

https://dspace.org/

Modelling equity returns in South Africa: a data mining approach to the variable selection problem in asset pricing

Mhlanga, Wayne

2015
http://hdl.handle.net/10539/18797

Downloaded from a DSpace repository. https://dspace.org/

MODELLING EQUITY RETURNS IN SOUTH AFRICA: A DATA MINING APPROACH TO THE VARIABLE SELECTION PROBLEM IN ASSET PRICING

WAYNE MHLANGA
Supervisor: YUDHVIR SEETHARAM

A research report submitted to the School of Economic and Business Sciences, Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment (50%) of the requirements for degree of Master of Commerce in Finance.

Date of submission:
March 2015

DECLARATION

I, Wayne Mhlanga, declare that this research report is my own unaided work. It is submitted in partial fulfilment of the requirements for the degree of Master of Commerce in Finance at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

Signed:_____ Date: _____

ACKNOWLEDGEMENTS

It takes many sailors to command a galleon.

To my patrons, who offer so much more than their patronage, landsmen now but former seafarers themselves.

To my sailing master Yudhvir Seetharam, who let me set the course, man the helm, and command the rigging, but always ensured that I had a little less rope than was necessary to hang myself.

And to Locke, who really has no idea how to sail a ship but who lied so that I might believe it is possible.

Whatever the value of its contents, the ship is in port.

Thank you.

TABLE OF CONTENTS

DEFINITION OF KEY TERMS ---------- ix

ABSTRACT---------- 1

1. INTRODUCTION---------- 2

1.1 Research Objective and Hypotheses---------- 8

1.1.1 Primary Hypothesis ---------- 8

1.1.2 Secondary Hypotheses---------- 8

1.1.3 Tertiary Hypothesis ---------- 9

2. LITERATURE REVIEW---------- 10

2.1 On The Efficiency of Markets---------- 10

2.2 On Asset Pricing Theory and Pricing Factors ---------- 13

2.2.1 Arbitrage Pricing Theory---------- 14

2.3 On Macroeconomic Factors and Equity Returns---------- 18

2.4 On Data Mining and Feature Selection---------- 20

2.4.1 Least Angle Regressions (LARS) ---------- 23

2.4.2 Decision Trees and Random Forests ---------- 24

2.5 On Artificial Neural Networks ---------- 27

2.5.1 General Regression Neural Network (GRNN)---------- 29

2.6 Summary of the Literature Review ---------- 31

3. THEORETICAL MOTIVATIONS FOR THE EMPIRICAL FRAMEWORK ---------- 32

3.1 Modelling the Explanatory Variables ---------- 32

3.2 Determining the Model Features---------- 33

3.3 Modelling Ideals ---------- 36

3.4 Summary of Theoretical Motivations ---------- 37

4. DATA AND METHODOLOGY ---- 38
4.1 Data Collection and Sampling ---- 38
4.2 Pre-Processing ---- 43
4.3 Feature Selection and Modelling ---- 45
4.3.1 Linear Variable Selection ---- 46
4.3.2 Nonlinear Variables Selection ---- 47
4.4 Evaluation of Hypothese ---- 50
5. RESULTS ---- 54
5.1 Descriptive Statistics ---- 54
5.2 Modelling the Selected Features In the Training Period ---- 60
5.2.1 Linear Multifactor Models ---- 60
5.2.2 GRNN Multifactor Models ---- 71
5.3 Test Period Performance ---- 78
5.3.1 Diebold-Mariano Tests ---- 82
5.4 General Regression Neural Networks with Linear the Features ---- 84
5.5 Summary of Results and Remarks ---- 86
6. CONCLUSION ---- 88
7. REFERENCES ---- 91
APPENDICES ---- 102
Appendix 1 ---- 102
A: R code for Spline ---- 102
B: R code for Feature Selection using Random Forests ---- 102
C: R code for the construction of the GRNN ---- 103
Appendix 2 ---- 105
A: Histograms and Plots of Index Returns ---- 105

B: Correlation Matrices for the Industry Indices Variable sets. ---------- 108

TABLE OF FIGURES

Figure 2.1 Overview of the steps in the data science process. 21
Figure 2.2 Basic decision tree for ALSI returns. 25
Figure 2.3 Network architecture of a three layer feed forward neural network. 27
Figure 2.4 Network architecture of a generalised regression neural network 30
Figure 4.1 Overview of the methodology used in this study. 38
Figure 5.1 The correlation-matrix for the ALSI explanatory variables. 59

TABLE OF TABLES

Table 2.1 Literature pertaining to the relationship of macroeconomic variables and returns. 18
Table 4.1 Variable set from which relevant features are to be selected.....41
Table 4.2 The sample periods, sub-periods and the number of associated data points per variable. 43
Table 5.1 Summary statistics for the ALSI and Industry index log return series being modelled. 55
Table 5.2 Summary statistics for the macroeconomic explanatory variables. 57
Table 5.3 Summary statistics for the Gold, Oil and international stock indices. 58
Table 5.4 Linear APT models for the ALSI over the training periods. 61
Table 5.5 Linear APT models for the J500, J510 and J520 over the training periods. 64
Table 5.6 Linear APT models for the J530, J540 and J550 over the training periods. 66
Table 5.7 Linear APT models for the J560, J580 and J590 over the training periods 68
Table 5.8 Tally of the number of variables chosen by LARS and the number with significant coefficients.....70
Table 5.9 Variables selected for GRNN modelling of the ALSI. 72
Table 5.10 Variables selected for GRNN modelling of the J500, J510 and J520. 73
Table 5.11 Variables selected for GRNN modelling of the J530, J540 and J550. 75
Table 5.12 Variables selected for GRNN modelling of the J560, J580 and J590. 76
Table 5.13 Tally of the number of variables chosen by Boruta and the common variables. 77
Table 5.14 Test period performance metrics for the ALSI and industry indices models. 79
Table 5.15 Diebold-Mariano tests for testing period performance of the indices linear and neural network models. 82
Table 5.16 Training period diagnostics for GRNN with linear feature set. 84
Table 5.17 Test period performance metrics for GRNN with linear feature set. 85
Table 5.18 Diebold-Mariano tests on the forecasts of the GRNN models with the linear feature sets. 85

DEFINITION OF KEY TERMS

Artificial Neural Network – computational models inspired by the processing capabilities of the human brain. These are capable of pattern recognition, linear and nonlinear forecasting.

Bootstrap aggregating (bagging) – is a machine learning algorithm that creates a bootstrapped sample by sampling from the original training set with replacement.

Boruta Algorithm – a machine learning algorithm used for feature selection. It is wrapper algorithm for Random Forests and solves the difficulty in deciphering which variables are important.

Data Mining – a field of computer science that is concerned with the discovery of patterns in large data sets.

Data Science – a blanket term for a number of techniques from different fields. The term itself has come to represent a field encompassing the extraction of knowledge from data.

Decision Tree – a machine learning technique that builds out a tree like structure to map input variables to possible outputs. This can be used as a nonlinear variable selection tool.

Feature Selection – is the selection of a relevant variable subset from a wider set of variables.

General Regression Neural Network – a neural network that requires only one-pass training and converges to the underlying nonlinear or linear regression in the data.

Least-Angle Regression – a regression algorithm that is optimised for large data sets by using a unique method to iterate over data and discover linear relationships. This can be used as a linear variable selection tool.

Machine Learning – a broad term in computer science that describes sets of algorithms that can learn by iterating over data.

Model Appropriateness – in the context of this study model appropriateness refers to the model fit.

Ockham's Razor – is a scientific principle that suggests that among competing theories of equal probability one should always choose the theory with the fewest assumptions.

Parsimony – an ideal drawn from Ockham's razor which in the context of modelling suggests that fewer variables are preferable to more variables.

Random Forests – an ensemble data mining technique that uses a number of randomised decision trees to extract relevant features from data sets.

Stepwise Regression – this is the process of building a model in a piecemeal fashion by sequentially adding or removing variables based on the significance of their estimated coefficients.

ABSTRACT

This study demonstrates the use of data mining techniques for variable selection in the construction of macroeconomic time-series models. The Arbitrage Pricing Theory (APT) provides a broad based framework for modelling equity returns, but a significant drawback is that the theory itself says nothing about the identity and the number of variables necessary to adequately price returns. Variable selection is a common modelling problem for which the APT represents an acute example. This study borrows two computationally efficient variable selection algorithms from the field of data science, Least-Angle Regression and Random Forests, and demonstrates their efficacy at the selection of relevant variables that lead to appropriate multifactor time-series models for the JSE All-Share index and the JSE Industry indices. The resultant time-series models for the JSE All-Share index show good in-sample fit and perform better than a naïve model in test periods. The resultant models for the JSE Industry indices show a large degree of variation in training period model fit, but overall they suggest that the variables selected for modelling by the feature selection techniques are appropriate. There is strong evidence that the variables selected by the feature selection techniques are both period and industry dependent. Period and industry dependence implies that the subset of variables that comprise the true return generating process for JSE equity indices is not constant, which is symptomatic of an adaptive framework for returns.

1. INTRODUCTION

The ability to provide a robust and appropriate model for stock returns is the focus of asset pricing theory. Numerous asset pricing models have been proposed highlighting the myriad of potential pricing factors. Many of these models are coupled with impressive theoretical bases, and have demonstrated robust in-sample performance. However, few models have proven sufficiently robust for generalisation. Model appropriateness is necessitated by the inherent need to show justifiable explanatory power. However, the need for the most appropriate model – one with the best fit – is often balanced against the need for parsimony.

Parsimony is an ideal that predates much of the scientific literature, and originates from a maxim often attributed to William of Ockham (c. 1285 - 1347): "entities are not to be multiplied beyond necessity", implying that scientific best practices favour models with the less assumptions to models with more (Forster, 2004). The caveat is best expressed by Neal (1995) when he states that "deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex" (Neal, 1995, p. 104). This suggests that the appropriateness (or fit) of a model is the principal test of model viability and is always more important than parsimony.

In an asset pricing context, parsimony is important to ensure that a model can be easily applied. This is perhaps a key reason that the Capital Asset Pricing Model (CAPM) remains in widespread use. Robustness is likely the most obvious model requirement. It goes to the heart of scientific study – the principle of falsification. Amongst renowned scientific philosopher Karl Popper's early conclusions is the suggestion that "confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory" (Popper, 1963, p. 48). Robustness examined through an attempt at falsification of a model ensures that a confirmed model gives a pervasive explanation of returns. An additional practical requirement could be that of interpretability. Conceivably, practitioners would be loath to frequently utilise a model that cannot be easily interpreted.

Modelling ideals are important, but the primary challenge in asset pricing is not in the fulfillment of admirable ideals, but in the increasingly difficult task of selecting appropriate explanatory variables. The impetus for the manner in which variable selection is conducted in much of the traditional asset pricing literature is the Efficient Market Hypothesis (henceforth, EMH) outlined by Fama (1970). Fama (1970) establishes markets as efficient with respect to the available information. This implies that assets are priced quickly and accurately as new information emerges. If stock prices represent the equilibrium condition – the result of a multitude of investors rationally assessing risk based on the available information – then the corollary purports that there exists a set of risk factors of principal import to market participants. The fundamental question is thus: What factors provide an appropriate model for asset returns? The consequent task of asset pricing is the modelling of these relevant factors. Even if investors are not entirely rational as Fama (1970) assumes, or if the market is not fully efficient, an asset pricing model may still be pertinent if there are stable and predictable relationships between the changes in explanatory factors and returns – that is to say, so long as returns are not random. Indeed, perhaps an appropriate model may be one which expresses returns as a function of both variables that proxy for risks and others that proxy for the extent mispricing due to investor psychology[1].

Regardless, the traditional approach to variable selection for asset pricing has typically involved the examination of a limited set of risk factors that are routinely selected based on hypothesised or observed efficacy. Ross (1976) provided the Arbitrage Pricing Theory (henceforth APT) as a general framework for asset pricing models, incorporating the prospect of multiple key pricing factors. However, the APT stops short of identifying exactly which factors are most important for modelling returns or even how many factors are required. The APT is the emblematic example of the variable selection problem in asset pricing. The literature on variable selection for APT multifactor models broadly fits into two groups: 1) studies that view the true nature of relevant factors as unobservable and thus employ a factor-analytic approach (see Connor & Korajczyk, 1988; Dhrymes, Friend, & Gultekin, 1984; Lehmann & Modest, 1988; Reinganum, 1981; Roll &

[1] See Hirshleifer (2001). Also see Ho and Hung (2009) who shows that conditioning asset pricing models on investor sentiment results in better explanations of returns. Lakonishok, Shleifer, and Vishny (1994) argue that value investment strategies (buying stocks with low ratios of price to some measure of fundamental value) work because they are contrary to naïve investment strategies, hence using these factors in asset pricing may proxy elements of investor psychology.

Ross, 1980; van Rensburg & Slaney, 1997); and 2) studies that aim to pre-specify relevant factors for modelling returns based on theoretical associations (see Asprem, 1989; Bilson, Brailsford, & Hooper, 2001; Clare & Thomas, 1994; van Rensburg, 1996, 2000). This study takes the view that the true relevant factors are indeed latent, but they can be approximated by relevant proxies. Regardless, so long as there is ambiguity about the identity of the true relevant pricing factors or the most suitable proxies, variable selection remains a fundamental component of asset pricing.

The critical shortcoming that encumbers the widespread adoption of the APT is the lack of consensus about which explanatory variables (or risk factors) offer the best explanation for returns. Perhaps more importantly, this shortcoming is coupled with the absence of a framework for the identification of suitable explanatory variables. The developing field of data science provides significant advances in the evaluation of large quantities of data and the production of relevant succinct models. In short, data science tools provide a means of variable selection. Viewed from a data science perspective the selection of relevant variables for an asset pricing model such as the APT is a common problem, one for which there are a large number of computational solutions. Most notably, the field of data science provides a clear framework through which such problems may be resolved objectively. This study uses data science techniques to examine a large set of financial and economic factors as potential explanatory variables for stock returns, utilising the subset of relevant variables to build time-series models. The goal is to demonstrate the use of variable selection techniques from the field of data science in time-series multifactor modelling.

Of the many suitable algorithmic variable selection techniques provided by the field of data science, two relatively simple techniques are chosen to conduct variable selection in this study, Least-Angle Regression (henceforth, LARS) and Random Forests. The two techniques are chosen largely because their underlying processes are relatively easy to comprehend and they allow the selected variables to retain their economic meaning, which aids in the interpretation of the models produced. These techniques are also robust to high dimensional data - where m, the number of possible variables, is greater than n, the number of observations. They also show a reduced proclivity to overfit the data and a high degree of computational efficiency. LARS and Random

Forests also have distinctly different variable selection criterion and can potentially identify different sorts of relationships between the explanatory variables and returns.

LARS is employed as a variable selection (feature selection)[2] technique to establish a subset of linearly related variables. LARS is a computationally efficient regression algorithm for high dimensional data. It has been shown to represent a suitable method of determining relevant variables for asset pricing (Wang & Tan, 2009). LARS is a constrained version of ordinary least squares regression, utilising an efficient algorithm to iterate over the entire set of input variables in a stepwise fashion (Efron, Hastie, Johnstone, & Tibshirani, 2004). Random Forests, on the other hand, are employed to select variables with both linear and possible nonlinear relationships with returns. To understand the implementation of Random Forests one must first appreciate the technique on which they are based, that of inductive decision trees. Decision trees use a value test to split the data into subsets based on the set of variables, essentially learning the optimal tree from the data through a recursive process which stops when no more value can be gained through splits (Kuhn & Johnson, 2013). A determination of the relevant variables can be made by assessing the variables on which the data is split. Inductive decision trees, have been shown be a useful variable selection tool for developing asset pricing models (Thawornwong & Enke, 2004). Decision trees are suitable for assessing nonlinear relationships due to the scope of their splitting criteria (Breiman, Friedman, Olshen, & Stone, 1984). Random Forests can be viewed as ensemble feature selection techniques which average over a number of decision trees (Breiman, 2001). As Breiman (2001) notes, Random Forests avoid much of the overfitting that characterises many decision tree algorithms through the Law of Large Numbers.

The variables selected by LARS can be easily modelled as part of a linear time-series APT type model. To model the variable sets selected by Random Forests, which may have nonlinear relationships with returns, a roughly equivalent nonlinear model is required. Basic regression methods can be manually adjusted to incorporate nonlinear trends in data through exponentiation,

[2] Note that due to the differing conventions in the computer science and statistical literature the terms ‘variable selection’ and ‘feature selection’ will be used interchangeably throughout this study. Both terms refer more strictly to variable subset selection.

but this requires one to know the exact nature of the nonlinear relationship. Artificial neural networks (henceforth, ANN) present a powerful tool for modelling nonlinear relationships. In this study models that incorporate possible nonlinear interactions are computed through the use of neural networks. Cybenko (1989) shows that neural networks can theoretically be used to approximate any continuous nonlinear function. They are a black-box method but have the advantage of requiring no *a priori* assumptions about the data generating process (Kuhn & Johnson, 2013). Cao, Parry, and Leggio (2011) show that using ANNs to estimate the CAPM and the Fama and French (1992) explanatory variables leads to better return forecasts than traditional linear estimations. Specht (1990a) develops the General Regression Neural Network (GRNN). It is a novel neural network that does not require backpropogation and can be trained in a single pass. This study uses a GRNN to construct neural network models for returns. Specht (1990a) demonstrates that the GRNN converges to the underlying regression surface (linear or nonlinear) making the GRNN useful as an APT-type model.

Defining appropriate, parsimonious and robust models for asset returns is a difficult task. It is made more difficult by the concurrent challenge of relevant variable selection. This study aims to illustrate computational solutions to the variable selection problem, and demonstrate that the variables selected lead to viable time-series models of asset returns. LARS and Random Forests are employed as variables selection techniques to determine viable models for the Johannesburg Stock Exchange All-Share index (ALSI) and the main industry indices (J500, J510, J520, J530, J540, J550, J560, J580 and J590) over a sixteen year time period (1998 to 2014). Model viability is tested by assessing the in-sample fits of the resultant models. At the time of writing, no known paper has employed the variable selection techniques used here to determine relevant variables for asset pricing in South Africa. This study is, in truth, ancillary to asset pricing theory. Models are presented, but the purpose is simply to demonstrate the viability of the selected features and not to assess why selected features are relevant or address any number of similar asset pricing considerations. A study aiming to make definitive statements on asset pricing considerations would require cross-sectional analysis in order to be more directly comparable to the traditional asset pricing literature. This study uses time-series analysis which enables a clearer representation of the use of the variables selection techniques outlined.

The results in this study indicate that LARS and Random Forests select subsets of variables that result in viable linear and GRNN models of the ALSI and industry indices. Variables selected in all the linear ALSI models include returns to the international stock indices the Mexican Bolsa IPC Index, FTSE 100 and Bovespa Brasil, and changes in the coincidental economic indicator published by the South African Reserve Bank. Displaying some commonality with the variables selected by the LARS algorithm, the international equity indices are also predominant in the GRNN models of the ALSI. Variables selected by Random Forests to be used in GRNN models tended to be more redundant. Amongst the industry indices only one variable is in every selected subset – the ALSI. The relationship between the ALSI and industry indices is undoubtedly two-way but if one thinks of the industry indices as akin to weighted portfolios, the intuition suggests the usefulness of a beta coefficient as in the CAPM. The intercept is insignificant in all of the linear models. The variables selected for both the linear and GRNN models are sample specific and industry specific. Test period analysis showed the models to be unreliable predictors of returns – as suggested by the error metrics. However, most of the models performed better than a naïve model in the test period.

The rest of the study proceeds as follows: Chapter 2 comprises a review of the relevant literature; Chapter 3 provides a discussion of some of the motivations for the techniques employed in the study; Chapter 4 outlines the methodology utilised, highlighting the explanatory variables employed; Chapter 5 presents the results and a discussion on their possible implications; and Chapter 6 concludes the study.

1.1 RESEARCH OBJECTIVE AND HYPOTHESES

This study aims to demonstrate the efficacy of data mining techniques, specifically Least Angle Regressions (LARS) and Random Forests, in selecting appropriate explanatory variables for time-series models of equity index returns within a multifactor linear framework and a multifactor neural network framework. The consequent objective is to demonstrate a methodology for the inclusion of objective algorithm based variable subset selection techniques in the asset pricing field.

1.1.1 PRIMARY HYPOTHESIS

Hypothesis 1: The subset of variables selected by the feature selection algorithms (LARS and Random Forests) can be used to form viable models of equity indices returns.

> *Hypothesis 1.1*: The subset of explanatory variables selected by LARS can be used to form viable linear models.
>
> *Hypothesis 1.2*: The subset of explanatory variables selected by Random Forests can be used to form viable GRNN models.

1.1.2 SECONDARY HYPOTHESES

Hypothesis 2: The linear models, based on the subset of variables selected by LARS, perform equally well in the test periods as the equivalent GRNN models, based on the subset of variables selected by Random Forests [3].

> *Hypothesis 2.1*: The GRNN performs equally well with the subset of explanatory variables selected by LARS as the linear model does.

[3] *Hypothesis 2* is in fact a joint hypothesis, evaluating the performance of the feature selection techniques (LARS and Random Forest) jointly with the performance of the modelling techniques OLS-regression (linear) and GRNN (nonlinear). *Hypothesis 2.1* and *2.2* attempt to separate these effects.

Hypothesis 2.2: The GRNN performs equally well with the subset of explanatory variables selected by LARS as it does with the subset of explanatory variables selected by Random Forests.

1.1.3 TERTIARY HYPOTHESIS

Hypothesis 3: Variables selected by the feature selection techniques are period and industry independent.

Hypothesis 3.1: The subset of explanatory variables selected for each index are the same regardless of the sub-period evaluated.

Hypothesis 3.2: The subset of explanatory variables selected for each index are the same regardless of the industry index evaluated.

2. LITERATURE REVIEW

2.1 ON THE EFFICIENCY OF MARKETS

This study is not aimed at providing a significant commentary on the efficiency of markets. However, it is inevitable that a discussion of returns and specifically the mechanism through which they are determined entails some postulations on market efficiency. In a sense, modern financial theory begins with the Efficient Market Hypothesis. In offering a succinct definition of market efficiency and some testable implications Fama (1970) initiated much of the current thinking in financial theory. Fama (1970) defines an efficient market as one in which prices fully reflect all available information. He notes that this strict definition yields too few testable implications and so he devises his renowned three forms of informational efficiency – weak, semi-strong and strong form. Fama (1970)'s assumptions about the conditions under which equilibrium prices are obtained are notably stringent. Nonetheless, his theory suggests that investor's trade on information which they process accurately and efficiently. This theory implies that the ideal market has no noise traders. Perhaps an acceptable concession could see noise traders producing no overall effect on the market as he later contends (Fama, 1998).

Tests for the weakest form of informational efficiency in South Africa appear to show the JSE to be at least weak form efficient (Jefferis & Smith, 2005; Smith, Jefferis, & Ryoo, 2002). However, opponents of market efficiency may find support from another key observation, the presence of noise traders whose trades appear not to be sufficiently offset. This disputes Fama (1970)'s important assumption of investor rationality. Studies in the U.S. document the ability of noise traders, who by definition trade on no information at all, to have significant systematic effects on the market (De Long, Shleifer, Summers, & Waldmann, 1991; French & Roll, 1986).

