of each firm at the beginning of the 12 month period was used.

The sample calculation below has been completed for the period 29 June 2007 to 31 July 2007 with weightings taken according to 29 June 2007.

$$r_{weighted} = \sum_{i=1}^{n} w_i r_i$$

$$r_{weighted} = [0.0405 \times 0.0228 + \cdots + 0.175 \times (-0.0757)]$$

$$r_{weighted} = -0.0734 = 7.34\%$$

When calculating the excess monthly returns the above number would then be compared to the benchmark over the same period, as shown below:

$$r_{excess} = r_{weighted} - r_{benchmark}$$

$$r_{excess} = -0.0734 - 0.00789$$

$$r_{excess} = -0.0813 = -8.13\%$$

To determine the share's variance (over a 12 month period) monthly stock returns were used. The same methodology as the one described above was used to calculate the monthly returns for all the companies and with the returns calculated it was then possible to determine the sample variance and covariance between each firm's monthly returns. The values for the variance and covariance between the eight companies are shown in Table B-6.

Table B-6: Sample Calculation Monthly Return Variance and Covariance

	DTC	CND	ERM	MRF	PET	PMV	TDH	ZCI
DTC	0.0151							
CND	0.0056	0.0138						
ERM	0.0073	0.0049	0.0132					
MRF	-0.0015	0.0033	0.0051	0.0200				
PET	0.0074	0.0109	0.0043	-0.0095	0.0288			
PMV	0.0038	0.0216	0.0045	-0.0033	0.0242	0.0490		
TDH	0.0017	0.0015	0.0000	-0.0047	0.0014	0.0029	0.0061	
ZCI	0.0065	0.0026	0.0065	0.0035	0.0022	-0.0066	-0.0016	0.0259

With all these values it was then possible to calculate the portfolio standard deviation using Equation 11 below.

$$\sigma_{port} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov(R_i, R_j)}$$

$$\sigma_{port} = 0.0771$$