Perhaps noise traders are simply acting on behavioural biases. Kahneman and Tversky (1979) notably evidence the presence of behavioural biases such as loss aversion in human decision

making under conditions of risk. It has also been shown that behavioural biases have significant effects on asset returns (Coval & Shumway, 2005). Accordingly, DeBondt and Thaler (1985) document evidence in the U.S. of market overreaction and subsequent mean reversion, suggesting inaccurate or at least inefficient information processing on the part of investors. This corresponds to the evidence on the presence of noise traders and points to possible systematic effects resulting from their positions being inadequately offset. Market overreaction and mean reversion has been shown to be evident in the JSE (Hsieh & Hodnett, 2011; Page & Way, 1992) and it is often credited with the profitability of contrarian investment strategies. Similarly, Jegadeesh and Titman (1993) find evidence of momentum effects in stock returns. They show abnormal returns to a strategy of buying stocks ranked as winners and selling losers (Jegadeesh & Titman, 1993). It has been suggested that momentum is simply the result of a delayed overreaction (Daniel, Hirshleifer, & Subrahmanyam, 1998; Hong & Stein, 1999). When considered in this way momentum may fall into the overreaction hypothesis as a precursor to mean reversion (Daniel *et al.*, 1998; Hong & Stein, 1999; Jegadeesh & Titman, 2001). Taken together there is clear evidence of time-series predictability in stock returns, going against the principle of weak form efficiency, and challenging the validity of the EMH as a general framework for equity markets.

Grossman and Stiglitz (1980) make the nuanced argument that for there to be even a small degree of efficiency in markets there must first be exploitable inefficiencies in order to encourage price discovery. Black (1986) describes how attentive investors (the so called 'smart money') profit from exploiting noise traders and the arbitrage opportunities they create. This assessment implies that without noise traders the market cannot be efficient whilst at the same time should their trades not be offset then the market is by definition inefficient. Behavioural finance incorporates elements of sociology and cognitive psychology, providing an outline within which the systematic effects of noise traders can be better understood. Shiller (2003) highlights the development of behavioural finance noting early behavioural models which include feedback models – for which he cites evidence extant in the natural experiments of bubbles and crashes that often characterise markets. Shiller (2003) further posits that prices are a bad forecaster of future cash flows as they are more volatile than the variable they attest to forecast (namely, dividends). Shiller (2003) proceeds to suggest reasons why the strategies of noise traders are not fully offset: First, short sale restrictions

may prevent arbitrageurs from taking opposing views; Second, there is a psychological and potential real cost to short sales – due respectively to the fact that losses may be unlimited and arbitrageurs do not know when the market will acknowledge mispricing – which is related to the ideas of delayed arbitrage and synchronisation risk (Abreu & Brunnermeier, 2002); Third, rational utility maximising investors may be concerned with the risk associated with irrational investors – a notion corroborated by the findings of De Long, Shleifer, Summers, and Waldmann (1990), who show that in some cases it may indeed be rational for investors to follow the strategies of noise traders. Stein (2009) illustrates how arbitrageurs face the 'crowd-trade effect' which is a coordination problem because traders do not know how many other market participants are trying to exploit the same opportunity resulting in uncertainty about the scope of the opportunity. Behavioural finance offers an attractive avenue through which to explore asset pricing. It is assumed in this analysis that behavioural factors and risk factors can be priced in precisely the same way using appropriate variables to proxy their exact nature.

The absence of an overarching behavioural theory is perhaps central to Fama's refutation of much of behavioural finance (Fama, 1998). With the Adaptive Market Hypothesis (henceforth, AMH) Lo (2004) attempts to provide a theory that reconciles the EMH and behavioural finance. Behavioural finance suggests that anomalies represent systematic divergence from rational utility maximising behaviour. The proponents of the EMH argue that such divergence is limited by arbitrage, preventing such effects from becoming systematic. Lo (2004) suggests an evolutionary approach where investors are viewed as organisms whose behaviour is shaped by natural selection dependent on the environment in which they are evolving[4]. As such, the AMH suggests that investors make decisions which are satisfactory and not necessarily optimal, a concept he dubs "satisficing". Prices in such a market reflect information conditional on the environment and the number and nature of investor groups. The AMH explains the deviations from expected utility maxing behaviour (or rationality) as an orthodox component of a market since investors need not optimise utility but aim rather to "satisfice". Kim, Shamsuddin, and Lim (2011) find evidence for the AMH in the time varying return predictability of the Dow Jones Industrial Index. They base their analysis on what they regard to be the two testable implications of the AMH: "First, the

[4] Appreciation of the nuances of this argument may necessitate a full reading of Lo (2004).

degree of market efficiency fluctuates over time. Second, the degree of market efficiency is governed by market conditions" (Kim *et al.*, 2011, p. 869). Similar evidence has been shown in a number of Asian and Latin American capital markets (Bonilla, Romero-Meza, & Hinich, 2006; Todea, Ulici, & Silaghi, 2009). At present no similar study has been published in South Africa. Nevertheless, evidence in this study showing that the subset of variables chosen by the feature selection tools is period dependent may constitute evidence for the first testable implication of the AMH.

2.2 ON ASSET PRICING THEORY AND PRICING FACTORS

The CAPM is built on the foundation of mean-variance optimisation and has since become a stable of asset pricing. Markowitz (1952) developed modern portfolio theory instituting the concepts of diversification and portfolio optimisation. Where Sharpe (1964) approached the problem of equilibrium asset pricing from an investor's perspective, Lintner (1965) analogously approached the problem from the point of view of a firm. Both arrived at the same conclusions, setting the backbones of a model that would go on to dominate asset pricing theory for decades. Under the CAPM asset returns are said to be a function of the return of the market portfolio. The key metric is thus the sensitivity of an asset to the return of the market portfolio, known as beta (β). The beta, Sharpe (1964) ascertains, represents a catchall measure for the sensitivity of an assets returns to systematic risk. The higher the systematic risk the higher the expected return. To derive this theory he first makes a series of stringent assumptions about the nature of the market and the nature of investors. He concedes that his assumptions are "highly restrictive and undoubtedly unrealistic" (Sharpe, 1964, p. 434), but contends that the test of a theory is not in its assumptions but in the validity of its implications. While the implications of the CAPM are generally acceptable, fitting neatly into financial theory, the assumptions however have since proved prohibitive leading to challenges of the theory itself.

Almost since its inception, numerous return anomalies not accounted for by the CAPM's beta have been observed. Banz (1981) makes the empirical observation that firms with small market

capitalisations are associated with higher returns. Graham and Dodd (1934)'s early work expressed the ability of value stocks to generate higher returns than growth stocks. Several studies have since documented the positive relationship between value stocks – as measured by high value to price multiples such as book-to-market equity, cash flow to price, or earnings yield ratios – to generate higher returns than growth stocks (Auret & Sinclaire, 2006; Connor & Korajczyk, 1988; Fama & French, 1992; Graham & Uliana, 2001; Lakonishok *et al.*, 1994; Rosenberg, Reid, & Lanstein, 1985; van Rensburg & Robertson, 2003). Based on similar evidence Fama and French (1992) build their three factor model (henceforth FF3). Their analysis shows size and BTM to be separate pricing factors and they show their model, which adds size and value factors to the traditional CAPM, provides relatively robust explanations for equity returns in the U.S. To generalise the Fama and French three-factor approach and explain bond returns, Fama and French (1993) add two factors (proxies term structure and default factor) to the model. They concede that the inclusion of factors to account for the size and value effects is simply because they have been shown to work as opposed to any coherent theory on their relation to systematic risks. In South Africa van Rensburg and Robertson (2003) aim to identify relevant style-based factors in returns. They too find evidence of a size and value effect that is not sufficiently explained by CAPM's beta. Basiewicz and Auret (2010) provide a serviceable version of the FF3 for the JSE. Carhart (1997) adds a momentum factor to FF3, constructing a four-factor model to explain cross-sectional variation in stock returns and demonstrating power another potential explanatory factor.

2.2.1 ARBITRAGE PRICING THEORY

The multitude of potential pricing factors evokes the need for a general framework by which asset returns may be modelled in using multifactor approach. Arbitrage pricing theory represents one such framework. Ross (1976) proposes the Arbitrage Pricing Theory (APT) as a general theory for the modelling of asset returns, and alternative to the CAPM. Its usefulness is in part due to its practical nature. The APT contends that the expected returns of an asset can be expressed as a function of a number of factors that account for possible sources of arbitrage (Ross, 1976). The APT is based on a no-arbitrage assumption, modelling returns in a manner to best ensure that there are no riskless returns (positive errors) available. The APT models returns to security *i*, with the following linear multifactor model:

$$r_i = b_{i,0} + b_{i,1}F_1 + b_{i,2}F_2 + \cdots + b_{i,J}F_J + e_i \quad i = 1,2,3,\ldots,N \qquad (1)$$

For a universe of *N* assets the explanatory variables are the factors F_j *(j = 1,2,...,J)*, with factor loadings $b_{i,j}$ and error term e_i. If in equilibrium all arbitrage opportunities are exhausted then the APT suggests that expected return of an asset *i* can be given by:

$$E(r_i) = r_f + (\delta_1 - r_f)b_{i,1} + (\delta_2 - r_f)b_{i,2} + \cdots + (\delta_J - r_f)b_{i,J} \qquad (2)$$

A risk free asset with return r_f is assumed to exist, and $\delta_j - r_f$ is the risk premium of factor *j* for an orthogonalised set of factors. Many multifactor models can be expressed as a special case of the APT. The CAPM can be expressed similarly as a special case, where $j = 1$ and $F_1 = r_m$. The trouble with the APT is also precisely the aspect that makes it flexible and reduces the need for a plethora of assumptions. The APT shows that there exists a method of modelling returns but is decidedly silent on which linear combination of factors best explain returns or even how many factors are required. The variety of factors used within the literature highlights the fact that there is little consensus about the most appropriate pricing factors. Proponents of market efficiency adhere to the belief that all factors proxy systematic risks. Nevertheless, so long as returns are not randomly generated, an optimal set of variables that provide the best explanation for returns is likely to exist. Ultimately, if such variables exhibit sufficiently steady relationships with returns then this implies the existence of an appropriate model for returns.

Two main schools of thought are prevalent in the literature concerning the selection of relevant variables for multifactor asset pricing. The first, is the factor-analytic approach. It contends that the true pricing factors (or risk factors) are unobservable and so statistical procedures such as principle component (PCA) and factor analysis (FA) are required to extract the common factors. The significant drawback of these techniques is that in the process of extracting commonalities, the factors lose their economic meaning. Gehr (1978) was perhaps the first to champion this approach for variable selection in the APT. There is a significant volume of literature supporting the factor-analytic approach (see Connor & Korajczyk, 1988; Dhrymes *et al.*, 1984; Lehmann & Modest, 1988; Reinganum, 1981; Roll & Ross, 1980; van Rensburg & Slaney, 1997). In an early

South African study Page (1986) provides evidence for the factor-analytic approach. van Rensburg and Slaney (1997) employ a novel adaptation of the procedure allowing for the identification of observable proxies for the analytic factors. They find that a two factor model offers an adequate description of returns and that JSE Actuaries Gold Index and JSE Industrial Index could act as proxies for the first two factors of the analytic model. The factor-analytic approach resolves the problem of the identification of relevant factors, but the literature shows divergence in the estimation of the optimal number of factors.

The second approach to the selection of relevant variables is characterised by the pre-specification of variables. In perhaps the quintessential study in this area Chen, Roll, and Ross (1986) attempt to identify the economic state variables that affect stock returns. They contend that "no satisfactory theory would argue that the relationship between financial markets and the macroeconomy is entirely in one direction" (Chen *et al.*, 1986, p. 384). Indeed, subsequent studies in South Africa, Asia, Europe and the U.S. have shown the relationship between macroeconomic variables and stock returns to be complex and dynamic involving a degree of cointegration (Cheung & Ng, 1998; Kwon & Shin, 1999; Lettau & Ludvigson, 2001; Moolman & Du Toit, 2005; Nasseh & Strauss, 2000; van Rensburg, 2000). Of their limited set of economic and non-equity asset variables Chen *et al.* (1986) find industrial production, changes in risk premium and changes in the yield curve to have the offer the most robust relationships with returns. They find frequently touted variables such as changes in oil prices and consumption to be insignificant. The idea that economic factors affect equity returns opens up a considerably large set of potential explanatory variables and a challenging debate over what exactly constitutes an economic state variable. Notionally, any economic variable with a significant pervasive effect on stock returns could be considered a descriptor of the overall economic state, and hence a state variable.

Clare and Thomas (1994) examine the ability of macroeconomic variables to explain beta and size sorted portfolios as part of an APT model. The evidence points to a number of macroeconomic variables being significant for beta sorted portfolios but only two factors prove significant for size sorted portfolios. Asprem (1989) investigates the use of APT-type models to explain the returns of ten European countries using international equity indices and a number of local and aggregated

macroeconomic variables. The results indicate that many macroeconomic variables show significant relationships with equity returns in European countries. Local variables are shown to be of particular importance. Evidence for the pre-specification approach can also be found in emerging market studies. Singh (2008) investigates the dollar/rupee exchange rate, wholesale price index and call money rate amongst several possible explanatory variables for the Indian stock market. The call money rate and the exchange rate are discovered to be significant. Bilson *et al.* (2001) define an APT based model to describe the returns of twenty emerging stock markets (ESMs). They find that the microeconomic effects of fundamental variables (price-earnings and dividend yield ratios) dominate but other local macroeconomic variables are also significant. In South Africa van Rensburg (1996) applied the pre-specification approach to the APT demonstrating that unexpected changes in the Dow-Jones Industrial Index, short term interest rates and term structure provided significant relationship with the JSE All-Share index returns. van Rensburg (2000) appends his previous methodology devising a hybrid factor-analytic pre-specification approach. The results suggest that in the two factor-analytic proxies subsume the observed significant effects of gold prices, long-term bond rates, the Dow-Jones Industrial Index and the level of reserves (gold and foreign exchange).

This study takes the view that the true relevant innovations are indeed latent and likely unobservable, but that a relevant subset of variables may be used as a proxy for the true variables. Chen and Jordan (1993) directly compare the factor-analytic and pre- specification approaches on a sample of sixty-nine industry portfolios and thirty equally-weighted portfolios in a hold-out sample. Their observations suggest that both approaches appear to construct viable models. Their factor-analytic approach does lead to fractionally better models in-sample due to the factors (from PCA or FA) being sample specific. Results from their out-of-sample analysis suggests that the pre-specification may lead to more generalisable models. Chen and Jordan (1993) conclude that there is little lost in moving between the two approaches.

2.3 ON MACROECONOMIC FACTORS AND EQUITY RETURNS

It is worthwhile noting that this analysis is not aimed at making particular deductions on the nature of the relationship between specific macroeconomic variables and returns. Rather evidence of such relationships is viewed in the broader scheme of a description of the economic state. Nevertheless the literature –part of which has been appraised in Section 2.2 – provides a clear indication that such specific relationships exists. Table 2.1 briefly highlights some of the literature related to the specific variables used in this study in support of their inclusion.

Table 2.1 Literature pertaining to the relationship of macroeconomic variables and returns.

Variable	Studies	Comments
Commodity Prices		
Oil	Cheung and Ng (1998), Chen *et al.* (1986) and Clare and Thomas (1994).	A negative relationship may be plausible under the premise that oil prices are negatively correlated with real economic activity. A positive relationship is also plausible if returns are also directly related to oil prices e.g. oil industry index.
Gold	Clare and Thomas (1994),van Rensburg (1995), , van Rensburg (2000).	Gold is frequently spoken of as a 'risk-off' investment safe haven. The ALSI however is heavy on mining stocks. van Rensburg (2000) observes a positive correlation between unanticipated gold price changes and returns.
Inflation Metrics		
CPI	Bodie (1976), Bilson *et al.* (2001), Chen and Jordan (1993), Chen *et al.* (1986), Fama (1981) ,Humpe and Macmillan (2009) and van Rensburg (1996).	The literature documents a persistent negative relationship between inflation and equity returns. Fama (1981) suggests that this is likely due to the effect of inflation on real economic activity. None of the studies directly used PPI as an explanatory variable, but there is no reason to suggest that the intuition would be dissimilar.
PPI		
Government Bonds		
Long Term Government Bonds	Asprem (1989), Clare and Thomas (1994)	The literature suggests a positive relationship.
Short-Term Treasury Bills	Clare and Thomas (1994), Humpe and Macmillan (2009),	
Term Structure	Asprem (1989), Clare and Thomas (1994), Chen and Jordan	The relationship between term structure and is in general found to be negative.

	(1993),Chen *et al.* (1986), van Rensburg (1996),	
Monetary Supply		
M1	Cheung and Ng (1998), Flannery and Protopapadakis (2002), Humpe and Macmillan (2009)	The theoretical relationship between monetary aggregates and stock returns is complicated.
M2	Bilson *et al.* (2001)	
Real Activity		
Employment/ Unemployment	Asprem (1989), Clare and Thomas (1994), Flannery and Protopapadakis (2002)	It is expected that an increase in employment represents an increase in the economic activity which should be a positive for equities. The empirical relationship an interpretation is rather more complex.
Consumption	Cheung and Ng (1998), Chen *et al.* (1986), Humpe and Macmillan (2009)	While Chen *et al.* (1986) prominently find no significant relationship other studies using a multifactor approach document a marginally positive. A positive relationship would appear to be in line with the notion that consumption growth is a good economic signal.
Balance of Trade	Asprem (1989)	Asprem (1989) shows significant relationships between European stock indices and changes in exports and imports.
Industrial Production	Asprem (1989), Bilson *et al.* (2001), Chen and Jordan (1993),Chen *et al.* (1986), Clare and Thomas (1994), Humpe and Macmillan (2009)	Increases in industrial production are generally viewed as a positive for economy representing an uptick in real activity. The literature reflects a positive relationship between industrial production and stock returns.
Retail Trade Sales	Clare and Thomas (1994)	Growth in retail sales is commonly assumed to be a positive indicator for the overall economic state. Clare and Thomas (1994) find a positive relationship.
Credit Expansion		
Domestic Credit	Clare and Thomas (1994)	The only paper reviewed which evaluated credit expansion looked at growth in private credit, finding an indeterminate relationship. The general intuition however suggests that growth in credit is a good indicator of possible future spending and therefore a positive indicator of real activity.
International Indices		
Dow-Jones Industrial Index/ S&P	Asprem (1989), van Rensburg (1995), van Rensburg (1996) and van Rensburg (2000)	Studies that looked at U.S. market indices as explanatory variables for international stock returns have generally found a significant

		positive relationship in line with the idea of international equity co-movements (Bekaert, Hodrick, & Zhang, 2009). Perhaps the international equity index returns act as a proxy for the state of the global.
World Indices	Bilson *et al.* (2001)	Bilson *et al.* (2001) evaluates the use of the MSCI World Index as part of a macroeconomic APT model.

A large number of macroeconomic variables have been shown to have a significant relationship with stock returns. The methodology used in this analysis enables the simultaneous evaluation of all the macroeconomic variables presented.

2.4 ON DATA MINING AND FEATURE SELECTION

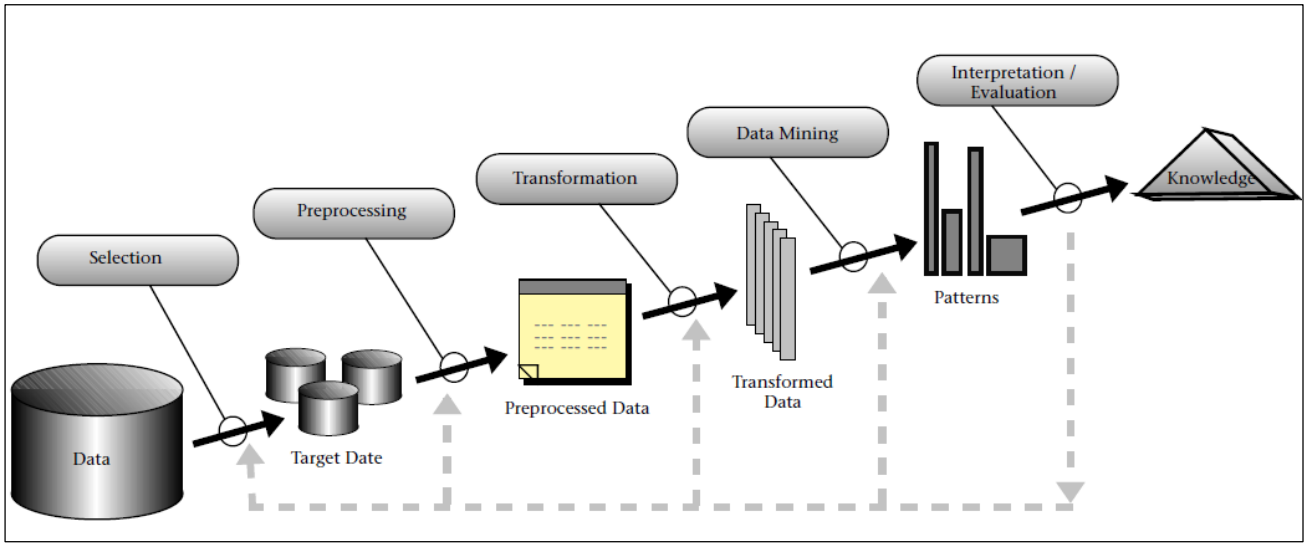
The term data mining has been used as an unfortunate pejorative in the past. An example of this can be found in Schwert (1990), where on discovery of corroborating evidence for Fama's (1981) findings the author suggests that "it is unlikely that [sic] 'data mining' could explain Fama's results" (Schwert, 1990, p. 1256). More recently, the term has come into common usage as a legitimate scientific framework for extracting knowledge from data. Data mining offers a blend of statistics and computer science tools and is at the core of the developing field of data science. Data mining tools are good at finding relationships. However, they are not good at defining why the relationships exist. For that, the impetus returns to conventional theory.

The birth of data science could be attributed to Tukey (1962), a manuscript in which he illustrates the limitations of statistics as it was, and its evolution into something incorporating a plethora of new techniques. He makes a number of inferences on the future adoption of new methods and suggests that computers will be 'important but not vital', putting primary importance on "our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria, and abstract results" (Tukey, 1962, pp. 63-64). Notwithstanding a number of significant intervening contributions to the field, Fayyad, Piatetsky-Shapiro, and Smyth

(1996) first articulated the idea that data mining and many other algorithmic techniques are in fact part of a greater field of study which they called Knowledge Discovery in Databases (henceforth KDD). They write: "In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process" (Fayyad *et al.*, 1996, p. 39); A step in which patterns are extracted from data using algorithms. Cleveland (2001) more overtly proposes the evolution of a field at the juncture of computer science and statistics, recommending a pathway for this eventuality.

The basic data science process consists of five main steps illustrated in Figure 2.1. These are: 1) initial data selection or data base construction; 2) pre-processing and analysis; 3) transformation of the data to better enable inferences; 4) data mining; 5) interpretation and evaluation (Fayyad *et al.*, 1996). Depending on the nature of the analysis, the data, and research question, the steps may be combined and a number of iterations or loops between steps may be necessary.

Figure 2.1 Overview of the steps in the data science process.



(Source: Fayyad *et al.* (1996))

Of particular import is the data mining step, into which the two key elements of this study, modelling and most importantly feature selection, fit. Feature selection is a term used in the computer science literature to describe the process of selecting a subset of variables from a wider set.

Berry, Burmeister, and McElroy (1988) address the problem of variable selection for an APT model by outlining some criterion. Their criterion is threefold: First, variables must have pervasive effects on the asset returns being modelled. Second, the variables must be unpredictable at the beginning of the period. Finally, variables must have a significant influence on expected returns. The third criteria is arguably the most important. The first criterion appears to be severely restrictive as it suggests that only systematic factors, which cannot be diversified away, qualify. This then requires a determination of what constitutes a systematic risk factor. As the literature demonstrates, some firm-specific variables such as BTM and firm size have significant influences on returns. However, it is not clear what systematic factors these variables proxy for. The second criteria is also restrictive. It suggests that all variables must be generated by a random process. If this criterion is followed strictly then a finding that the changes in a particular market index do not follow a random walk disqualifies the use of that index in an APT model. This seems to contradict the fact that market indices, such as the Dow-Jones Industrial Index, have been shown to have significant influences on South African stock indices (van Rensburg, 1995, 1996).

In data science feature selection has been divided into three classes of algorithms: a) filters, b) wrappers, and c) embedded methods[5]. Filters select features based on some sort of defined criteria that is independent of any supervised learner or classifier (Guyon & Elisseeff, 2003). An example of a basic learner is an OLS-regression and an example of a common criteria is Pearson's correlations. Wrappers use a learner as a black box to concurrently examine the usefulness of the feature subset and select the relevant features (Guyon & Elisseeff, 2003; Kohavi & John, 1997). Kohavi and John (1997) explain that the training set (input/output set) is presented to an induction algorithm which induces the learner. The feature selection algorithm searches for a good feature subset using the induction algorithm that evaluates the subsets within the native function (Kohavi & John, 1997). Some studies have attempted to detect patterns in stock returns using genetic algorithms as wrappers (Kim & Shin, 2007; Kim & Han, 2000). A common criticism of wrappers is that they quickly become computationally demanding (Guyon & Elisseeff, 2003). Embedded methods may offer a computationally efficient alternative for evaluating large data sets. Rather

[5] Readers unfamiliar with the classification of feature selection techniques may find it useful to read Liu and Motoda (1998)

than treating the learner as a black box embedded methods select feature subsets as part of the process of training of the learner (Guyon & Elisseeff, 2003). Decision trees are prominent examples of embedded feature selection techniques. Another prominent example the LARS algorithm used in this study.

2.4.1 LEAST ANGLE REGRESSIONS (LARS)

The most common form of feature selection in asset pricing literature is the forward stepwise regression (see for example Auret & Sinclaire, 2006; Fama & French, 1992; Fama & French, 1993). LARS was conceived by Efron *et al.* (2004) as a novel algorithm that solves the main problem associated with the classic forward stepwise regression. A stepwise regression involves an initial linear regression with the explanatory variable most correlated to the response. Subsequent variables are added in a stepwise fashion projected in sequence orthogonally on the growing variable subsets. The disadvantage with this process is that it often leads to a greedy fit disregarding useful variables correlated with the variable subset even if they add explanatory power (Efron *et al.*, 2004; McCann & Welsch, 2007). A forward stage-wise regression advances in a more guarded manner taking smaller steps towards the final model. Efron *et al.* (2004) procedurally place LARS somewhere in between a forward stage-wise regression and a forward stepwise regression. The geometry of the LARS algorithm means that a simple modification provides the least absolute shrinkage and selection operator (LASSO) as outlined by Tibshirani (1996). Accordingly, the LARS algorithm can be used as a more efficient way to fit Lasso models – a property that is especially useful for high dimensional optimisation problems (Kuhn & Johnson, 2013). The LARS algorithm has demonstrated the ability to select an appropriate subset of covariates from a large set of data in a number of different fields (Efron *et al.*, 2004; Keerthi, 2005; McCann & Welsch, 2007; Yuan & Lin, 2006). As an example of the computational efficiency of the LARS algorithm Efron *et al.* (2004) show that for a complex quadratic model an unmodified Lasso uses one hundred and three steps versus only sixty-four steps for LARS.

LARS uses Mallows' (1973) C_p estimate of prediction error as an optimisation criterion, leading to an optimal choice of selected factors. The LARS algorithm only examines linear relationships

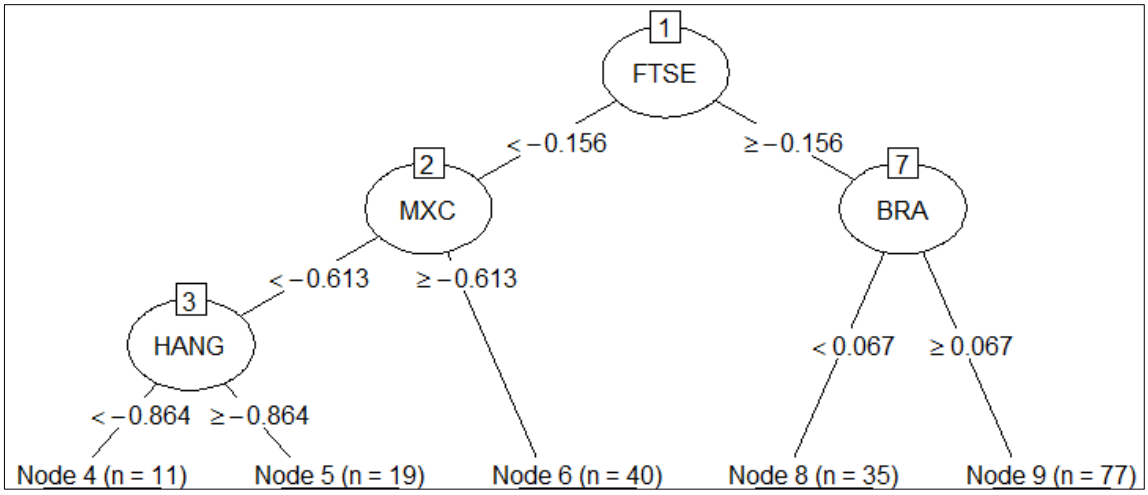
and therefore the selected variables can be placed directly into an APT framework. Wang and Tan (2009) use LARS as a feature selection technique to determine an appropriate subset of explanatory factors for an APT model. Feature selection is conducted on sixty-five variables from the balance sheets, income statements and cash flow statements of firms listed on the major U.S. exchanges from 1999-2003. The returns of the companies were used as response variables to construct models for each year. The results point to the market value of equity and earnings-per-share as the most pervasive factors although the overall model for returns selected by LARS appears to differ every year. Their conclusion is that LARS is an effective method of feature selection for the APT. Cakmakli and van Dijk (2010) investigate the predictability of U.S. monthly excess stock returns employing LARS to conduct feature selection for a set of one hundred and six (106) macroeconomic variables. Their results are supportive of LARS as a feature selection technique and of the use of macroeconomic variables in an APT framework.

2.4.2 Decision Trees and Random Forests

Decision trees are a machine learning technique adopted with the aim of inductively learning a description of a function from a series of inputs and outputs. Like LARS, inductive decision trees can be used to identify an optimal set of variables. However, unlike LARS they are capable of modelling nonlinear relationships even when variable types differ (Breiman, 2001; Breiman *et al.*, 1984; Breslow & Aha, 1997). Decision trees do not require any assumptions about the nature of the relationship between the input and output variables they simply learn the most appropriate tree from the data based on their splitting criteria (Kuhn & Johnson, 2013). Inductive decision tree learning involves splitting the data into smaller more homogeneous groups by determining an explanatory variable (feature) on which to conduct a split and then repeating this process recursively on each new cluster to 'grow' a tree (Breiman, 2001; Kuhn & Johnson, 2013). The variables which provide the best splits are the variables that result in the largest homogeneous groups of data. These are added to interior nodes in a hierarchical manner. Each interior node in a tree corresponds to one input variable, but not all input variables are used to construct interior nodes (Kuhn & Johnson, 2013). Decision trees stop growing when additional splits no longer divide sufficient amounts of data. A way to regulate this is to define a minimum size of terminal node. Figure 2.2 illustrates the basic structure of a decision tree. The tree is based on modelling

ALSI returns as a function of a group of international stock indices which include the FTSE 100 (FTSE), Hang Seng (HANG), Mexican Bolsa IPC Index (MXC), and Bovespa Brasil Mexican (BRA). This serves to illustrate the interior nodes, splitting criterion, and terminal nodes. Whilst the basic structure of decision trees remains the same many algorithms have been developed to provide different splitting criterion and implementations.

Figure 2.2 Basic decision tree for ALSI returns.



Note: The features on which splits occur are encircled and the splitting criterion is defined on each branch. The number of observations in each terminal node are shown at the base of the tree. This particular decision tree uses an evolutionary algorithm to optimise feature selection but the structure is the same for most decision trees..

The Classification and Regression Trees algorithm (CART) is a commonly used decision tree learning algorithm that uses a Gini impurity measure or Gini index to compute the best splits (Breiman *et al.*, 1984). Gini impurity is a misclassification measure that is computed by summing the probability of each item being chosen and multiplying it by the probability of a mistake in categorising that item. Quinlan (1993) develops the *C4.5* decision tree learning algorithm based on his *ID3* algorithm. It uses the information gained by splits, based on information entropy measures before and after the split, as splitting criteria. A key problem with the decision tree algorithms is that they rely on greedy heuristics (for splitting) tending towards over-fitting the data. As such a litany of techniques to reduce the complexity of trees exist. For example, Breslow and Aha (1997) show a number of techniques used to simplify and 'prune' decision trees. Gray and Fan (2005) show the efficacy of using an evolutionary algorithm to optimise the splitting process reducing some of the problems associate with decision trees. Grubinger, Zeileis, and Pfeiffer (2011) introduce a decision tree learning algorithm which searches for globally optimal decision

trees using an evolutionary algorithm for optimisation over the global search space. As embedded feature selection techniques, feature selection is merely one component of what decision trees do. An inductive tree which has been trained on a data set may be used to conduct predictions. Indeed, some studies have used decision trees for both feature selection and prediction (Hargreaves & Hao, 2013). Other studies have simply utilised the feature selection capabilities leaving estimation and modelling to other techniques. Thawornwong and Enke (2004) for example first employ a decision tree using the C4.5 algorithm and then model the variables subset using neural network ensembles and linear regressions.

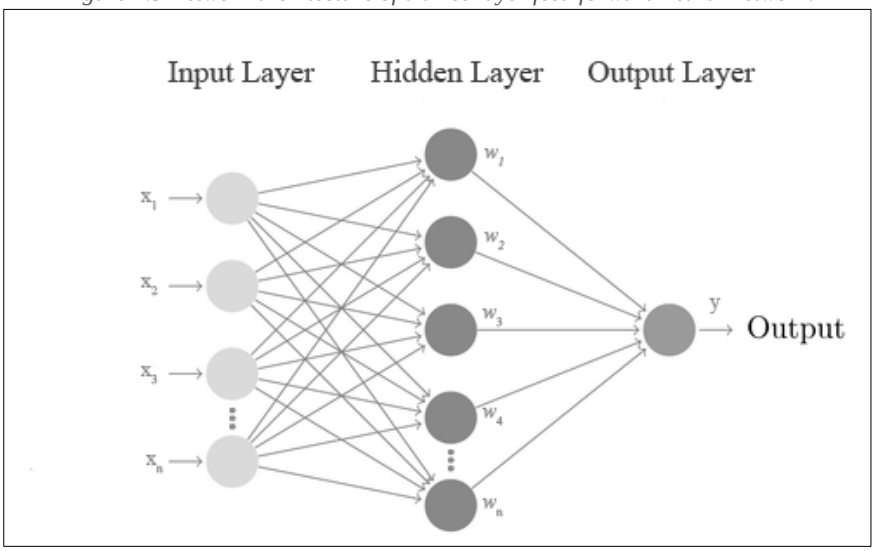
Breiman (2001) proposed Random Forests as a novel way to avoid the over-fitting problems that have characterised single decision trees. Random Forests combine Breiman (1996)'s concept of bagging with the random selection techniques championed by Amit and Geman (1997). A Random Forest is an ensemble feature selection technique which averages over a number of decision trees grown using the CART methodology (Breiman, 2001). In using a number of trees as opposed to a single tree the Law of Large Numbers implies that Random Forests will not overfit the training data. Breiman (2001) demonstrates how the use of bagging and the introduction of random features produces good classification results. One of the disadvantages of Random Forests is feature redundancy. The splitting at each node is only locally optimal, so there is no inherent test to ensure that the split is performed on a feature dissimilar to previous splits (Deng & Runger, 2012). The obvious solution would be to augment the decision tree algorithm compelling global optimality tests at splits. Following that principle new algorithms have been proposed such as regularised Random Forests (Deng & Runger, 2012), and guided regularised Random Forests (Deng & Runger, 2013). At present, these methods are fairly new and untested. The concern is that such adjustments may reduce redundancy at the expense of selecting a sub-optimal feature subset. The Boruta algorithm is a wrapper for Random Forests which adds additional randomness to further elucidate the relevant features[6] (Kursa & Rudnicki, 2010). In this study Random Forests are implemented using the Boruta algorithm.

[6] Details on the specific mechanism by which this is done can be found in Chapter 4.

2.5 ON ARTIFICIAL NEURAL NETWORKS

Numerous studies illustrate the superior ability of nonlinear modelling of stock returns, specifically using Artificial Neural Networks (ANN's) (Cao, Leggio, & Schniederjans, 2005; Cao *et al.*, 2011; Gibbons, 1982; Qi, 1999; Refenes, Zapranis, & Francis, 1994)[7]. An ANN is a group of linked nodes in sparse mimicry of the immense networks of neurons that make up the brain. Cybenko (1989) shows how sigmoidal transfer functions enable neural networks to theoretically approximate any nonlinear relationship. This is a particularly powerful capability and it means that ANN's can model data without making *a priori* assumptions about the nature of the data generating process (Schalkoff, 1997). Consequently, ANN's are ideal for modelling data where the exact nature of the DGP is unknown. Rosenblatt (1958) devised the perceptron as the algorithm that defined input – output relationship based on the learning theory hypothesised by Hebb (1949). It was not until the derivation of the backpropagation algorithm by Werbos (1974) that neural networks research really gained momentum.

Figure 2.3 Network architecture of a three layer feed forward neural network.



A multi-layer perceptron is a basic neural network that is capable of mapping an input vector onto an output vector. Figure 2.3 demonstrates the architecture of a three layer neural network. The input layer accepts the data pattern. There may be any number of hidden layers in between the

[7] Also see Dase and Pawar (2010) and Atsalakis and Valavanis (2009) for reviews.

input and output layers. These layers serve to retrieve output from the input layer re-weighting them and passing them on via an activation function. The output layer obtains values from the hidden layers weights them appropriately and produces the target values. ANN's model outcomes using an intermediary set of hidden variables which are linear combinations of the original predictors with no hierarchical structure (Schalkoff, 1997). The linear combination of predictors in the hidden layer are oft transformed by a transfer function - typically a nonlinear function such as a sigmoidal function *g(u)* (Cybenko, 1989; Kuhn & Johnson, 2013).

$$f_k(x) = g\left(b_{0k} + \sum_{i=1}^{n} x_j b_{jk}\right), \qquad (3)$$

where $g(u) = \frac{1}{1+e^{-u}}$

Equation (3) illustrates the function by which an ANN defines targets. The coefficients b_j are equivalent to regression coefficients and coefficient b_{jk} is the weight of the jth predictor on the kth hidden unit. The model is often fitted using some method of minimising the sum of squared residuals. The backpropagation algorithm has shown particular efficiency at this operation (Rumelhart, Hinton, & Williams, 1988). Network topology varies with the number of layers in the ANN (Schalkoff, 1997). For example, an ANN with no hidden layers is able to place a hyperplane in the input space. An ANN with one hidden layer is able to define a decision boundary comprising a region of the input space. ANN's with two hidden layers can describe an arbitrary decision boundary. The use of a sigmoid activation function enables ANN's with even one hidden layer to approximate any decision boundary (Cybenko, 1989; Schalkoff, 1997).

Once a neural network has been fitted (trained) it can then be used to provide forecasts in the same way as an estimated linear model. Studies by Kryzanowski, Galler, and Wright (1993), Qi (1999)[8], Tkacz (2001), and Callen, Kwan, Yip, and Yuan (1996), *inter alia*, show the ability of neural networks to forecast financial data. White (1988) was amongst the first researchers to use neural networks to forecast stock returns. He focuses on a single stock IBM and shows that feed-forward neural networks can decipher undetected regularities in asset returns. In more recent studies Leigh, Paz, and Purvis (2002) and Armano, Marchesi, and Murru (2005) evaluate neural network

[8] Although the conclusions he reaches from his initial empirical tests are disputed, see Racine (2001).

forecasts for US equity indices and find that their forecasts are serviceable. Patel and Marwala (2006) develops a neural network based system which is used to provide forecasts for next day returns for the Dow-Jones Industrial Average, NASDAQ 100, Nikkei 225 and JSE All-Share indices. A particularly noteworthy study by Cao *et al.* (2005) examines the usefulness of neural networks in predicting stock returns in China. The study takes the intriguing step of using factors from the FF3 and CAPM as input to their neural network and then comparing these forecasts to that of linear versions of the models. They find that in all cases the neural networks are better at forecasting, and the addition of variables in going from the CAPM to the FF3 only serves to strengthen this observation (Cao *et al.*, 2005). Refenes *et al.* (1994) evaluates the ability of neural networks to model stock performance using an APT-type framework. Their examination with dynamic APT-type models shows neural networks significantly outperforming linear regressions in both goodness of fit tests and out-of-sample generalisation tests. In a study on stock ranking Refenes, Azema-Barac, and Zapranis (1993) find that neural networks outperform traditional APT models. They surmise that the smooth interpolation properties neural networks possess enable better in-sample fit and out-of-sample generalisation. Qi (1999) estimates linear and neural network models of S&P 500 index returns constructing direct comparisons between neural networks and linear OLS estimations. In all cases the neural networks are shown to be superior.

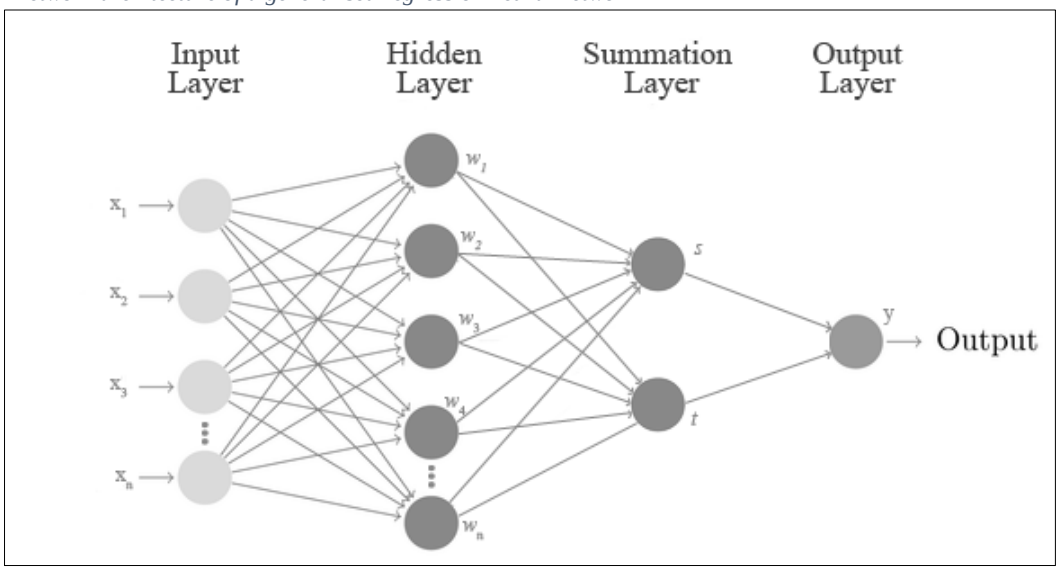
2.5.1 General Regression Neural Network (GRNN)

The particular ANN used in this study is the GRNN. The GRNN is a nonparametric feed forward neural network proposed by Specht (1990a). It is based on nonlinear regression theory and so can be used when the assumption of linearity is not justified. The functional implementation of the GRNN is an adaptation of the radial basis function[9]. The GRNN consists of four layers: 1) An input layer for which the number of nodes is dependent on the number of explanatory variables. 2) A hidden layer (pattern layer) in which a training pattern is presented and used to develop outputs. 3) A summation layer (also a hidden layer) which comprises two neurons *s* and *t* that take output from the pattern layer and perform summation and single division, respectively. The unit *s*

[9] See Powell (1977), and Park and Sandberg (1993) for details and empirical evaluations of RBF's.

sums the weighted responses from the pattern layer while the unit *t* computes un-weighted responses. 4) The output layer divides the output from *s* by that the output from *t*.

Figure 2.4 Network architecture of a generalised regression neural network



The network architecture of the GRNN is displayed in Figure 2.4. As opposed to backpropagation, which may take a large number of iterations before convergence, Specht (1990a) proposes a one-pass learning algorithm that converges to the underlying surface regression. Like most other neural networks the GRNN requires no prior knowledge of the functional form. A suitable probability density function is determined from the observed data through kernel density estimation, or more specifically Parzen's window estimation (Parzen, 1962).

The GRNN is incredibly robust and has been used in a wide range of forecasting and classification problems in a number of fields (see for example Kulkarni, Chaudhary, Nandi, Tambe, & Kulkarni, 2004 and; Li, Bovik, & Wu, 2011). However, the literature on financial applications of the GRNN is relatively thin. Chen (1994) attempts to estimate daily trends in closing prices of the Dow-Jones Industrial Average. In his study the GRNN is shown to outperform linear models in trend estimation and prediction. In a fascinating study Steiner and Wittkemper (1997) develop a nonlinear dynamic capital market model under the preconception that the linear forms of the CAPM and APT are too restrictive. The investigation incorporates thirty-one German stocks and

first estimates parameters for the coherent market hypothesis model[10], then performs out-of-sample forecasts using a GRNN. Forecasts are used to rank stocks and they demonstrate that the best eight ranked stocks achieve annual returns 25% higher than the market portfolio.

Mei (2010) uses a GRNN to forecast the closing prices of thirty-seven Taiwanese equities. The GRNN is shown to result in generally viable forecasts (Mei, 2010). A GRNN is amongst the neural networks used by Enke and Thawornwong (2005) in their application of data mining tools to forecast financial data. As in this study, and indeed their own previous study (Thawornwong & Enke, 2004), they initially use decision trees to conduct feature selection for their APT-type model. Their findings indicate that a GRNN performs better than a linear model in forecasting returns, but they also find that other neural networks such as Specht's (1990b) probabilistic neural network perform better. No view on the superiority (or otherwise) of a GRNN over comparable neural networks is presented here. Of the multitude of neural networks available the GRNN is used simply because it provides a function that best mimics regressions frequently used in modelling the APT, and therefore leads to more straight forward comparisons.

2.6 SUMMARY OF THE LITERATURE REVIEW

This chapter begins with a review of some of the seminal asset pricing literature and progresses to incorporate the broader applications of the APT framework. In particular, literature pertaining to the pre-specification approach to variable selection is reviewed along with a brief analysis of some of the many variables found to have significant effects on returns. Altogether, the literature on macroeconomic variables suggests that there is an assortment of variables which represent possible pricing factors. The variable selection techniques used in this study are outlined together with the limited empirical work that has thus far incorporated them in the financial field. Lastly, a brief review of neural network literature is conducted and the chapter concludes with a look at the research relating to the particular neural network used in this study, the GRNN.

[10] See Vaga (1990) for the full description of the coherent markets hypothesis and the related implications.

3. THEORETICAL MOTIVATIONS FOR THE EMPIRICAL FRAMEWORK

In this chapter theoretical motivations are provided for decisions made on some of the more contentious issues of the empirical framework. Empirical study necessitates a number of judgement calls on both contentious and more mundane aspects, requiring one to decide upon a perspective. In the main, perspectives are highlighted and motivated within the literature review presented in Chapter 2 and the methodology outlined in Chapter 4. This Chapter highlights a few perspectives held as undercurrents in the study but are not overtly discussed and motivated elsewhere. The aim of this chapter is to provide further clarifications. As this investigation demonstrates a methodology for the incorporation of feature selection techniques in the multifactor time-series modelling of asset pricing in South Africa, lucidity is essential for researchers to adapt and revise the presented methodology, and perhaps extend it to conventional cross-sectional APT models.

3.1 MODELLING THE EXPLANATORY VARIABLES

The assertion made in this study in regards to macroeconomic variables takes its cue from Chen *et al.* (1986) who posit that macroeconomic variables are descriptors of the economic state. This study advances that view in South Africa by suggesting that the local stock indices actively price in the economic state. This takes place in real-time and because relevant economic state data is not available continuously, it is conceivable – or perhaps it is anticipated – that market participants, in lieu of relevant announcements, continually price in their own estimates of the fundamental and economic state. Indeed this is the presumption that leads researchers to create unanticipated explanatory innovations (Chen *et al.*, 1986; van Rensburg, 1996, 2000). The intuition here is that only the unanticipated component of macroeconomic variables represent relevant risk factors as the anticipated part should already be incorporated in returns. The conjecture that investors continuously price in their estimates of the economic state is perhaps an explanation for the observed high trading day variances (Black, 1986; French & Roll, 1986). It could quite possibly be a manifestation of the price discovery process. A process that perhaps requires a degree of

mispricing *á la* Grossman and Stiglitz (1980). Altogether, this appears consistent with French and Roll (1986) who show evidence corroborating the hypothesis that the process of trading introduces noise. To make more definitive suppositions on the subject would require one to delve into the poorly defined domain of market microstructure, and the inevitable debate between Hasbrouck's (1995) information shares model and Gonzalo and Granger's (1995) common factor component weight model. Regardless, the overall thrust is that the macroeconomic variables are only relevant to the degree that they represent the economic state. Furthermore, the belief held in this study is that the variables used are not risk factors in and of themselves, but rather proxies for the true descriptors of the economic state. Since the goal here is to investigate whether the variable selection techniques used can identify the descriptors that proxy for the overall economic state, it may be redundant to decompose the possible features into their unanticipated and anticipated components. Therefore, the multifactor time-series models in this study are not strictly APT models pricing risk factors, but rather APT-type models modelling relationships. They take the spirit of the APT to price returns of index j as a function of descriptors of the economic and fundamental state.

$$R_j = f(proxy\ i\ for\ economic\ state, proxy\ k\ for\ fundamental\ state) \qquad (4)$$

An intriguing concurrent investigation examines if the function $f(\bullet)$ is sub-period dependent. If this is the case then, *ceteris paribus*, this may be deemed to constitute evidence for the AMH. Additionally if test period performance of the models is period dependent then this may also constitute evidence for an adaptive framework. In a study in the U.S. Kim *et al.* (2011) examine one of the testable implications of the AMH – time-varying market predictability. If in this study returns can be better predicted in one period than the other then this may be said to constitute evidence for the AMH.

3.2 DETERMINING THE MODEL FEATURES

Feature selection provides significant benefits to modelling complex problems. However, there is a distinction in the ultimate outcome of the process depending on the methods used. Some feature

selection techniques focus on delivering a subset of useful features, while others identify all relevant features (Guyon & Elisseeff, 2003; Kohavi & John, 1997). The distinction is minor, but if the goal is to determine explanatory models for complex variable relationships then the consequences for the eventual model may be significant. Ideally, the feature selection mechanisms should give the exact subset of features that result in the best model. Unfortunately the best model is not easily defined. The definition of relevance and irrelevance is critically important in the determination of the best subset of features (John, Kohavi, & Pfleger, 1994). Perhaps one of the more common definitions states that variable X_i is relevant *iff* there exists a x_i and y for which $p(X_i = x_i) > 0$ and $p(Y = y \mid X_i = x_i) \neq p(Y = y)$ (Gennari, Langley, & Fisher, 1989, Section 5; as cited by, John *et al.*, 1994). This definition is extraordinarily broad and perhaps not effective for feature selection. John *et al.* (1994) describe the necessity of dividing the definition of relevance into two components: weak relevance, which comprises features that sometimes but not always contribute to an explanatory model, and strong relevance, which incorporates features that in no instances can be discarded without loss to the overall model. It is then trivial to deduce the definition of feature irrelevance as the antithesis of the combined relevant sets.

With the emphasised definition for relevance it is apparent that feature selection can fail in two key ways. Over-selection comprises selecting more features than there are in the true subset that gives the best model. Under-selection comprises selecting fewer features than in the true subset. Perhaps more significant than feature irrelevance is the concept of feature redundancy. In the context of feature selection an exact definition of redundancy gets convoluted when one considers that redundancy has to incorporate features that are correlated with other features but are also weakly relevant. Yu and Liu (2004) construct a two-fold definition for redundancy. First they define a Markov blanket for a feature X_i in a set of possible features X as follows: For feature X_i where $M_i \subset X$ ($X_i \notin M_i$), M_i is said to be a Markov blanket for X_i *iff*

$$P(X - M_i - \{X_i\}, Y \mid X_i, M_i) = P(X - M_i - \{X_i\}, Y \mid M_i) \qquad (5)$$

,where Y is the vector of outputs. For the set of features X a feature is redundant *iff* it is weakly relevant and has a Markov blanket M_i, as defined by equation (5), within X (Yu & Liu, 2004, p. 1208). The feature selection techniques used are theoretically capable of selecting all relevant features but on the issue of redundancy their performance would appear to be less robust.

For LARS the use of Mallows C_p as a barometer for the feature selection provides a convenient test of relevancy. It allows the definition of the best model to be the one that results in the best fit (lowest C_p). In effect LARS is being used to both identify the relevant features and model them iteratively defining the best model. All relevant features from the theoretical best model are then passed to the APT for re-estimation in an OLS framework. This mitigates the chances of feature selection failure[11]. The mechanism by which LARS iterates over the data limits redundancy but with highly correlated explanatory variables a high degree of redundancy is still possible. This is because at each step LARS proceeds equiangular to the combined growing set of variables and not equiangular to each of the individual variables themselves.

The particular implementation of Random Forests used in this study (see Section 4.2) is designed to maximise the chances of providing all relevant features. The number of trees can be increased to a level where the separation between weakly relevant and irrelevant becomes more evident. Nevertheless, Random Forests suffer when it comes to the issue of redundancy. The CART decision tree algorithm underlying the Random Forests performs locally optimal splits (Breiman, 2001). Consequently it does not test the features used to perform splits directly against the other selected features. This inherently introduces the prospect of feature redundancy. In this study the belief is that the GRNN can appropriately weight the variables to compensate for possible redundancy. The GRNN however is a black-box method and so the implicit disadvantage is that the manner in which the features are ultimately used cannot be observed. Attempts are also made (see Section 4.4) to discern and eliminate highly correlated features which ostensibly represent the same economic effects. In spite of this, the unavoidable question remains: of the many other possible nonlinear feature selection techniques available why use Random Forests if they may result in a subset with a large degree of redundancy? The chief motivation for the use of Random Forests is the interpretability of the underlying decision trees which – like the use of C_p in LARS

[11] Feature selection failure is not model failure. While a feature selection technique may not accurately identify all relevant features a model containing some relevant features may still prove adequate. Indeed, the differences between the modelling techniques may mean that the best model as determined by the feature selection method is not necessarily the best model for the estimation technique (OLS and GRNN).

– provides a foundation for the incorporation of data mining techniques that is easily comprehensible.

3.3 MODELLING IDEALS

The emphasis in this study is primarily on model fit. Model interpretability is of secondary import and is upheld through the use of the pre-specification approach ensured by the chosen feature selection techniques. Regardless of whether one subscribes to the EMH, the AMH or one of any number of ancillary market theories, the empirical literature has provided few reasons to believe that the relationship between the economy and stock returns is anything but complex. Thus the relationship between stock returns and proxies for the economic state likely reflects that complexity. Indeed, the notion that "deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex" (Neal, 1995, p. 104), could be said to apply directly to the tendency in economics to seek parsimony as a primary objective.

Parsimony, a principle propelled by Ockham's razor, has intuitive appeal as it relates to the concepts of over-fitting and generalisability. A paradox exists in the application of Ockham's razor in economic modelling. Ockham's razor seems to only be applied in relation to parsimony and not in relation to the simplifying assumptions[12]. For example, if F is the number of features in the true subset of relevant explanatory variables, a strict observance of parsimony suggests that if feature set f, where $f \subset F$, can give a suitable explanation for returns, then feature set f is preferable to even the true subset of relevant variables F. Part of the intuition is that when modelling the features only assumptions about the small parsimonious subset of features n, are now required. The problem with this interpretation is that it suggests that the $F - f$ simplifying assumptions necessary to form a model based on only f do not matter (for complex models it can be the case that $F - f \gg f$). Empiricists pass up this discrepancy on the pretext that the conclusions, if accurate, are more important than the simplifying assumptions. An example of this can be found in the justification for the CAPM assumptions provided by Sharpe (1964). He states that "the proper test of a theory

[12] See the musings of Hurtford (2012) from whence this notion is adapted.

is not the realism of its assumptions but the acceptability of its implications" (Sharpe, 1964). In disputing the CAPM detractors often take issue with the assumptions, highlighting the reality that simplifying assumptions are significant. The viewpoint taken with regards to parsimony in this study is that it is only important in its relation to interpretability.

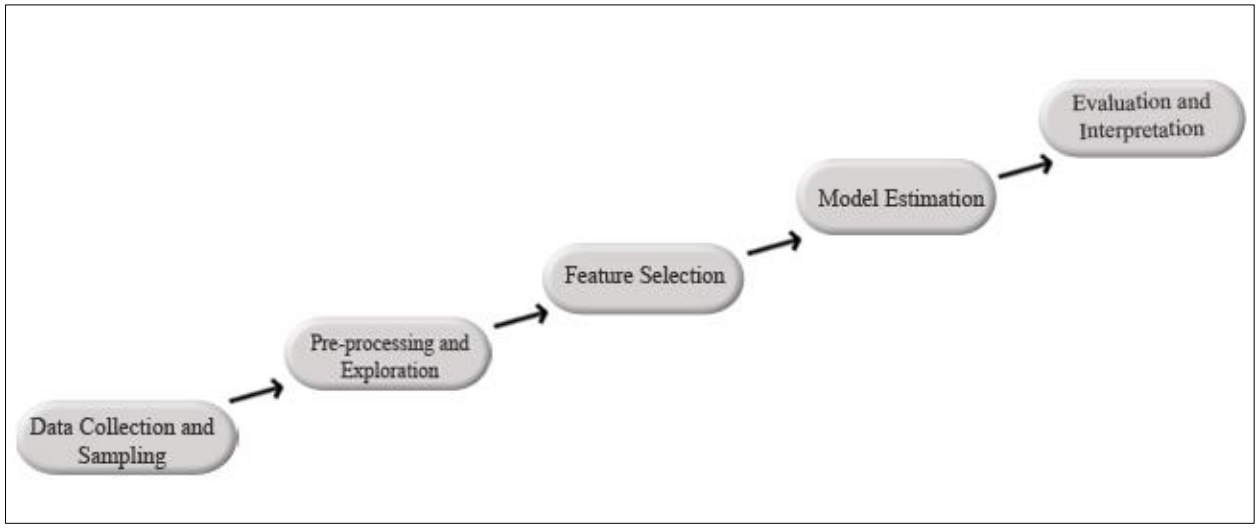
3.4 SUMMARY OF THEORETICAL MOTIVATIONS

This chapter highlights the views taken in this study that are not explicitly motivated elsewhere. Specifically, the chapter emphasises the viewpoint of macroeconomic variables as proxies for the true (unknown) underlying descriptors of the economic state for which investors continuously price in their own estimates. The chapter also discusses the intricacies of feature selection, deliberating over the issues of variable relevance, irrelevancy and redundancy. It is suggested that the Random Forests may be characterised by a degree of redundancy, but they are still useful in demonstrating a basis for the incorporation of data mining techniques in asset pricing. The final section in this chapter motivates the pursuit of the most relevant variables in preference to the best parsimonious set of relevant variables.

4. DATA AND METHODOLOGY

The methodological framework used in this study follows the modified KDD process outlined in Figure 4.1. The following steps are employed 1) data collection and sampling, 2) data pre-processing and exploration, 3) feature selection, 4) model estimation, 5) evaluation and interpretation. Modifications from the KDD process displayed in Figure 4.1 include the combination of the transformation and pre-processing, and the division of the data mining step into features selection and model estimation steps. The changes are made in a bid to clarify the empirical process applied, highlighting the use of the feature selection techniques.

Figure 4.1 Overview of the methodology used in this study.



4.1 DATA COLLECTION AND SAMPLING

The complete sample period covers fifteen years and ten months from January 1998 through September 2014. The sample period is selected to comprise a time after the implementation of the automated trading platform the Johannesburg Equities Trading system (JET) and the electronic clearing system, Shares Transactions Totally Electronic (Strate). The sample also coincides with a period of (relative) political stability, well past the post-apartheid re-entry of South Africa into the international community. This may be important for the interpretation of global economic

affects. The sample covers a number of periods of international economic expansions and downturns – most notably the global financial crisis and the subsequent recovery. Monthly closing values for the indices in the study are obtained from the INET/BFA database. In addition, market capitalisation, earnings and dividend yield data for the ALSI and industry indices is also obtained from INET/BFA. Unfortunately market capitalisation data for the industry indices is only available after 2002 when the JSE adopted the ICB criterion. Monthly data is of principle concern in this study because it corresponds with the frequency in which most of the macroeconomic variables are reported. The indices for which data is collected include the value weighted FTSE/JSE All-Share Index, or ALSI (J203), and the following industry indices: the J500 (Oil and Gas), J510 (Basic Materials), J520 (Industrials), J530 (Consumer Goods), J540 (Healthcare), J550 (Consumer Services), J560 (Telecommunications), J580 (Financials), and the J590 (Technology). The J570 (Utilities) obtains a technical designation as an industry index under the Industry Classification Benchmark (ICB) system adopted by the JSE, however in South Africa too few utilities are listed over the period under study for accurate construction this index. Consequently, the return series under consideration consist of the ten aforementioned indices (excluding the J570). According to the FTSE/JSE Ground Rules the industry indices consist of all FTSE/JSE All-Share constituents belonging to a particular industry (FTSE/JSE, 2015). Thus, it is reasonable to suggest that they offer appropriate insights into the listed firms within the 10 ICB industries (with the obvious exclusion of utilities). The inclusion of industry indices is motivated by a desire to assess whether the macroeconomic factors offer different models for different indices and thus examine whether the models are industry dependent.

In modelling returns as a function of economic state proxies, a large number of possible explanatory variables are available in economic and financial data. Rather than reviewing individual variables a large set of possible features $F = [f_1, f_2, \ldots f_d]$ is constructed (Table 4.1), and the relationship of those factors with returns is interrogated using data mining techniques to discern optimal feature subsets. As the feature selection techniques are robust to high dimensional data further studies may select to review even larger data sets. One of the key ideals in the models proposed in this analysis is interpretability, and hence the variables selected to represent the possible set of features are limited to three broad classes: 1) variables for which an empirical

relationship with returns has been shown (see Section 2.2 and 2.3), 2) leading or coincidental economic variables for which a plausible theory for them representing the (global or local) economic state can be devised, and 3) speculative variables contrived to represent known stock return anomalies. Note that the classes are for classification and clarification purposes only. They are by no means definitive. The complete list of explanatory variables considered can be found in Table 4.1. The data for these variables is obtained from a combination of resources as shown in Table 4.1. The majority of macroeconomic variables are obtained from OECD data sourced from the Federal Reserve Bank of St. Louis (FRED) database, with several exceptions. Retail trade sales, passenger car sales and housing permits data are obtained from Statistics South Africa publications. Monetary aggregates, leading and coincidental indicator indices, and credit expansion data is obtained from the South African Reserve Bank. Data for the RMB/BER Business Confidence Index and the FNB/BER Consumer Confidence Index are attained from the Bureau for Economic Research publications. Closing values for the international indices (the S&P 500, CAC 40, DAX 30, Nikkei 225, NASDAQ Composite, Hang Seng, VIX, Mexican Bolsa IPC Index, FTSE 100 and Bovespa Brasil) are obtained primarily from Bloomberg and supplemented with data from Yahoo! Finance.

Most of the variables in Table 4.1 are common, but some warrant additional clarification. The leading and coincidental business cycle indicators obtained from the Reserve Bank are composite indicators integrating various economic time-series into a single indicator. The current incarnation of the leading indicator contains a weighted sum of a diverse set of indicators which include factors such as the average hours worked per factory worker and the growth rate in job advertisements in the Sunday Times. Although the leading indicator contains a smoothed six month growth rate of the price of all classes of shares, it is still believed to provide a novel innovation as this singular component only represents what stock prices should be based on what they have previously been. The composite coincidental indicator contains variables such as the production utilisation and value of retail, wholesale and new vehicle sales. Like the Bureau of Economic Research the OECD also tracks business confidence and consumer confidence in OECD countries based on survey evidence. The resultant time-series are recorded as BCIO and CCIO.

Table 4.1 Variable set from which relevant features are to be selected

	Variable	Abbrev.	Class	Source	Freq.
Macroeconomic	Gross Domestic Production	GDP	2	F	Q
	Consumption Expenditure	CE	1	F	Q
	Formal Employment Count	E	2	F	Q
	Consumer Price Index	CPI	1	F	M
	Producer Price Index	PPI	1	F	M
	M2	M2	1	R	M
	M1	M1	1	R	M
	Manufacturing Production	MANP	1	F	M
	Mining Production	MINP	2	F	M
	Construction Production	CONP	2	F	M
	Retail Trade Sales	RTS	2	S	M
	Passenger Car Sales	PCS	3	S	M
	Domestic Credit	DC	1	R	M
	Consumer credit	CC	2	R	M
	Housing Permits	HP	2	S	M
	Composite Leading Indicator	LI	3	R	M
	Composite Coincidental Indicator	CI	3	R	M
	Balance of Trade	BoT	1	F	M
	RMB/BER Business Confidence Index	BCIS	2	E	Q
	FNB/BER Consumer Confidence Index	CCIS	2	E	Q
	OECD Business Confidence Index	BCIO	2	F	M
	OECD Consumer Confidence Index	CCIO	2	F	M
Commodities	Oil Price	Oil	1	B	M
	Gold Price	Gold	1	B	M
Bonds	91-day Treasury Bills	TB	1	F	M
	Long term (10YR) Government bonds	LTB	1	F	M
Fundamentals	Earnings Yield	EY*	1	I	M
	Dividend Yield	DY*	1	I	M
	Market Capitalisation	S*	1	I	M
International Indices	S&P 500	SNP	2	B	M
	NASDAQ Composite	NAS	2	B	M
	VIX	VIX	3	B	M
	Nikkei 225	NIK	2	B	M
	Hang Seng	HANG	2	B	M
	FTSE 100	FTSE	2	B	M
	DAX 30	DAX	2	B	M
	CAC 40	CAC	2	B	M
	Mexican Bolsa IPC Index	MXC	2	B	M
	Bovespa Brasil	BRA	2	B	M

Note: Class 1 refers to variables which have been shown to have an empirical relationship with returns, Class 2 refers to leading and coincidental macroeconomic variables for which a plausible theory either exists or can be devised. Class 3 encompasses speculative variables motivated by theory. The data is of the following frequencies: Q – Quarterly data, M – Monthly data. Sources for the data include: I – INET/BFA, S – Statistics South Africa, R – The South African Reserve Bank, F - The Federal Reserve Bank of St. Louis (FRED) database, E – Bureau for Economic Research, and B – Bloomberg and/or Yahoo Finance.

**Variables are index dependent, the naming convention used is variable"index" e.g. for J500 earnings yield: EYJ500.*

Financial markets have changed considerably since van Rensburg (1996). It is debatable whether the Dow Jones Industrial Index still offers a suitable impression of U.S. equity markets. The slow adjusting process and strict inclusion criterion of the DOW has meant that it has fallen out of step with the growth of many industries such as the technological sector. A prominent example is the exclusion of Apple Inc. from the DOW which in the years 2009 to 2014 was amongst the top two largest companies in the world by market capitalisation[13]. The S&P 500 is investigated in this study as a more robust proxy for U.S market effects. The motivation for the use of international stock indices is that these may proxy for international economic effects. Two emerging market indices are present amongst the international indices examined. They are perhaps reasonable proxies for the global appetite for investments in emerging economies, which is a function of the global economy.

The sample period is subdivided into two overlapping sub-periods, sub-period 1 which covers the period from February 1998 to September 2008, and sub-period 2 that covers the period from February 2004 to September 2014 (Table 4.2). Market capitalisation will only be used as an innovation in the second sub-period as data for industry indices is only available after 2002. The periods under investigation are further separated into training and test periods of following proportions: 85 to 90% training data and 10 to 15% test data. The training sets for the full period and sub-periods consist of 182 and 110 months, respectively. In each case the test sets consists of 18 months of data. In accordance with Looney (1996) for the full period 65% of the sample is used for model estimation, and 25% for cross-validation – all within the training sets. The test sets are thus 10% of the full sample period. For the sub-periods 60% of the sample is used for model estimation, 25% for cross-validation and 15% for models testing. Feature selection takes place on the training data in each period.

[13] At the time of writing Apple Inc. was finally poised to join the DOW despite being amongst the world's largest companies, by market capitalisation, for several years.

Table 4.2 The sample periods, sub-periods and the number of associated data points per variable.

	Period	No. of data points
Full sample period (f)	February 1998 - September 2014	182
Sub-period 1 (s1)	February 1998 - September 2008	110
Sub-period 2 (s2)	February 2004 - September 2014	110

4.2 PRE-PROCESSING

All data manipulation, statistical and computational analysis in this study is performed using R statistical software environment[14]. Monthly logarithmic returns for the ALSI, J500, J510, J520, J530, J540, J550, J560, J580, and the J590 are computed from the indices closing values. As the main results of this study are drawn from monthly data, an interpolation method is necessary for quarterly data in order to build a series contemporaneous with the majority of the macroeconomic series in this study. Cubic spline interpolation is used for this purpose [15]. The cubic spline is a piecewise polynomial which is able to estimate a smoothed series that passes through the known data points, often called 'knots' (Eilers & Marx, 2010; Reinsch, 1967). The key motivation for using spline interpolation is the avoidance of Runge's phenomenon where the interpolant leads to oscillations between data points[16]. A cubic spline works by drawing a unique cubic polynomial across each space (gap between the known data points). For example, with *r* quarterly values the number of spaces between them is *r-1*. A cubic polynomial is drawn across each space in a piecewise fashion (Forsythe, Moler, & Malcolm, 1977). Fitting the spline requires a solution to a problem with $(r - 1) \times 4$ unknowns. In assuming that the spline passes through all knots $(r-1) \times 2$ conditions can be met. A further $(r-2) \times 2$ conditions can be met by assuming that the series is smooth thus removing unnecessary oscillations (Forsythe *et al.*, 1977). Forsythe *et al.* (1977) show that the remaining two conditions can be solved by using the first four and last four points to define a cubic from which the third derivatives are used as a boundary condition. Having resolved the relevant coefficients for all the polynomials, values at the required intervals can be obtained by

[14] For discussion and replication purposes the customised technical functions can be found in Appendix 1.
[15] The R code used to perform the spline is displayed in Appendix 1A for interested readers.
[16] See Epperson (1987) for details on Runge's theory.

substitution (Forsythe *et al.*, 1977). Interpolation does not 'generate' new innovations. The interpolant is not estimating the unknown data points. It simply builds a series as a smoothed path through the known values. This minimises interpolation error but it also results in a series with a lower variance, which may be significant in attempting to model variables with a high variance such as equity returns.

After obtaining monthly values for GDP, CE, E, U, BCIS and CCIS through spline interpolation, these variables are transformed alongside the monthly data. In general, variables that are measured in percentages are differenced and variables which are unit counts, indices or prices are differenced after finding their natural logarithms.

The unemployment rate, U, is differenced so as to arrive upon the growth rate of unemployment. The FNB/BER Consumer Confidence Index, the OECD business confidence index and the OECD consumer confidence index for South Africa are obtained in percentage terms and so they are simply differenced in this study. The rates of 91-day treasury bills and long term government bonds are simply divided by twelve to give a monthly rate. The additional variables TERM and RHO are constructed as specified in Chen *et al.* (1986). The fundamental values for each index, earning yield (EY), dividend yield (DY) and market capitalisation (S) are logged, as is common practice for such variables in regression studies. New variables REY, RDY, RS and RCR are based on the notion that the best measure for a value or size effect may not be an absolute measure but rather a measure of size or value relative to some benchmark. For example, the obvious benchmark for the earnings yield of an industry index (such as EYJ500) is the EYALSI. These variables are constructed by dividing the industry EY, DY and S by that of the ALSI to obtain relative measures. Further studies may consider using historical averages of the EY, DY and S to obtain relative measures. RCR is a variable constructed to measure relative cumulative return. This represents an attempt to capture a possible loser-winner effect as outlined in DeBondt and Thaler's (1985) overreaction hypothesis. RCR is based on the twelve month cumulative returns (CR's) of the indices. Each CR is calculated by adding returns over the twelve months prior to the time in question. The CR's for each index are then divided by the ALSI cumulative returns over the same

period, to give a measure of relative cumulative return for each index. The idea is to examine whether there is mean reversion in the returns of the industry indices.

The DAX, CAC and the SNP are found to be highly positively correlated with one another and with the FTSE (>0.82). As a result they are removed as possible features based on the understanding that the effects they proxy for are likely captured in the FTSE. Before feature selection is performed each variable is standardised over the training and test period by subtracting its training period mean from each value then dividing by its training period standard deviation, essentially computing z-scores where $z = (x_t - \mu) / \sigma$. This does not affect the relative relationships of the variables but allows any residual scaling effects to be removed. Standardising in this manner also has the convenient consequence of allowing each standardised variable to have a mean of zero making coefficient comparisons more informative.

4.3 FEATURE SELECTION AND MODELLING

This study computes time-series multifactor models of the ALSI and industry indices. Feature selection is necessary to ensure that only relevant subsets of variables are used, aiding in the interpretability of the models. The training sets are used for variable selection by the feature selection techniques. Feature selection is repeated with one of the ALSI, J500, J510, J520, J530, J540, J550, J560, J580, or J590 as dependent variables. The test sets are reserved exclusively for the predictions, which will be used to assess the robustness of the models. The set of possible features examined for the ALSI excludes all relative variables. For the industry indices all the relative measures and the ALSI are added as possible features. Market capitalisation is only added as a possible feature for the second sub-period for which data is available. As the models are meant to be explanatory, contemporaneous variable series are used. This implies an inherent lag for some variables for which the data for a given month is only available the next month – altogether suggesting that the models are composed of data which is mostly available to investors.

4.3.1 LINEAR VARIABLE SELECTION

The LARS algorithm is used to find linear variables for a time-series multifactor model as shown in equation (2). Described in this way the APT shows a clear description of risk-return relationships. As the literature illustrates LARS is optimised for linear variable selection. The selection criterion for variables is the Lasso which minimises the sum of squares of the residuals subject to an upper bound on the absolute value of the coefficients b_j, as shown in equation (6).

$$\sum_{i=1}^{d} \left(ri - \sum_{j=1}^{d} f_{ji} b_j \right)^2 + \theta \sum_{j=1}^{d} |b_j| \qquad (6)$$

This is solved using the LARS steps. As outlined by Efron *et al.* (2004) the procedures of the LARS algorithm can be described as follows: It initiates with all coefficients equal to zero and finds the predictor variable x_{j1} with the highest correlation to the response variable. Rather than taking a complete step to estimating the model with this initial predictor, the LARS algorithm takes the largest possible step in the direction of the predictor until another predictor x_{j2} has just as much correlation with the produced residual. When this occurs LARS proceeds in a direction equiangular between the two predictors until another predictor x_{j3} has as much correlation with the residual as the other two predictors. The LARS algorithm continues in this vain proceeding equiangular between the predictors and building a "most correlated set" of input variables.

The LARS algorithm identifies the optimal subset of variables sequentially. To complete feature selection, Mallows C_p statistics are used to measure prediction error. Mallows (1973) defines the C_p measure for a model with p features selected (intercept included) from the set of possible features F as follows:

$$C_p = \frac{SSE_p}{MSE} - (n - 2p), \qquad (7)$$

SSE_p is the sum of squared errors of prediction for the model with p independent variables and MSE is the mean squared error for the full model with F features. Mallows C_p is a measure of goodness of fit, penalising redundant variables. The model with the most appropriate variables and thus the feature subset is the one that minimises C_p.

The features selected by LARS give the variables to be estimated by the following APT-type model:

$$y_t = \alpha + \sum_{j=1}^{n} f_j b_j + \varepsilon_t \qquad (8)$$

where *n* is the total number of coefficients, b_j is the coefficient of *n* LARS selected factors f_j. This model is estimated over the training period and validated over the cross-validation period. The robustness of the linear models are then evaluated on the test periods.

4.3.2 NONLINEAR VARIABLES SELECTION

General Regression Neural Networks with a sigmoid activation function are used to model the possible nonlinear relationships between the factors and returns. The word nonlinear is used loosely here as a GRNN is not strictly nonlinear, it converges to model whatever relationships exist whether they are linear or nonlinear. Neural networks are not markedly encumbered by the demands of variable selection. They can efficiently take all of the input variables as arguments, only learning the relevant pattern and ignoring irrelevant variables. However, ANN's are black box models, meaning that even after they have been successfully trained they offer no insight into which variables proved to have the highest predictive or explanatory power. In order to meet the relative constraints of parsimony and provide an interpretable model it is best to determine a subset of explanatory variables *a priori*. For this purpose Random Forests are employed.

Random Forests can be thought of as an ensemble of decision trees, but more technically they are an ensemble of tree based weak classifiers (Kursa, Jankowski, & Rudnicki, 2010). The idea is that by combining these in the final classification more robust classification can be performed. The procedure as described by Breiman (2001) begins with the growing of individual trees. Trees are grown using bootstrapped samples of the training data based on the concept of bootstrap aggregating or 'bagging' (Breiman, 1996, 2001). After each split a different subset of features is randomly selected for evaluation, and splits are ultimately performed on the feature that results in

the largest nodes - the attribute for which the most data can be split on (Breiman, 2001). Completed trees are used to obtain *out of bag* error estimations and to measure the importance of features for a single tree. This is done through a system of feature ranking through tallies and randomised repetitions (see Breiman, 2001). The overall importance of features is attained by averaging the importance measures for single trees. Despite the name the importance score generated by Random Forests is not a definitive measure of feature importance. Kursa and Rudnicki (2010) suggest that the problem has to do with Breiman's (2001) assumption that the low correlation between single trees meant that importance was normally distributed. The suggestion is that normality is an inaccurate assumption and so a reference point is necessary to determine importance from irrelevance.

The Boruta algorithm is a wrapper for Random Forests that compares the importance of variables selected by a Random Forest with that of randomised ones, providing a better test for variable importance (Kursa *et al.*, 2010; Kursa & Rudnicki, 2010). The procedure for the algorithm is described in detail by Kursa and Rudnicki (2010) and can be summarised as follows: Boruta first extends the system by creating randomised copies of the original features (called shadow features) by shuffling the values of the original features. Importance scores are obtained by running Random Forests on the extended set. Any feature with an importance score higher than the maximum score among the shadow features (MSSF) is deemed important. Features with lower scores than MSSF are therefore deemed unimportant and are permanently removed from the system. All shadow features are then removed. The process is repeated until either all the remaining features are deemed important or a predefined end point is reached. The Boruta algorithm is implemented using R statistical software[17].

The Random Forests give a subset of variables with relationships to returns that may include some nonlinearity. The following model describes the incorporation of variables selected by the Random Forests into an ANN:

[17] The R code used to implement the Boruta algorithm is displayed in Appendix 1B.

$$y_t = f[(f), a, b\,] = \sum_{j=1}^{n} a_j logsig\left(\sum_{i=1}^{m} b_i f_i + b_{0j}\right) \quad (9)$$

where f is a vector of variables selected by the Random Forests, n is the number of units in the middle layer, m is the number of features selected, a is a vector of coefficients from the middle to output layer neurons in the neural network and b is a matrix of coefficients from input to middle layer neurons.

The GRNN's used in this study are built using the R software environment[18]. As specified 65% of the sub-periods are allocated to training of the GRNN, and 25% to cross validation. The GRNN is much simpler than many other neural networks requiring only one adjustable parameter – the smoothing factor σ, where $0 < \sigma \leq 1$. The most appropriate value of σ can be determined through cross-validation meaning that the GRNN can be implemented autonomously reducing the chances of subjective biases. This study uses the cross validation set to determine the most appropriate value for the smoothing parameter, σ. In the GRNN each training input-output set (row of data in a data frame), represents its own cluster. Specht (1990a) defines the GRNN training process as follows: On the introduction of new input pattern *z* to the GRNN, the training pattern *y* of cluster *i* assigns *z* a value λ_i as defined in equation (10):

$$\lambda_i = exp^{\left(-\frac{D_i^2}{2\sigma^2}\right)} \quad (10)$$

, with Euclidean distance $D = D(x, u_i)$, where $D^2 = (x - u_i)^T(x - u_i)$, and $0 < \sigma \leq 1$.

$$\hat{y} = \frac{\sum_{i=1}^{n} \lambda_i w_i}{\sum_{i=1}^{n} \lambda_i} \quad (11)$$

The target value or estimation $\hat{y}$ is computed by equation (11). The GRNN is trained and cross-validated on the training data and then initially used to estimate the target values on which it is trained. This allows some training period comparisons to the linear models. The GRNN models are then evaluated over the test periods to assess their robustness.

[18] The code for the GRNN function used can be found in Appendix 1C.

4.4 EVALUATION OF HYPOTHESE

The simplest hypothesis to test is *Hypothesis 3* as it involves an analysis of the variables selected by LARS and Random Forests to model the ALSI and industry indices. Should the subset of variables selected for modelling a particular index differ in the full period and sub-periods then this constitutes a rejection of *Hypothesis 3.1*. Should the subset of variables selected for modelling the industries indices differ then this constitutes a rejection of *Hypothesis 3.2.*

Hypothesis 1 posits that the feature selection techniques (LARS and Random Forests) can select variables which result in appropriate models. Describing a model as appropriate does not imply that the model is necessarily parsimonious or even that a better model cannot be defined. It is simply a statement that suggests that the model appears to give a suitable explanation for the series being modelled. To test *Hypothesis 1* the variables selected are modelled in the training period and the fit is investigated. The fit of the linear models is primarily examined by assessing the significance of the coefficients and the adjusted-R^2. Although adjusted-R^2 is a better overall measure, the correlation of the fitted and actual values is also reported. The correlations can be compared to those of the GRNN models for which no equivalent to the adjusted-R^2 exists. The outcome of a Sign test on the fitted and actual values is evaluated to test the null hypothesis that the median difference between the fitted and actual values is zero. Finally, the distributional properties of the fitted and actual values are assessed using a Kolmogorov-Smirnov test, which is a nonparametric test with the null hypothesis that the fitted and actual values have the same distribution (Smirnov, 1948). When correctly modelled, the returns of the ALSI and the industry indices should be a function of a deterministic component and a stochastic error. The OLS regression assumes that the residuals represent stochastic error. Residual plots are assessed to ensure that there is no heteroscedasticity, nonlinearity or other biases in the residuals. Random errors often follow a normal distribution and hence Shapiro and Wilk's (1965) test for normality is performed on the residuals. It must be noted that non-normality in the residuals does not necessary disqualify the usefulness of a regression model. As long as the residuals are independent useful inferences can be drawn. Ljung and Box (1978) develop a joint test for serial correlation up to a given lag order. Ljung-Box tests are performed on the residuals to test for serial correlation up

to the fifth lag. For neural networks testing the training period fit is substantially more difficult. The trained GRNN is used to produce fitted values by applying it to the training data. The correlation of the fitted values with the actual values is computed as well as a Sign test with the null hypothesis that the median difference between the actual and fitted values is zero. As in the linear models a Kolmogorov-Smirnov test is performed to compare the distributions of the fitted and actual values. The in sample residuals of the GRNN models are also assessed using residual plots, Ljung-Box tests and Shapiro-Wilk tests.

Hypothesis 2 can only be tested to a limited extent by evaluating model fit. To draw stronger comparisons between the linear models and the neural networks models, the test period model performance is investigated. Over the testing periods (10-15% of the full sample or sub-periods as indicated in Table 4.2) predictions are computed for equations (8) and (9). These are predictions in this sense that they are conducted over periods beyond the initial model estimation (training). However, they are not extrapolative (conditional on historical data only). Instead, they are conditional on the 'new' explanatory variable data in the test periods. The fact that this data is known means that the predictions are akin to estimates. Nevertheless, prediction error is still the difference between actual future value of the return of the index being predicted, and the estimated value of that return. To assess the accuracy of predictions, as in Qi (1999), Callen *et al.* (1996) and Cao *et al.* (2011), the prediction errors are interrogated using the following metrics:

Mean Absolute Error (MAE) = $\frac{1}{N-n}\sum_{t=n+1}^{N}|y_t - \hat{y}_t|$ `

Mean Absolute Percentage Error (MAPE) = $\frac{1}{N-n}\sum_{t=n+1}^{N}\left|\frac{y_t-\hat{y}_t}{y_t}\right|$

Root Mean Squared Error (RMSE) = $\sqrt{\frac{1}{N-n}\sum_{t=n+1}^{N}\left(\frac{y_t-\hat{y}_t}{y_t}\right)^2}$

Pearson correlation coefficient actual and predicted/fitted (CORR) = $\frac{\sum(y_t-\bar{y})(\hat{y}_t-\bar{\hat{y}})}{\sqrt{\sum(y_t-\bar{y})^2}\times\sqrt{\sum(\hat{y}_t-\bar{\hat{y}})^2}}$

Ratio of correctly predicted Signs (SIGN) = $\frac{1}{T}\sum z_i$ where z is the count of correctly predicted signs.

In addition, a Sign test, with the null hypothesis that the median difference between the actual and predicted values is zero, is performed.

Hypothesis 2.1 and *2.2* attempt to clarify whether any difference in the model performance of neural networks and linear models is driven by the underlying variable selection techniques or the modelling techniques. Since neural networks can estimate linear models but OLS-regressions do not account for possible nonlinear relationships, then a simple way to investigate *Hypothesis 2.1* and *2.2* is to use the GRNN to model the linear variable sets selected by LARS. This is done over the full sample period for all ALSI and industry indices. The results are compared to the standard linear models and the standard neural network models (i.e. with variables selected by Random Forests) using the aforementioned prediction error metrics.

Like Cao *et al.* (2011) and Qi (1999) this study also computes Diebold-Mariano tests to examine whether the linear model or the neural network provided the better test period performance. The simple test conceived by Diebold and Mariano (1995) considers the loss associated with prediction *i* to be a function *g(•)* of the prediction error. The loss functions for a linear (*L*) and neural network (*N*) model are expressed as $g(e_L)$ and $g(e_N)$, respectively. The loss differential *d* between the two models is defined by equation (12).

$d_t = g(e_{Nt}) - g(e_{Lt})$ (12)

The mean loss differential is defined by equation (13).

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(e_{Nt}) - g(e_{Lt})] \qquad (13)$$

For the linear and neural network prediction comparisons the null hypothesis of the Diebold-Mariano tests asserts that the test period linear and GRNN loss functions are indistinguishable from each other, H_0: $E(d_t) = 0$. The alternative is that the test period linear and GRNN loss functions differ, H_A: $E(d_t) \neq 0$. If the null is rejected the model with the better prediction error metrics can be considered a better predictor. The linear and GRNN predictions are also compared to those of a naïve model (a random walk). To test this, Diebold-Mariano tests are employed with the null hypothesis that the linear and GRNN predictions are, in turn, indistinguishable from those of a naïve model, H_0: $E(d_t) = 0$. The alternative hypothesis in this case is that the loss function of the naïve model is worse than that of the linear or GRNN model being tested.

5. RESULTS

5.1 DESCRIPTIVE STATISTICS

Summary statistics for the indices examined are displayed in Table 5.1. The tests in Table 5.1 are performed before the standardisation of variables for use in modelling and they represent statistics for the log returns of the indices. Histograms and plots of the series over the full sample period can be found in Appendix 2A. The mean monthly return for the ALSI over the full period under study is 0.4% with a considerable standard deviation of 2.6% highlighting the high degree of volatility associated with equity returns. Indeed, the mean monthly returns for all the industry indices are lower than their respective standard deviation measures. The industry index with the highest mean return is the J530 (Consumer Goods) with 0.64%, it is followed by the J560 (Telecommunications) at 0.57% and the J540 (Healthcare) at an average of 0.57% per month. They are joined by the J500 (Oil and Gas), J520 (Industrials), and the J550 (Consumer Services) as industry indices that showed higher average monthly returns than the ALSI. At 0.10%, the J590 (Technology) has the lowest mean industry index return over the period under study. Along with the J510 (Basic Materials) and the J580 (Financials) it has mean returns lower than the ALSI. Incidentally, in addition to having the lowest mean return the J590 also has the highest standard deviation at nearly 5%. The industry standard deviations range from 2.75% for the J540 to the 4.94% observed for the J590, but without exception they are all higher than that of the ALSI, pointing to the benefits of industry diversification. Also evident in the indices returns is a persistent negative skew which manifests in the general tendency for one month minimum returns to be larger than one month maximums. Implicit in this is the potential for greater losses in any one month than gains. The summary statistics indicate some truth to this conjecture but a precise measure of this is dependent on the distributional properties of the returns. The J510 and J520 are the only exceptions to this tendency boasting higher one month maximums than minimums. The J590 shows the largest one month loss of 2.35% while the J510 shows the largest one month gain of 15.17%. Consistent with the common observation of leptokurtosis in equity returns, the series here largely demonstrate a high degree of excess kurtosis. Nowhere is the skewness and kurtosis more evident than in the J580 (see Appendix 2A, Figure 2A9). The only clear exception is the J500 which only has a marginally positive excess kurtosis.

Table 5.1 Summary statistics for the ALSI and Industry index log return series being modelled.

Variable	Mean	Std.	Min./ Max.	Skewness	Kurtosis	JB test	Shapiro test	LB test (lag = 5)	LB test (lag = 12)	ADF test w/no trend	KPSS test
ALSI	0.0042	0.0261	-0.1515	-1.227	5.570	308.35	0.935	7.633	11.766	-9.437	0.057
			0.0570			(<0.01)	(<0.01)	(0.178)	(0.465)	(<0.01)	(>0.1)
J500	0.0047	0.0348	-0.0916	-0.091	0.469	2.336^	0.991^	11.488^	17.325	-9.440	0.064
			0.1230			(0.302)	(0.311)	(0.043)	(0.138)	(<0.01)	(>0.1)
J510	0.0036	0.0344	-0.1111	-0.118	2.049	36.16	0.972	2.953	11.835	-9.072	0.111
			0.1517			(<0.01)	(<0.01)	(0.707)	(0.459)	(<0.01)	(>0.1)
J520	0.0044	0.0278	-0.0540	-1.068	3.717	153.42	0.947	5.224	10.655	-9.115	0.104
			0.0609			(<0.01)	(<0.01)	(0.389)	(0.559)	(<0.01)	(>0.1)
J530	0.0064	0.0301	-0.1078	-0.626	1.945	45.12	0.962	11.749^	22.877^	-9.050	0.087
			0.0874			(<0.01)	(<0.01)	(0.038)	(0.029)	(<0.01)	(>0.1)
J540	0.0053	0.0275	-0.1234	-0.786	2.195	61.18	0.965	2.771	10.174	-9.434	0.320
			0.0749			(<0.01)	(<0.01)	(0.735)	(0.601)	(<0.01)	(>0.1)
J550	0.0046	0.0300	-0.1766	-1.389	5.858	349.51^	0.926^	10.070	17.287	-8.548	0.237
			0.0634			(0.463)	(0.432)	(0.073)	(0.139)	(<0.01)	(>0.1)
J560	0.0057	0.0408	-0.1718	-0.270	2.431	52.417^	0.966^	7.782	13.233	-9.596	0.061
			0.1377			(0.604)	(0.467)	(0.169)	(0.352)	(<0.01)	(>0.1)
J580	0.0032	0.0292	-0.2223	-2.221	17.259	2629.01	0.855	2.802	13.217	-10.545	0.069
			0.0940			(<0.01)	(<0.01)	(0.731)	(0.354)	(<0.01)	(>0.1)
J590	0.0010	0.0494	-0.2348	-1.169	3.665	157.57	0.927	4.001	16.045	-9.271	0.287
			0.1303			(<0.01)	(<0.01)	(0.549)	(0.189)	(<0.01)	(>0.1)

Note: p-values are in brackets. ^Atypical results in relation to the other series.

The results of Jarque-Bera and Shapiro-Wilk normality tests are also reported in Table 5.1. Normality, while not as important an assumption as stationary, imbues a series with convenient properties. Most notably, a normally distributed series only requires information about the mean and standard deviation for probabilities and hence expectations to be derived. The ALSI is found not to follow a normal distribution, with both the Jarque-Bera and Shapiro-Wilk tests rejecting their null hypothesis of normality (or more strictly for the JB test that skewness and kurtosis is zero). Of the industry indices the J500, J550 and J560 are found to be normally distributed over the sample period with both tests failing to reject the null hypothesis. Ljung-Box tests for serial correlation are conducted up to lag orders five and twelve. The Ljung-Box test is a portmanteau test with a null hypothesis of no serial dependence in the data (independently distributed). The test statistic reported in Table 5.1 is in fact the χ^2 statistic. Using the Ljung-Box tests the ALSI is found to exhibit no serial correlation up to either lag order. The results for serial correlation tests within the industry indices are more nuanced. The J500 displays serial correlation up to lag order five but not when twelve lags are considered. Ljung-Box tests on the J530 indicate that the consumer goods

index has significant serial correlation up to both five and twelve lags. The rest of the industry indices do not exhibit serial correlation in the tests performed.

The stationarity of the index returns is investigated as non-stationary series are prone to the spurious regression problem most adeptly described by Granger and Newbold (1974) in relation to time series. Augmented Dickey-Fuller (ADF) tests are conducted to examine stationarity. These test have a null hypothesis that the series has a unit root. In this study the distribution of the test statistics is obtained from Hamilton (1994). Table 5.1 clearly demonstrates that the ALSI and all the industry returns are found to be stationary through the rejection of the ADF null hypothesis. Results for ADF tests with a trend or drift added (not reported) corroborate these findings. Kwiatkowski, Phillips, Schmidt, and Shin (1992) design tests with a null hypothesis of stationarity. In all cases the test fails to reject this null.

The summary statistics for the explanatory variables are presented in Tables 5.2 and 5.3. As with the indices statistics (Table 5.1) these statistics are computed on the transformed series before standardisation. The first four variables (GDP, CE, E and U), BCIS and CCIS are products of the cubic spline interpolations conducted on the quarterly data. As a result, the degree to which the summary statistics characterise the true monthly representation is disputable. For one, the technique undeniably results in a decrease in the overall standard deviations by smoothing the series. Nevertheless, the cubic splines do not remove information and they are common in econometrics. At 0.248% the mean monthly GDP growth rate appears plausible suggesting an annualised rate of around 3%. The average monthly change in the unemployment rate is 0.02% emphasising a stubbornly stagnant labour market and implying an annualised increase in the unemployment rate of 0.24%. Also troubling is the positively skew associated with changes in unemployment. Table 5.2 reports an average monthly inflation rate of 0.44% corresponding to an annual rate of 5.28%. Amongst the production metrics mining production (MINP) is anomalous in showing an average monthly decline. A likely demonstration of the decline in the mining sector, gold mining in particular, that has occurred over the sample period.

Table 5.2 Summary statistics for the macroeconomic explanatory variables.

Variable	Mean	Std.	Min./ Max.	Skewness	Kurtosis	JB test	Shapiro test	LB test (lag = 5)	LB test (lag = 12)	ADF test w/no trend	KPSS test
GDP	0.0025	0.0020	-0.0059	-0.9570		11.34	0.949	447.16	554.40	-3.502	0.356
			0.0065			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(0.096)
CE	0.0030	0.0026	-0.0048	-0.2780	0.3305	3.62	0.981	570.66	798.42	-3.164	0.271
			0.0089			(0.164)	(0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
E	0.00020	0.0026	-0.0088	-0.6691	0.6722	18.78	0.972	238.74	254.12	-9.059	0.524
			0.0057			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
U	0.00024	0.0046	-0.0134	0.1421	1.2626	14.47	0.978	198.74	214.96	-15.110	0.080
			0.0168			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CPI	0.0044	0.0048	-0.0114	0.5159	1.5321	28.90	0.969	67.22	81.79	-3.835	0.121
			0.0235			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
PPI	0.0050	0.0081	-0.0149	3.8599	33.3893	9706.16	0.742	48.74	54.67	-5.429	0.058
			0.0782			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
M2	0.0096	0.0149	-0.0327	-0.0857	-0.4271	1.537^	0.993^	4.5749^	21.02^	-6.868	0.494
			0.0413			(0.463)	(0.432)	(0.470)	(0.05)	(<0.01)	(0.043)
M1	0.0100	0.0255	-0.0588	0.1129	0.2362	1.007^	0.993^	12.13	27.08	-9.555	0.236
			0.0802			(0.604)	(0.467)	(0.033)	(0.008)	(<0.01)	(>0.1)
MANP	0.0006	0.0271	-0.0735	0.1900	1.2725	15.20	0.979	59.23	90.47	-15.692	0.053
			0.1016			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
MINP	-0.0003	0.0464	-0.1658	-0.0133	2.6554	59.55	0.949	60.83	67.19	-14.489	0.031
			0.1797			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CONP	0.0014	0.1884	-0.5898	3.0266	2.3156	48.44	0.971	49.44	54.09	-19.009	0.060
			0.8721			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
RTS	0.0032	0.0176	-0.0797	-0.4077	2.9867	80.60	0.963	31.10	50.91	-12.141	0.181
			0.0676			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
PCS	0.0029	0.0586	-0.1761	0.2307	1.6182	24.19	0.980	14.12	21.87	-12.013	0.151
			0.2455			(<0.01)	(<0.01)	(0.015)	(0.039)	(<0.01)	(>0.1)
DC	0.0096	0.0143	-0.0416	0.2399	1.4260	19.41	0.982	18.92	36.97	-6.379	0.327
			0.0614			(<0.01)	(0.013)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CC	0.0092	0.0090	-0.0154	1.7335	9.8907	911.00	0.882	87.11	207.76	-4.171	0.6604^
			0.0689			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(0.017)
HP	0.0077	0.1445	-0.5847	-0.0035	2.6759	60.46	0.949	78.47	112.87	-18.031	0.111
			0.4896			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
LI	0.0017	0.0114	-0.0384	-0.0186	0.8369	6.22	0.985	142.55	161.11	-5.205	0.070
			0.0363			(0.045)	(0.035)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CI	0.0020	0.0079	-0.0234	-0.6559	1.0709	24.14	0.963	178.41	223.02	-4.977	0.421^
			0.0277			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(0.068)
BoT	0.0000	0.1231	-0.4492	0.6684	5.6025	276.55	0.924	102.69	108.96	-15.771	0.035
			0.6758			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
BCIS	0.0008	0.0651	-0.2194	0.1789	0.9461	8.93	0.983	190.36	217.75	-10.774	0.105
			0.1985			(0.012)	(0.018)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CCIS	4.6242	9.8418	-17.0000	0.2113	-0.9961	9.15	0.959	593.99	1014.58	-4.775	1.046^
			23.2049			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(0.01)
BCIO	-0.0046	0.1972	-0.5000	-0.3794	-0.0020	4.73^	0.958	249.00	257.20	-7.050	0.086
			0.5000			(0.094)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)
CCIO	-0.0113	0.2634	-0.8000	0.2780	1.5336	22.73	0.959	172.71	176.10	-12.435	0.091
			1.0000			(<0.01)	(<0.01)	(<0.01)	(<0.01)	(<0.01)	(>0.1)

Note: p-values for the tests are reported in brackets. ^Atypical results in relation to the other series.

Table 5.3 Summary statistics for the Gold, Oil and international stock indices.

Variable	Mean	Std.	Min./ Max.	Skewness	Kurtosis	JB test	Shapiro test	LB test (lag = 5)	LB test (lag = 12)	ADF test w/no trend	KPSS test
Oil	0.0079	0.0929	-0.3110	-0.7451	0.8452	24.56	0.964	8.291^	21.49	-8.557	0.058
			0.2007			(<0.01)	(<0.01)	(0.141)	(0.044)	(<0.01)	(>0.1)
Gold	0.0077	0.0393	-0.1239	0.2686	1.4047	19.37	0.981	4.59^	10.97^	-9.869	0.577
			0.1639			(<0.01)	(<0.01)	(0.468)	(0.531)	(<0.01)	(0.025)
NAS	0.0044	0.0754	-0.2601	-0.6201	1.1211	23.58	0.973	2.366^	7.531^	-9.491	0.074
			0.1987			(<0.01)	(<0.01)	(0.797)	(0.821)	(<0.01)	(>0.1)
VIX	-0.0022	0.1789	-0.3851	0.5683	0.7060	15.08	0.979	6.693^	11.445^	-11.236	0.045
			0.6458			(<0.01)	(<0.01)	(0.245)	(0.491)	(<0.01)	(>0.1)
NIK	-0.0020	0.0592	-0.2722	-0.7211	1.2763	31.18	0.968	5.531^	9.647^	-8.569	0.100
			0.1209			(<0.01)	(<0.01)	(0.355)	(0.647)	(<0.01)	(>0.1)
HANG	0.0028	0.0775	-0.3482	-0.4911	2.3805	55.90	0.965	1.388^	7.903^	-9.308	0.056
			0.2532			(<0.01)	(<0.01)	(0.926)	(0.793)	(<0.01)	(>0.1)
FTSE	0.0021	0.0430	-0.1395	-0.6979	0.5413	18.69	0.965	6.875^	12.161^	-9.871	0.093
			0.0849			(<0.01)	(<0.01)	(0.230)	(0.433)	(<0.01)	(>0.1)
MXC	0.0128	0.0695	-0.3498	-0.9523	3.2172	116.84	0.955	4.310^	10.850^	-9.662	0.054
			0.1766			(<0.01)	(<0.01)	(0.506)	(0.542)	(<0.01)	(>0.1)
BRA	0.0101	0.0924	-0.5034	-1.2405	4.6196	229.00	0.935	6.226^	13.550^	-9.399	0.074
			0.2155			(<0.01)	(<0.01)	(0.285)	(0.330)	(<0.01)	(>0.1)

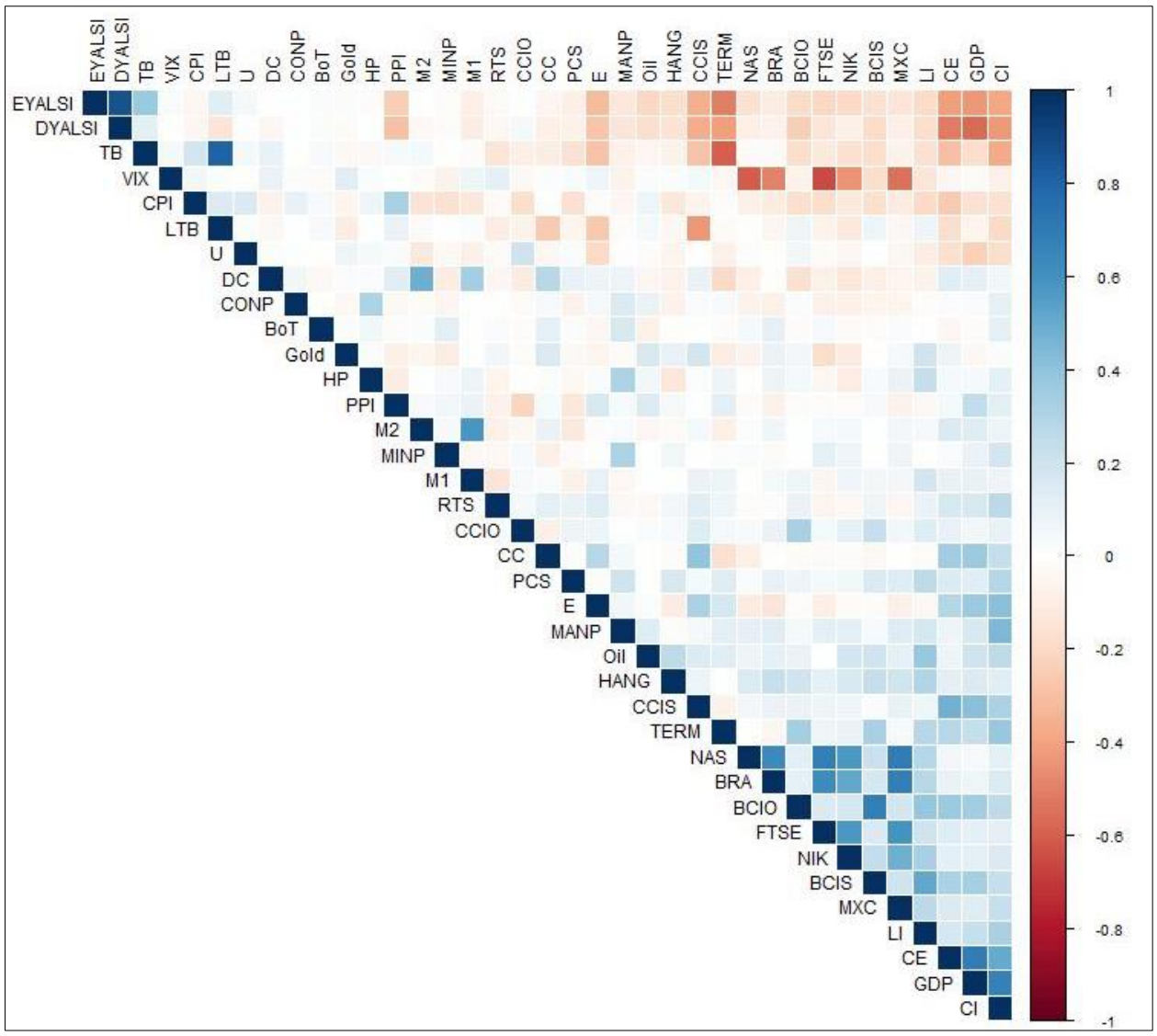
Note: p-values for the tests are reported in brackets. ^Atypical results in relation to the other series.

A comparison of the monthly mean returns of the ALSI (Table 5.1) with those of other markets shows the ALSI recording lower mean monthly returns than the Mexican Bolsa IPC Index (MXC) and Bovespa Brasil (BRA) but higher average monthly returns than the FTSE 100, Hang Seng Indices and Nikkei 225. The returns of the ALSI are roughly in line with those of the Nasdaq Composite Index (NAS) however the NAS has a higher standard deviation. The usefulness of direct mean return comparisons is doubtful in light of foreign exchange fluctuations, but the foreign indices may still be useful as proxies for factors in the global economic environment, which includes foreign exchange effects.

The results from normality tests (Table 5.2) show the two monetary aggregates M1 and M2 to be normally distributed. The Jarque-Bera test finds changes BCIO to be normally distributed but the Shapiro-Wilk test rejects a normal distribution for changes in this index. Ljung-Box tests on the macroeconomic variables find few with serial correlations up to the fifth and twelfth lag. The actively traded variables such as oil, gold and the international stock market indices all show significant serial correlation. In this the international indices stand in contrast to the ALSI and the

majority of JSE industry indices that show no serial correlation. The more important correlations to consider are the cross-correlations between explanatory variables illustrated in Figure 5.1.

Figure 5.1 The correlation-matrix for the ALSI explanatory variables.



The correlations for the ALSI set of possible features is illustrated in Figure 5.1. The differences between the ALSI feature set and the industry indices feature sets are sufficiently minor for Figure 5.1 to be a good general representation. The correlation matrices for the variables sets used for industry indices can be found in Appendix 2B. All variable Pearson cross-correlation coefficients are below 0.70, save that of TB and LTB at 0.812, and DY and EY (>0.8 in all cases).

5.2 MODELLING THE SELECTED FEATURES IN THE TRAINING PERIOD

This section displays the training period models computed based on the features selected for each index by the LARS and Random Forests feature selection techniques.

5.2.1 LINEAR MULTIFACTOR MODELS

The features selected by the LARS algorithm are fitted by OLS regression. The resultant models for the ALSI and the industry indices over the full sample period and the two sub-periods are shown in Table 5.4, Table 5.5, Table 5.6 and Table 5.7. The models are only fitted in the training sets of the respective sample periods leaving the test sets for robustness testing. Model fit is examined by evaluating the significance of coefficients, the adjusted R^2, the correlation of the fitted and actual values, the outcome of a Sign test on the fitted and actual values, and an assessment of their distributional properties via a Kolmogorov-Smirnov test. The correlation between the fitted and actual values is perhaps not as useful a measure of fit as the adjusted R^2. However, correlations are reported for the purposes of comparisons to the GRNN models for which it is difficult to compute a meaningful alternative to adjusted-R^2 measures. The assumptions of the regression, chiefly the independence of the residuals, are examined by conducting Ljung-Box test on the residuals to test for serial correlation, a Shapiro-Wilk test on the residuals assess possible normality and the analysis of residual plots (not reported).

It is immediately evident in the linear models that not all coefficients in the models are significant. Moreover, the LARS variable selection appears to be industry and period dependent – such that the set of relevant variables depends on the index being modelled and the period in which the modelling takes place. Assuming the variables selected to maximise the LARS fit (minimising C_p) are all relevant, a finding of insignificant coefficients in OLS regressions suggests a degree of redundancy in the selected variables. The LARS algorithm is applied equally with the same constraints to all the periods and industry indices. Following the notion that the explanatory variables merely proxy the true state factors that explain returns, if these factors remain the same over time, one would still expect a number of consistent explanatory variables to emerge that are

both industry and period independent. Consequently, industry and period dependency suggests that the underlying factors that determine equity returns change over time. However, this conjecture hinges on the ability of the variables selected by LARS to appropriately model returns.

Table 5.4 shows ALSI returns to be partially explained by proxies for the global economy and global investment – that is the international stock indices. In particular, index returns in emerging markets (BRA and MXC) and the FTSE 100, appear to be significantly positively related to the ALSI. The Nikkei 225 is chosen by LARS as a relevant variable but no significant coefficient is found in the OLS regression. The ALSI returns evidently reflect the global economy in which the exchange is embedded, but the model produced underlines the importance of the local economy. Macroeconomic variables with significant coefficients in the ALSI return model include the changes in the coincidental indicator (CI) produced by the South African Reserve Bank, growth rate in the monetary supply M1, and changes passenger car sales (PCS). LARS selected two additional macroeconomic variables (CCIO and Oil) and two fundamental variables (the ALSI earnings yield, EYALSI and dividend yield, DYALSI) for the ALSI model, but their coefficients are insignificant when modelled with the other variables over the full period.

Table 5.4 Linear APT models for the ALSI over the training periods.

Features Selected by LARS and Resultant Linear APT Models for the ALSI									
			Model Fit				Residuals		
Index	Period	Linear Models	adj- R^2	CORR	SIGN	KS-test	LB(5)	Shapiro
ALSI	f	**R** = -0.008 + 0.327**BRA*** + 0.286**MXC*** + 0.196**FTSE*** + 0.084**NIK** - 0.026**EYALSI** + 0.125**CI*** + 0.059**Oil** - 0.107**M1*** - 0.063**DYALSI** - 0.102**PCS*** - 0.0613**CCIO**	0.6446	0.8162	0.4945 (0.941)	0.1099 (0.222)	15.608 (0.08)	0.9923 (0.4482)
ALSI	s1	**R** = 0.023 + 0.24**MXC*** + 0.3**BRA*** + 0.21**FTSE*** + 0.09**NIK** + 0.16**EYALSI** + 0.16**CI*** - 0.32**DYALSI*** + 0.10**Oil** - 0.16**M1*** - 0.08**VIX** - 0.12**PCS*** + 0.09**Gold**	0.6550	0.8324	0.4727 (0.634)	0.1182 (0.426)	7.6718 (0.175)	0.9916 (0.74)
ALSI	s2	**R** = -0.02 + 0.43**FTSE*** + 0.30**BRA*** + 0.14**MXC** + 0.15**CI*** - 0.13**CCIO*** - 0.07**LTB**	0.6757	0.8328	0.4727 (0.634)	0.1 (0.641)	19.961^ (0.001)	0.9923 (0.494)

** Coefficients significant at a 5% significance level. ^Null hypothesis is rejected.*

It is intriguing that the coincidental indicator (CI) shows a positive relationship with ALSI returns. One might expect the leading indicator (LI) to have a stronger relationship with the ALSI given that equity returns are often touted as a leading economic indicator and indeed the leading indicator. There are few plausible explanations for the negative relationship between PCS and returns. One would expect increases in passenger car sales to reflect increases in real economic activity and therefore have a positive relationship with stock returns. The negative coefficient on

M1 may reflect its relationship with the real economy, or perhaps a view that increases in monetary supply result in future tightening of monetary policy. Overall the linear model using variables selected by LARS appears to be appropriate for the ALSI return data. Of the eleven variables selected by LARS, six have significant coefficients in OLS regressions. The adjusted-R^2 of the ALSI model implies that 64% of the ALSI variance is explained by changes in the variables selected by LARS.

The proposed ALSI model in sub-period 1 (Table 5.4) is very similar to the model for the full sample period. Most of the same factors are significant but unlike the full period ALSI model the coefficient for DYALSI is significant. This points to the relevant set of features being period dependent. The ALSI model for sub-period 1 also contains an additional factor as Gold and the VIX index replace CCIO from the full period model. Neither of these have significant coefficients. The adjusted-R^2 for the sub-period 1 ALSI model is slightly higher (0.66) than that of the full period model and analysis of the model fit and the residuals suggest the model is a good fit. The ALSI model in sub-period 2 (Table 5.4) contains different variables from the models suggested by LARS for the full period and sub-period 1. LARS selects fewer features in the sub-period 2 ALSI model. Of the six variables selected by LARS four are significant, and while variables common to all three periods FTSE, MXC, BRA, and CI persist, unlike in the other two periods MXC has an insignificant coefficient. Overall the sub-period 2 ALSI model has the highest adjusted-R^2 of the three linear ALSI models at 0.68 but the residuals exhibit a significant degree of serial correlation up to lag order five suggesting some bias in the model. Despite having different explanatory variables all the linear ALSI models achieved a reasonable fit to their training sets. Although the sub-period 2 model showed evidence of bias and failed to adhere to the model assumptions, two of the three linear ALSI models appeared wholly appropriate.

An analysis of the linear industry indices models emphasises that the variables that best explain returns are industry dependent. In a nod to the CAPM the ALSI is the only variable consistently selected by LARS for all the industry models. It also has a significant coefficient in all the models. Regardless, all the other variables appear to be model dependent. For the J500 (Oil and Gas) variables selected by LARS with significant coefficients include the ALSI, Oil, NAS, MXC,

MANP, TB, BCIO, PPI, M1 and DC (Table 5.5). Oil prices are evidently important to firms in the oil and gas industry as a positive coefficient suggests. Increases in MANP may reflect increased demand for energy but it is not immediately clear why decreases in domestic credit are associated with higher J500 returns. Unlike in the ALSI model M1 has a positive coefficient in the J500 model. Eight out of the thirteen variables selected by LARS have significant coefficients. The J500 model has the highest adjusted-R^2 value, 0.75, of all the full period models in the study. In a consistent theme in this study the J500 sub-period 1 and sub-period 2 models (Table 5.5) of variables selected by LARS differ markedly from the J500 model over the full sample period. In the J500 sub-period 1 model the TB selected for the full period model is replaced by LTB which has a significant coefficient. DC and BCIO no longer have significant coefficients. Curiously, changes in oil prices do not appear to be significantly related to sub-period 1 J500 returns. Only three of the twelve variables selected have significant coefficients. Altogether, the sub-period 1 J500 model retains the high adjusted-R^2 of 0.75 shown by the full period model, and goodness of fit tests indicate that this model is a good fit for J500 returns. In the sub-period 2 J500 model, LARS selects fewer variables (Table 5.5), and the model has a lower adjusted-R^2 than in the other two periods (0.67). Variables common to all three periods include the ALSI, MANP, Oil and BCIO. For the sub-period 2 J500 model MANP and BCIO have significant coefficients. However, as in the sub-period 1 J500 model the coefficient of the variable for the changes in oil prices is not significant. For the sub-period 2 J500 model LARS adds retail trade sales (RTS) as an explanatory variable. Its coefficient is negative and significant. Perhaps this is plausible if retail sales decrease as oil prices increase, suggesting some redundancy in the use of the two variables. From this viewpoint RTS may better proxy for the underlying effects of oil price changes (which may contain a leading or lagging effect) that are positive for energy producers and negative for retailers. The J500 models appear to be appropriate models and continue the general observation of the period dependency of variables.

Only three of the eleven variables selected by LARS for the J510 (Basic Materials) full period model have significant coefficients – ALSI, Oil and NAS (Table 5.5). However, the J510 model still achieves a higher adjusted-R^2 (0.68) than the ALSI full period model. Basic materials are inputs for production in the real economy. CONP and MINP are both selected by LARS as

explanatory variables for this model but their coefficients are surprisingly negative and not significant. Changes in housing permits is also selected as an explanatory variable for this industry index model but it too has a coefficient that is inexplicably negative and not significant. The Shapiro-Wilk test rejects the null hypothesis of normality of the residuals. Normality is only important in relation to the independence of individual residuals. Whilst the Ljung-Box test shows no serial correlation the residual plots for this model indicate a degree of nonlinearity. The model for the J510 (Table 5.5) in sub-period 1 has only three significant coefficients (ALSI, DY510 and NAS) from the nine variables selected by LARS. An adjusted-R^2 of 0.62 is lower than that of equivalent models over the full period and in sub-period 2. The model for J510 (Table 5.5) in sub-period 2 has three significant coefficients (ALSI and CC) of the eight variables selected by LARS. Nevertheless, this model has an adjusted-R^2 of 0.78 which is the highest of all the linear models tested in this study. R^2 alone is not the indisputably definitive measure for model fit but results of the Sign test, Kolmogorov-Smirnov test and analysis of the residual plots confirm that this model provides a good fit for the data. The J510 models for all three periods appear to fit appropriately and they show further evidence of period dependence and industry dependence of the variable subset. Only two variables are common to all three J510 models – ALSI and NAS.

Table 5.5 Linear APT models for the J500, J510 and J520 over the training periods.

Features Selected by LARS and Resultant Linear APT Models for the Industry Indices (J500, J510 and J520)										
			Model Fit				Residuals			
Index	Period	Linear Models	adj- R^2	CORR	SIGN	KS-test	LB(5)	Shapiro
J500	f	**R** = -0.025 + 0.99**ALSI*** + 0.11**Oil*** - 0.014**EY500** - 0.09**REY500** - 0.15**NAS*** + 0.09**MANP*** + 0.09**TB*** - 0.01**BCIO*** + 0.05**PPI** + 0.14**M1*** - 0.14**DC*** - 0.15**MXC*** -0.05**Gold**	0.7514	0.8771	0.4780	0.0879	2.6229	0.9855
					(0.604)	(0.483)	(0.758)	(0.057)
J500	s1	**R** = -0.044 + 0.97**ALSI*** + 00.08**Oil** + 0.02**CCIS** + 0.11**LTB*** - 0.22**NAS*** - 0.12**REY500** - 0.09**Gold** - 0.04**DC** - 0.03**U** - 0.13**BCIO** - 0.08**HP** - 0.05**CC**	0.7535	0.8835	0.4545	0.1182	2.7623	0.9853
					(0.391)	(0.426)	(0.737)	(0.272)
J500	s2	**R** = -0.028 + 0.75**ALSI*** + 0.13**MANP*** + 0.09**Oil** + 0.01**DY500** - 0.11**RTS*** - 0.14**EY500** - 0.15**BCIO*** - 0.06**CCIO**	0.6672	0.8316	0.4818	0.0818	5.7317	0.988
					(0.775)	(0.855)	(0.333)	(0.435)
J510	f	**R** = -0.02 + 0.89**ALSI*** + 0.08**Oil*** - 0.03**DY510** - 0.204**NAS*** + 0.11**LTB** - 0.08**MINP** - 0.06**CONP** - 0.06**PCS** + 0.06**BCIS** - 0.039**HP** + 0.07**CC** - 0.13**RDY510**	0.6818	0.8394	0.4890	0.0879	3.8542	0.9676^
					(0.824)	(0.483)	(0.571)	(<0.01)
J510	s1	**R** = 0.03 + 0.74**ALSI*** - 0.15**DY510*** + 0.08**Oil** + 0.09**BCIS** - 0.09**CONP** - 0.16**NAS*** - 0.07**M1** - 0.09**PCS** - 0.09**CCIS**	0.6170	0.8053	0.4818	0.0636	2.3709	0.9481^
					(0.775)	(0.979)	(0.796)	(<0.01)
J510	s2	**R** = -0.02 + 0.81**ALSI*** - 0.06**REY510** + 0.14**CC*** + 0.03**BRA** - 0.08**MINP** - 0.08**BoT** - 0.11**RDY510** + 0.06**Gold**	0.7815	0.8931	0.4909	0.0636	7.7246	0.9850
					(0.924)	(0.979)	(0.172)	(0.256)
J520	f	**R** = -0.054 + 0.6**ALSI*** + 0.14**BRA*** + 0.03**MXC** + 0.1**FTSE** - 0.29**RDY520*** + 0.02**NIK** + 0.06**EY520** + 0.07**BCIO** - 0.26**RCR520*** - 0.09**REY520** + 0.06**HP** - 0.15**TB*** + 0.04**CI**	0.7021	0.8506	0.5385	0.0714	5.356	0.9903
					(0.335)	(0.742)	(0.374)	(0.258)
J520	s1	**R** = 0.03 + 0.68**ALSI*** + 0.20**BRA*** - 0.14**RDY520*** + 0.03**MXC** - 0.14**REY520*** + 0.1**HP** - 0.17**RCR520***	0.7357	0.8676	0.5273	0.0727	9.7877	0.9806
					(0.634)	(0.933)	(0.081)	(0.109)
J520	s2	**R** = 0.1 + 0.41**ALSI*** + 0.16**FTSE** + 0.19**NAS** + 0.05**MXC** - 0.44**RDY520*** + 0.11**BCIO** - 0.27**RCR520*** + 0.14**RS520** - 0.31**S520** + 0.17**TERM** - 0.04**CPI**	0.6476	0.8266	0.5364	0.1364	7.0544	0.9826
					0.505)	(0.258)	(0.217)	(0.161)

** Coefficients significant at a 5% significance level. ^Null hypothesis is rejected.*

The J520 (Industrials) model in Table 5.5 includes thirteen variables five of which have significant coefficients. Variables with significant coefficients include ALSI, BRA, RDY520, RCR520 and TB. The model has an adjusted- R^2 of 0.70 but with thirteen explanatory variables it is amongst the models with the highest number of factors. The fact that RCR520 is significant and negative for this model suggests an implicit form of mean reversion – past cumulative returns is negatively related to current returns. A negative relationship between the yield on 91-day treasury bills and J520 returns may indicate the importance of economic effects of the costs of debt. The seven variables selected for the sub-period 1 J520 (Table 5.5) are all common to the full period model. Of the five variables with significant coefficients (ALSI, BRA, RDY520, REY520 and RCR520) all except REY520 are also significant in the full period model. This large overlap suggest that the important variables for the J520 remain fairly constant over time. With an adjusted-R^2 of 0.74 the sub-period 1 model for the J520 is favourably comparable to the full period model, providing a higher adjusted-R^2 with fewer variables. The sub-period 2 J520 model (Table 5.5) appears to have a worse fit than that of the sub-period 1 and full period, with an adjusted-R^2 of 0.65. It is however comparable to the other industry R^2's for the period. Of the eleven variables selected by LARS only three have significant coefficients, but all three significant coefficients are for variables also found to be significant in the sub-period 1 J520 model and the full period J520 model. Of all the linear models, the J530 models share the most common factors. These are ALSI, MXC, RDY520 and RCR520. There is no substantial evidence of bias in the residuals of any of the J520 linear models. Together with the model fitness tests this points to the conclusion that the models are appropriate descriptors of J520 returns.

The linear full period J530 (Consumer Goods) model contains eight variables (Table 5.6), of which ALSI, NAS and Gold have significant coefficients. The adjusted-R^2 this model is relatively poor at 0.54 and while there is little evidence of heteroscedasticity there does appear to be some bias in the residual plots. An explanation for the negative relationship with changes in gold prices could be found in the interpretation of gold as a safe haven asset. In economic downturns and periods of significant risks investors often buy gold while consumers cut back on discretionary items which is deleterious to retailers and other consumer goods companies. Of the nine variables chosen by LARS the linear sub-period 1 J530 (Table 5.6) has six significant coefficients (ALSI, FTSE, NAS,

DY530, EY530, and BCIO). At 0.60 the model adjusted-R^2 is higher than that of the equivalent full period model but lower than the sub-period 2 model. The J530 sub-period 2 model (Table 5.6) has four variables (ALSI, FTSE, BRA and HP) with significant coefficients in a model of twelve factors. The adjusted-R^2 of the model (0.62) is better than that of sub-period 2 and the full period model. While the adjusted-R^2 of the J530 models are not as high as those of other industry index models the J530 sub-period 1 and sub-period 2 models show no substantial bias in the model fit. The sub-period 1 and sub-period 2 J530 models appear to be appropriate models for the J530 returns in the respective sub-periods.

Table 5.6 Linear APT models for the J530, J540 and J550 over the training periods.

Features Selected by LARS and Resultant Linear APT Models for the Industry Indices (J530, J540 and J550)										
			Model Fit				Residuals			
Index	Period	Linear Models	adj- R^2	CORR	SIGN	KS-test	LB(5)	Shapiro
J530	f	**R** = 0.02 + 0.48**ALSI*** + 0.14**FTSE** + 0.17**NAS*** - 0.112**EY530** - 0.11**Gold*** - 0.1**CONP** - 0.06**DY530** - 0.08**CCIO**	0.5419	0.7497	0.5165 (0.711)	0.0879 (0.483)	7.357 (0.195)	0.9774^ (0.004)
J530	s1	**R** = 0.026 + 0.36**ALSI*** + 0.23**FTSE*** + 0.21**NAS*** - 0.17**DY530*** - 0.16**EY530*** + 0.10**NIK** - 0.05**CCIO** - 0.21**BCIO*** + 0.12**HANG**	0.5969	0.7938	0.5364 (0.505)	0.0818 (0.855)	8.6146 (0.125)	0.9879 (0.433)
J530	s2	**R** = 0.05 + 0.71**ALSI*** + 0.26**FTSE*** - 0.14**HP*** - 0.11**CONP** - 0.07**RDY530** - 0.10**CPI** - 0.08**CC** - 0.32**BRA*** + 0.06**M2** + 0.1**S530** - 0.09**PCS** - 0.08**RCR530**	0.6177	0.8123	0.4727 (0.634)	0.1 (0.641)	9.4907 (0.09)	0.9928 (0.838)
J540	f	**R** = 0.04 + 0.59**ALSI*** + 0.09**MXC** - 0.18**REY540*** - 0.13**TB** + 0.04**BCIO** - 0.03**VIX** - 0.13**Oil*** - 0.16**CCIS*** + 0.09**CCIO** - 0.05**CC** + 0.09**HANG** + 0.02**NIK**	0.5393	0.7549	0.5110 (0.824)	0.1099 (0.222)	7.876 (0.163)	0.9778^ (0.005)
J540	s1	**R** = 0.004 + 0.66**ALSI*** + 0.14**MXC** + 0.1**BCIO** - 0.08**TB** + 0.08**HANG** - 0.14**MANP*** - 0.1**Oil** - 0.06**REY540**	0.6748	0.8358	0.4727 (0.634)	0.1 (0.641)	16.67^ (0.005)	0.9837 (0.20)
J540	s2	**R** = 0.08 + 0.39**ALSI*** + 0.13**NAS** + 0.13**NIK** - 0.16**REY540** - 0.23**CCIS*** - 0.13**M1** - 0.11**BoT** + 0.16**CCIO***	0.3913	0.6603	0.5091 (0.924)	0.1545 (0.144)	2.3322 (0.802)	0.9897 (0.577)
J550	f	**R** = 0.01 + 0.52**ALSI*** + 0.22**NAS** + 0.05**MXC** - 0.22**LTB*** + 0.021**BRA** + 0.14**BCIO*** - 0.1**M1*** - 0.13**CC*** - 0.11**Oil*** + 0.03**REY550** + 0.089**Gold** - 0.12**TERM*** - 0.11**RDY550**	0.6359	0.8136	0.5110 (0.824)	0.1044 (0.275)	20.21^ (0.001)	0.9916 (0.367)
J550	s1	**R** = 0.023 + 0.45**ALSI*** + 0.10**MXC** + 0.18**NAS*** + 0.07**BRA** - 0.17**LTB*** + 0.08**Gold** - 0.17**TERM*** + 0.03**NIK** + 0.19**BCIO*** - 0.12**EY550*** - 0.11**M2***	0.6856	0.847	0.4727 (0.634)	0.1182 (0.426)	14.543^ (0.013)	0.9904 (0.63)
J550	s2	**R** = -0.002 + 0.34**ALSI*** + 0.16**FTSE** + 0.24**NAS**	0.4602	0.6892	0.4909 (0.924)	0.1545 (0.144)	7.8052 (0.167)	0.977 (0.054)

** Coefficients significant at a 5% significance level. ^Null hypothesis is rejected.*

The linear J540 (Healthcare) models display a large amount of variation in model fit and the J540 proved difficult to model. The full period J540 model has twelve variables (Table 5.6). The four variables with significant coefficients are ALSI, REY540, Oil and CCIS. The adjusted-R^2 of the model is 0.54 placing the model amongst the poorer industry index models for the full period, in terms of fit. There are a number of outliers in the residual plots and the Shapiro-Wilk test on the full period model rejects normality of the residuals. There is some evidence of bias in the residual plots for the full period J540 model. The sub-period 1 J540 model (Table 5.6) has eight factors,

two are significant (ALSI and MANP). The model appears to explain 67% of the variation in J540 returns which is the highest adjusted-R^2 amongst the J540 training period models. Unfortunately the residuals are shown to be serially correlated contesting the fit of the model. In the J540 models for the full period and sub-period 2 LARS could not decipher between the two consumer confidence indicators. Both are selected in the full period and in sub-period 2 and in sub-period 2 both also have significant coefficients (Table 5.6). From an explanatory point of view this is less than ideal but the correlation matrix computed in Figure 5.1 suggests that the two represent novel innovations. The ALSI is the only other variable with a significant coefficient in this model. The ALSI and the relative earnings yield measure REY540 are common to all the J540 models. At 0.39 the sub-period 2 J540 model has the lowest adjusted-R^2 of all the linear models in this study. This low adjusted-R^2 may suggest that this model is not appropriate. In addition, the serial correlation in the sub-period 1 model also suggests that a model that is not well suited to the data.

For the full period J550 (Consumer Services) model LARS selects thirteen explanatory variables (Table 5.6). Seven variable coefficients are significant (ALSI, LTB, BCIO, M1, CC, Oil and TERM) and the model has an overall adjusted-R^2 of 0.64. The residual plots indicate possible heteroscedasticity and consequently bias in the OLS regression. This result is corroborated by the Ljung-Box test which finds serial correlation in the residuals. The sub-period 1 model for the J550 (Table 5.6) contains seven variables with significant coefficients (ALSI, NAS, LTB, TERM, BCIO, EY550 and M2), out of the eleven variables selected by LARS. The adjusted-R^2 for the model is 0.69 which is the highest of the linear J550 models but there is evidence of serial correlation in the residuals suggesting some model bias. The sub-period 2 J550 model (Table 5.6) has the fewest explanatory variables of the models in the study. Of the three variables only the ALSI has a significant coefficient. This model does achieve a slightly higher adjusted-R^2 (0.46) than the sub-period 2 J540 but it is significantly worse than its full period and sub-period 1 equivalents. Evidence of serial correlation in the residuals of the full period and sub-period 1 models indicates that they may not be appropriate models. The sub-period 1 J550 model has slightly less problematic residuals but it has the lower adjusted-R^2.

The full period J560 (Telecommunications) has a low adjusted-R^2 of 0.40 (Table 5.7). This is the lowest of all the full period models and the lowest of the J560 models across periods – although none of the other J560 adjusted-R^2's are particularly high. In addition to the low adjusted-R^2 the full period J560 model also shows significant serial correlation in the residuals and evidence of bias in the residual plots. The sub-period 1 J560 model (Table 5.7) contains only six variables and four have significant coefficients (ALSI, NAS, DY560 and BCIO). This model has the lowest adjusted-R^2 of all the industry models in sub-period 1 at 0.48. The residual plots show some evidence of heterscedasticity in the residuals. The sub-period 2 J560 model (Table 5.7) is made up of twelve variables and eight have significant coefficients (ALSI, MXC, DC, CE, M1, HANG, RS560 and LI). In spite of the many explanatory variables, the model only explains 54% of returns. Across the periods considered, the J560 appears to be amongst the most difficult indices to model. Perhaps the examined economic variables do not provide suitable proxies for the underlying economic effects affecting telecommunications firms. Altogether the J560 models have relatively low adjusted-R^2's. The residual plots for the sub-period 2 and full period models reveal bias in the models.

Table 5.7 Linear APT models for the J560, J580 and J590 over the training periods

Features Selected by LARS and Resultant Linear APT Models for the Industry Indices (J560, J580 and J590)									
			Model Fit				Residuals		
Index	Period	Linear Models	adj- R^2	CORR	SIGN	KS-test	LB(5)	Shapiro
J560	f	**R** = -0.007 + 0.296**ALSI*** + 0.21**NAS*** + 0.18**BRA** + 0.06**MXC**	0.3989	0.642	0.4780 (0.604)	0.1154 (0.177)	11.083^ (0.05)	0.9674^ (<0.001)
J560	s1	**R** = -0.03 + 0.29**ALSI*** + 0.17**BRA** + 0.26**NAS*** - 0.24**DY560*** + 0.01**MXC** + 0.21**BCIO***	0.4816	0.714	0.5000 (>0.99)	0.1091 (0.530)	10.126 (0.072)	0.9822 (0.15)
J560	s2	**R** = -0.082 + 0.24**ALSI*** + 0.38**MXC*** + 0.06**NAS** + 0.35**DC*** + 0.1**MANP** + 0.22**CE*** - 0.25**M1*** + 0.14**HANG*** + 0.12**MINP** + 0.29**RS560*** - 0.15**LI*** + 0.12**CONP**	0.5394	0.768	0.5000 (>0.99)	0.0909 (0.754)	8.409 (0.135)	0.9895 (0.557)
J580	f	**R** = -0.004 + 0.62**ALSI*** + 0.13**MXC** - 0.18**M1*** - 0.05**FTSE** + 0.07**BRA** - 0.11**REY580*** + 0.09**NIK** - 0.14**Oil*** + 0.09**CONP** - 0.09**TB** + 0.06**HP** + 0.06**BCIO**	0.6786	0.837	0.5330 (0.415)	0.0714 (0.742)	4.646 (0.461)	0.9434^ (<0.01)
J580	s1	**R** = 0.006 + 0.67**ALSI*** + 0.16**MXC*** + 0.07**BRA** - 0.15**M1*** + 0.02**FTSE** - 0.14**Oil*** + 0.11**CONP***	0.7325	0.866	0.5091 (0.924)	0.1 (0.641)	2.2247 (0.817)	0.9825 (0.158)
J580	s2	**R** = 0.015 + 0.48**ALSI*** + 0.25**NAS*** - 0.08**FTSE** + 0.16**NIK** - 0.07**REY580** - 0.22**RDY580*** + 0.03**BCIO** + 0.04**CE** + 0.2**LTB*** + 0.07**PCS** + 0.10**HP**	0.6010	0.801	0.4545 (0.391)	0.1091 0.530)	7.1753 (0.208)	0.9123^ (<0.01)
J590	f	**R** = 0.005 + 0.41**ALSI*** + 0.41**NAS*** - 0.005**RCR590** + 0.11**DY590** - 0.17**EY590*** + 0.12**Gold*** - 0.12**BCIO*** + 0.11**PCS*** - 0.09**RTS** + 0.08**CONP** - 0.03**CE** - 0.05**LTB** + 0.05**MINP**	0.6096	0.799	0.5275 (0.505)	0.0604 (0.894)	1.1179^ (0.008)	0.9757^ (0.003)
J590	s1	**R** = 0.123 + 0.36**ALSI*** + 0.20**NAS** + 0.31**S590*** - 0.11**BoT** + 0.20**PCS*** - 0.11**CC** - 0.36**RCR590*** - 0.16**MANP** - 0.24**TB*** + 0.08**HANG** + 0.03**NIK** - 0.08**RTS**	0.6096	0.799	0.5364 (0.505)	0.1455 (0.195)	4.3837 (0.496)	0.9894 (0.55)
J590	s2	**R** = 0.001 + 0.45**ALSI*** + 0.44**NAS*** - 0.16**EY590*** + 0.24**Gold*** - 0.19**BCIO*** + 0.13**CONP*** + 0.02**CE** + 0.04**FTSE** - 0.03**PPI**	0.7087	0.8560	0.5364 (0.505)	0.0727 (0.933)	1.2418 (0.94)	0.9408^ (0.005)

** Coefficients significant at a 5% significance level. ^Null hypothesis is rejected.*

Twelve variables are selected for the full period J580 (Financial) model. Coefficients of four of those variables, ALSI, M1, REY580, and Oil, are significant. The model has an adjusted-R^2 of 0.68 which is higher than that of the sub-period 2 J580 but lower than that of the sub-period 1 J580 model. The model residuals are not normally distributed and the residual plots show that this is likely due to the presence of outlier suggesting a degree of bias in estimates. The sub-period 1 J580 model contains seven variables of which ALSI, MXC, M1, Oil and CONP have significant coefficients. This particular model appears to have a good fit with no evidence of substantial bias in the residuals and an adjusted-R^2 that implies the model explains 73% of the variance in sub-period 1 J580 returns. The J580 model in sub-period 2 contains eleven variables of which ALSI, NAS, RDY580 and LTB, have significant coefficients. LTB is notable as the positive coefficient suggests the returns to the financial industry index are positively related to long term government bond yields. This may relate to the underlying interest rate environment.

The J590 (Technology) model (Table 5.7) for the full period has thirteen variables with six significant coefficients (ALSI, NAS, EY590, Gold, BCIO and PCS). It is notable that the coefficient on the returns of the Nasdaq Composite Index (NAS) is positive and significant. As the Nasdaq is a leading technology heavy index one would expect a strong relationship with worldwide technology indices, reflecting the global growth in the industry. The full period J590 model has a relatively low adjusted-R^2 at 0.61 but there is evidence of autocorrelation in the residuals and the residual plot display likely heteroscedasticity. The coefficients of the ALSI, NAS, EY590, Gold, BCIO and CONP are significant in the sub-period 1 J590 (Table 5.7). These variables are amongst nine chosen by LARS for the model. Although the residuals are not normally distributed and plots appear to show some evidence of bias the overall adjusted- $R^2$of the model is 0.71. The sub-period 2 J590 model (Table 5.7) contains twelve variables. Five have significant coefficients including ALSI, S590, PCS, RCR590 and TB. As expected NAS is amongst the variables selected however its coefficient is not significant. The adjusted-R^2 of this model is 0.61 and the model diagnostics show the fit to be reasonable.

With the notable exceptions of the J540 and J550 indices, the fit of the models is generally better over the shorter sub-periods than over the full period. There appears to be a benefit to fitting the

models over smaller data sets. Intrinsically this makes sense as a shorter time period reduces the amount of variation a model has to explain.

Table 5.8 Tally of the number of variables chosen by LARS and the number with significant coefficients.

Overview of Linear Models				
	Full period	Sub-period 1	Sub-period 2	Common variables
ALSI	11 (6)	12 (7)	6 (4)	BRA, MXC, FTSE, CI
J500	13 (9)	12 (3)	8 (5)	ALSI, Oil
J510	12 (3)	9 (3)	8 (2)	ALSI, NAS
J520	13 (5)	7 (5)	11 (3)	ALSI, MXC, RDY520, RCR520
J530	8 (3)	9 (6)	12 (4)	ALSI, FTSE
J540	12 (4)	8 (2)	8 (3)	ALSI, REY540
J550	13 (7)	11 (7)	3 (1)	ALSI, NAS
J560	4 (2)	6 (4)	12 (8)	ALSI, NAS, MXC
J580	12 (4)	7 (5)	11 (4)	ALSI, FTSE
J590	13 (6)	9 (6)	12 (5)	ALSI, NAS

Note: Number of significant coefficients are in brackets.

The variable subsets selected by LARS to model the ALSI and industry indices display substantial variation in both the number of variables selected and the nature of the variables selected. An analysis of Tables 5.4, 5.5 and 5.6 and 5.7 shows that variable subsets are different for the different indices. For the ALSI the variables subsets differ across sub-periods. For industry indices the variable subsets also display this period dependence. Taken together, this constitutes substantive evidence for the rejection of *Hypotheses 3.1* and *3.2*. Indeed, only the ALSI is a common variable in every industry index model. Table 5.8 presents a tally of the number of variables in each linear model and highlights the variables common across sub-periods for each index. There are four variables which are consistently present in all the ALSI linear models. In light of the fact that these four variables are not always significant, and other variables which are significant in one period are not even deemed relevant in others, then these four variables do not seem to represent evidence for a sub-period independent return model. Indeed, the only substantial evidence of the same variables being consistently selected across the sub-periods could perhaps be found in the linear J520 models. Like the ALSI models four variables are in all the feature subsets selected by LARS for the J520, but in each of the estimated models at least three of these four variables has a significant coefficient. Furthermore, only one other variable, REY520, is significant in any of the J520 models. Nevertheless, on the weight of evidence, a rejection of *Hypotheses 3* seems to be the most sensible deduction.

The intercepts are not significant at a 5% significance level in any of the models. This, in addition to the generally adequate model fits, which are exemplified by the generally high adjusted-R^2 values observed, constitutes evidence that the models based on variables selected by LARS are viable. The models do not provide sufficient evidence to suggest the rejection of *Hypotheses 1.1*. For a more comprehensive evaluation of *Hypotheses 3* and for a test of *Hypotheses 1.2* GRNN models must be analysed.

5.2.2 GRNN MULTIFACTOR MODELS

Variables selected for the ALSI and industry indices, using the Boruta wrapper for Random Forests, are employed to train a GRNN. The selected variables for the GRNN models of the ALSI and the industry indices over the full sample period and two sub-periods are shown in Table 5.9, Table 5.10, Table 5.11 and Table 5.12. The black-box nature of neural networks means that many of the model tests easily applied to linear OLS-regressions have no neural network equivalent. For instance, neural networks provide no equivalent to adjusted-R^2 and therefore the onus is on less perceptible metrics to evaluate model fit. The GRNN is trained using the training sets of each period. The trained GRNN is used to estimate the target values in the training set, resulting in a set of fitted values. Tests of the model fit consist of evaluating the correlation of the fitted and actual values, performing a Kolmogorov-Smirnov test to assess the distributional properties of the fitted and actual values, and investigating the outcome of a Sign test on the fitted and actual values. Training a neural network comprises an intractable problem and hence no closed form expression exists for the fitted parameters. Instead, the pattern and summation layers of the GRNN pursue parameter optimisation, which is first localised then appropriately weighted and aggregated. The GRNN imposes a structure on the residuals. This likely connotes a non-normal distribution for the residuals, but the expectation that residuals consist of stochastic error remains. Similar to the linear models, the residuals of the GRNN models are evaluated by testing for serial correlation up to five lags, assessing possible normality, and examining residual plots. A suitable GRNN model should not display evidence of serial correlation, heteroscedasticity, or lingering nonlinearity in the residuals.

As with the linear models it is clear that the variables selected are both index and period dependent. There is also substantial variation in the number of variables selected for each model. The residuals for all the GRNN models are not normally distributed, as evidenced by the rejection of the null hypothesis of the Shapiro-Wilk test in each model. The ALSI model over the full period contains ten variables (Table 5.9). The correlation between the fitted and actual values in this model is 0.89. This compares to a correlation of 0.82 in the linear full period ALSI model, and correlations of 0.86 and 0.90 for the GRNN sub-period 1 and sub-period 2 ALSI models, respectively. The GRNN sub-period 1 and sub-period 2 models also contain ten variables. While the variables are not the same for each of the nonlinear ALSI models several commonalities are evident. Seven variables are common to all the GRNN ALSI models. As in the linear ALSI models the returns of international stock indices, BRA, MXC and FTSE are consistently selected as features by the Random Forests. The Random Forests also select other international equity indices such as the NAS, NIK, HANG and the volatility index VIX for the ALSI GRNN models in each period. CI which is regularly selected by LARS for linear ALSI models is selected for two of the three GRNN ALSI models. The GRNN ALSI models across periods are found to have fitted and actual values which follow the same distribution. There is no evidence of serial correlation heteroscedasticity or nonlinearity in the residuals of any of the ALSI GRNN models. However, there is evidence of outliers (numerous in the case of the sub-period 2 ALSI model) which may introduce bias to the models.

Table 5.9 Variables selected for GRNN modelling of the ALSI.

Features Selected by Boruta for the ALSI and Metrics for the GRNN Model Fit								
			Model fit			Residuals	
Index	Period	Symbolic Nonlinear Models	CORR	SIGN	KS-test	LB(5) tes	Shapiro
ALSI	f	**R ~ f (Oil, CI, NAS, NIK, HANG, FTSE, VIX, BRA, MXC, EYALSI)**	0.8910	0.4066	0.1043	8.110	0.7647^
				(0.014)	(0.27)	(0.150)	(<0.01)
ALSI	s1	**R ~ f (Gold, Oil, NAS, NIK, FTSE, VIX, BRA, MXC, EYALSI, DYALSI)**	0.8624	0.4455	0.1545	7.031	0.9204^
				(0.29)	(0.14)	(0.22)	(<0.01)
ALSI	s2	**R ~ f (DC, MANP, CI, NAS, NIK, FTSE, VIX, BRA, MXC, TERM)**	0.8971	0.4273	0.1182	4.6224	0.9483^
				(0.15)	(0.43)	(0.46)	(<0.01)

^ *Significant at a 5% significance level.*

The J500 GRNN model for the full period (Table 5.10) has twelve variables and a correlation of 0.85 between fitted the actual values. The GRNN J500 sub-period models have fewer variables (10 in the sub-period 1 model and 8 in the sub-period 2 model) and comparable correlations. In

the case of the sub-period 2 J500 its correlation of 0.87 between its fitted and actual values is higher than that of the full period model. When compared to the correlations between fitted and actual values in the linear models only the sub-period 2 GRNN model reports a higher correlation. This could suggest that in modelling the J500 a linear model is more appropriate. Furthermore, KS tests on the GRNN J500 full period and sub-period 1 models reject the null hypothesis that the fitted and actual values have the same distribution. Evidence of serial correlation can also be found in the GRNN sub-period 2 J500 model. The four variables are common to the J500 GRNN models (ALSI, CI, FTSE and BRA). Incidentally, changes in oil prices is not amongst the variables deemed relevant in all the GRNN J500 models. It is selected by Random Forests as a relevant variable for the full period model and for the sub-period 1 model, but not in the sub-period 2 GRNN model. One would expect changes in oil prices to play a critical role in all sub-periods, but perhaps other factors better represent the pertinent economic effects of changes in oil prices. The only variable that is selected consistently by both Random Forests and LARS for the modelling of J500 returns is the ALSI.

Table 5.10 Variables selected for GRNN modelling of the J500, J510 and J520.

Features Selected by Boruta for the Industry Indices (J500, J510 and J520) and Metrics for the GRNN Model								
			Model fit			Residuals	
Index	Period	Symbolic Nonlinear Models	CORR	SIGN	KS-test	LB(5) tes	Shapiro
J500	f	**R ~ f (ALSI, Oil, CI, NIK, HANG, FTSE, VIX, BRA, MXC, ALSI, EY500, REY500)**	0.8515	0.4121	0.1484^	5.337	0.9285^
				(0.02)	(0.04)	(0.376)	(<0.01)
J500	s1	**R ~ f (ALSI, Gold, Oil, CI, FTSE, BRA, MXC, CCIS, EY500, DY500)**	0.8496	0.4455	0.1818^	12.479^	0.9153^
				(0.29)	(0.05)	(0.03)	(<0.01)
J500	s2	**R ~ f (ALSI, MANP, CI, FTSE, VIX, BRA, DY500, RDY500)**	0.8657	0.4000	0.1727	1.6591	0.9117^
				(0.04)	(0.08)	(0.89)	(<0.01)
J510	f	**R ~ f (ALSI, Oil, BCIO, NAS, NIK, FTSE, VIX, BRA, MXC, GDP, DY510)**	0.8219	0.4560	0.1264	2.099	0.7887^
				(0.266)	(0.109)	(0.835)	(<0.01)
J510	s1	**R ~ f (ALSI, Oil, TB, BCIO, NAS, BRA, MXC, DY510)**	0.8906	0.5091	0.1273	3.6702	0.8134^
				(0.85)	(0.34)	(0.60)	(<0.01)
J510	s2	**R ~ f (ALSI, MINP, CI, NAS, NIK, FTSE, VIX, BRA, MXC)**	0.8774	0.5091	0.1454	4.132	0.9165^
				(0.92)	(0.19)	(0.53)	(<0.01)
J520	f	**R ~ f(ALSI, NAS, NIK, FTSE, VIX, BRA, MXC, CI, E, TERM, EY520, REY520, RDY520)**	0.8537	0.4890	0.1319	12.445^	0.7456^
				(0.82)	(0.084)	(0.029)	(<0.01)
J520	s1	**R ~ f(ALSI, NAS, NIK, FTSE, BRA, MXC, EY520, REY520, RDY520)**	0.8444	0.4091	0.1636	12.891^	0.9113^
				(0.07)	(0.11)	(0.02)	(<0.01)
J520	s2	**R ~ f (ALSI, CC, PCS, NAS, NIK, FTSE, VIX, BRA, MXC, GDP, TERM, RDY520)**	0.9309	0.5182	0.1	6.8089	0.9210^
				(0.78)	(0.64)	(0.24)	(<0.01)

^ *Significant at a 5% significance level.*

The full period J510 GRNN model (Table 5.10) contains eleven variables compared to eight for the J510 sub-period 1 model and nine for J510 sub-period 2 model. The fitted and actual values in the J510 full period model have a correlation of 0.82 which is lower than that of the linear J510 full period model and the equivalent GRNN sub-period 1 and sub-period 2 models (0.89 and 0.87, respectively). The GRNN sub-period 2 J510 model has a lower correlation between its fitted and actual values than the linear sub-period 2 J510. The opposite is true of the respective GRNN sub-period 1 and sub-period 2 models. The GRNN J510 models have only four variables in common. These are ALSI, NAS, BRA and MXC. Consequently, they share only two variables, ALSI and NAS, with the common variables selected by LARS for the linear J510 models. The Random Forests select a relatively lengthy feature set for the GRNN J520 full period model (Table 5.10), although the thirteen variables selected matches the number of variables selected by LARS for the J520 full period linear feature set. While this is the largest number of variables selected for any linear model, it is one shy of the total number of variables selected for the largest GRNN model, the full period J590 model. As with all indices the variables selected by LARS and Random Forests differ. The correlation of the fitted and actual values for the full period J520 model is 0.85, which is the same as that of the linear J520 full period model, but lower than the GRNN J520 sub-period 2 model (0.93) which achieves the highest correlation between fitted and actual values of any model in this study. The GRNN sub-period 1 J520 model has nine variables while the sub-period 2 J520 has twelve. The J520 models all have high correlations but the full period J520 and sub-period 1 J520 model have residual plots which indicate substantial bias. Serial correlation is also evident in the residuals for these models. The J520 models have many variables in common (ALSI, NAS, NIK, FTSE, BRA, MXC, RDY520). This may be evidence of an underlying constant model or it could simply represent variable redundancy.

The GRNN J530 full period model (Table 5.11) has eight variables as does the J530 model for sub-period 1. The J530 sub-period 2 model, however, has six variables (Table 5.11). Four variables are common to the GRNN J530 models (ALSI, NAS, NIK, and FTSE). For the full period model the correlation of the fitted and actual values is 0.79, however there the Kolmogorov-Smirnov test indicates that the actual and fitted values of this model are not from the same distribution. Similarly the Kolmogorov-Smirnov test for the sub-period 2 J530 model also rejects the null. The sub-period

1 J530 model has a correlation of 0.88 and appears to be an appropriate model although there is evidence of bias in the residuals. The ALSI, NAS, NIK and FTSE are the common variables in all the GRNN J530 models.

Table 5.11 Variables selected for GRNN modelling of the J530, J540 and J550.

Features Selected by Boruta for the Industry Indices (J530, J540 and J550) and Metrics for the GRNN Model								
			Model fit			Residuals		
Index	Period	Symbolic Nonlinear Models	CORR	SIGN	KS-test	LB(5) tes	Shapiro
J530	f	**R ~ f(ALSI, CPI, NAS, NIK, FTSE, VIX, TERM, EY530)**	0.7855	0.4780	0.1978^	20.263^	0.8985^
				(0.60)	(<0.01)	(<0.01)	(<0.01)
J530	s1	**R ~ f(ALSI, MINP, NAS, NIK, FTSE, BRA, MXC, EY530)**	0.8779	0.4909	0.1545	9.8076	0.9587^
				(>0.99)	(0.14)	(0.08)	(<0.01)
J530	s2	**R ~ f (ALSI, CONP, HP, NAS, NIK, FTSE)**	0.7753	0.4818	0.2091^	5.45865	0.9457^
				(0.78)	(0.02)	(0.36)	(<0.01)
J540	f	**R ~ f(ALSI, NAS, NIK, FTSE, VIX, BRA, MXC, BCIO, EY540, DY540, REY540)**	0.8491	0.4725	0.1374^	4.3245	0.8768^
				(0.824)	(0.06)	(0.50)	(<0.01)
J540	s1	**R ~ f(ALSI, HANG, FTSE, VIX, BRA, MXC, EY540, DY540)**	0.7933	0.4182	0.2182^	1.8425	0.9534^
				(0.10)	(0.01)	(0.87)	(<0.01)
J540	s2	**R ~ f(ALSI, CC, NAS, NIK, FTSE, VIX, BRA, MXC, CCIS, REY540, RS540)**	0.8084	0.4909	0.1364	3.7379	0.9100^
				(0.92)	(0.26)	(0.59)	(<0.01)
J550	f	**R ~ f(ALSI, Gold, NAS, NIK, FTSE, VIX, BRA, MXC, LTB, TB, EY550, REY550)**	0.8290	0.4780	0.0989	14.366^	0.6235^
				(0.60)	(0.34)	(0.01)	(<0.01)
J550	s1	**R ~ f(ALSI, Gold, LTB, NAS, NIK, FTSE, VIX, BRA, MXC)**	0.8876	0.4273	0.1727	15.288^	0.9586^
				(0.15)	(0.075)	(0.01)	(<0.01)
J550	s2	**R ~ f (ALSI, NAS, NIK, FTSE, VIX, BRA, MXC, CCIS, DY550, RDY550, RS550)**	0.8647	0.5273	0.1273	12.29^	0.9572^
				(0.63)	(0.34)	(0.03)	(<0.01)

^ *Significant at a 5% significance level.*

The GRNN full period J540 model (Table 5.11) and the sub-period 2 J540 model both contain eleven variables while the sub-period 1 equivalent contains eight variables. Five variables are present in all the J540 variables sets (ALSI, FTSE, VIX, BRA and MXC). The full period J540 has a correlation of 0.85 which ranks it amongst the better performing industry models in this period by this measure. The full period model has a higher correlation than either of the sub-period models. The full period model does not fare as well in the diagnostic tests, however. The fitted and actual values for the model have different distributional properties and residual plots appear show outliers and evidence of heteroscedasticity which may lead to model bias. The sub-period 1 model encounters similar problems but the sub-period 2 J540 GRNN model appears to be appropriate. Variables common to all the GRNN J540 models include ALSI, FTSE, VIX, BRA, and MXC. As Table 5.6 illustrates the J540 and J550 indices are difficult to model consistently using the linear

variables selected by LARS. The correlations for the J540 improve in the GRNN models. This is especially true for the sub-period 2 J540 model.

The full period J550 (Table 5.11) model contains twelve variables. The sub-period 1 J550 model contains nine variables and the sub-period 2 J550 eleven. The J550 models contain seven common variables. While the J550 models all have relatively high correlations of fitted and actual values, like their linear equivalents they show significant serial correlation in the residuals. The J560 models in Table 5.12, appear to be well specified. All three have relatively high correlations which are higher than their linear equivalents. The J580 full period model (Table 5.12) has eight variables and a correlation of its fitted values and the actual values of 0.78. This is relatively low compared to the correlation of 0.83 for the sub-period 2 J580 model and 0.94 for the sub-period 1 J580 model. The high correlation for the sub-period 1 J580 model echoes that of its linear counterpart. There are outliers in residual plots of the sub-period 2 and the full period J580 models. Still, all the J580 models appear relatively well specified. In a familiar theme the variables common to the J580 models consist of the ALSI and the international indices.

Table 5.12 Variables selected for GRNN modelling of the J560, J580 and J590.

Features Selected by Boruta for the Industry Indices (J560, J580 and J590) and Metrics for the GRNN Model								
			Model fit			Residuals	
Index	Period	Symbolic Nonlinear Models	CORR	SIGN	KS-test	LB(5) tes	Shapiro
J560	f	**R ~ f(ALSI, Gold, NAS, NIK, FTSE, BRA, MXC, CE, HP, DY560, RDY560)**	0.8004	0.5220 (0.604)	0.1154 (0.18)	6.623 (0.25)	0.8020^ (<0.01)
J560	s1	**R ~ f(ALSI, Gold, HP, NAS, NIK, FTSE, BRA, MXC, GDP)**	0.8832	0.4364 (0.25)	0.0727 (0.93)	7.167 (0.21)	0.6837^ (<0.01)
J560	s2	**R~ f (ALSI, DC, NAS, NIK, FTSE, BRA, MXC, TERM)**	0.8747	0.5000 (>0.99)	0.1091 (0.53)	5.759 (0.33)	0.8984^ (<0.01)
J580	f	**R ~ f(NAS, NIK, FTSE, VIX, BRA, MXC, TERM, ALSI)**	0.7763	0.4890 (0.82)	0.1264 (0.11)	5.401 (0.37)	0.7915^ (<0.01)
J580	s1	**R ~ f(ALSI, NAS, NIK, FTSE, VIX, BRA, MXC)**	0.9386	0.4364 (0.29)	0.0636 (0.98)	4.2788 (0.51)	0.6084^ (<0.01)
J580	s2	**R ~ f (ALSI, NAS, NIK, FTSE, VIX, BRA, MXC, EY580, REY580, RDY580)**	0.8289	0.4636 (0.5)	0.1091 (0.53)	9.7654 (0.08)	0.7909^ (<0.01)
J590	f	**R ~ f(LTB, TB, NAS, NIK, FTSE, VIX, BRA, MXC, ALSI, EY590, DY590, REY590, RDY590, RCR590)**	0.8967	0.4451 (0.16)	0.1428^ (0.049)	1.2524 (0.94)	0.9445^ (<0.01)
J590	s1	**R ~ f(ALSI, Gold, NAS, FTSE, BRA, MXC, EY590, REY590, RCR590)**	0.8808	0.4364 (0.25)	0.1 (0.64)	1.4801 (0.92)	0.6790^ (<0.01)
J590	s2	**R ~ f(ALSI, TB, NAS, NIK, FTSE, MXC)**	0.7709	0.4909 (0.92)	0.2455^ (<0.01)	1.1337 (0.95)	0.9337^ (<0.01)

^ *Significant at a 5% significance level.*

At fourteen, the J590 (Table 5.12) full period model contains the most explanatory variables of all the GRNN models. The correlation of the full period J590 model fitted values with the actual values is 0.90. This represents a stronger correlation than observed in the sub-period 1 and sub-period 2 J590 models. This is also a higher correlation than observed in any of the linear J590 models. Despite this, the Kolmogorov-Smirnov test rejects the null hypothesis, suggesting that the fitted and actual values for this model have different distributions. The sub-period 2 J590 model suffers from the same frailty whilst the sub-period 1 model appears to be appropriate.

Many more variables are common to GRNN models across periods for each index. This may be an indication that a GRNN approach better captures the relevant features. However, as a caution it is worth noting that this could also be a manifestation of a higher degree of variable redundancy if the variables proxy the same underlying effect. Regardless, the period dependency of the variable selection in both linear and GRNN models is clear.

Table 5.13 Tally of the number of variables chosen by Boruta and the common variables.

Overview of Nonlinear Models					
				Common Variables	
	Full period	Sub-period 1	Sub-period 2	Nonlinear Periods	All Nonlinear and Linear Periods
ALSI	10	10	10	NAS, NIK, HANG, FTSE, VIX, BRA, MXC	BRA, MXC, FTSE
J500	12	10	8	ALSI, CI, FTSE, BRA	ALSI
J510	11	8	9	ALSI, NAS, BRA, MXC	ALSI, NAS
J520	13	9	12	ALSI, NAS, NIK, FTSE, BRA, MXC, RDY520	ALSI, MXC, RDY520
J530	8	8	6	ALSI, NAS, NIK, FTSE	ALSI, FTSE
J540	11	8	11	ALSI, FTSE, VIX, BRA, MXC	ALSI
J550	12	9	11	ALSI, NAS, NIK, FTSE, VIX, BRA, MXC	ALSI, NAS
J560	11	9	8	ALSI, NAS, NIK, FTSE, BRA, MXC	ALSI, NAS, MXC
J580	8	7	10	ALSI, NAS, NIK, FTSE, VIX, BRA, MXC	ALSI, FTSE
J590	14	9	6	ALSI, NAS, FTSE, MXC	ALSI, NAS

Table 5.13 shows the number of variables selected by the Boruta algorithm as features for the respective models. The fact that the number and identity of the variables changes across industries and periods could be considered as evidence for the rejection of *Hypothesis 3*. This evidence needs to be balanced against the common variables observed in Table 5.13, which perhaps imply a possible underlying common model. Aside from the ALSI, the international stock indices and the VIX appear to be selected regularly. The fundamental question is therefore whether these common variables proxy independently for underlying factors, or are simply redundant variables. In the ALSI models it would appear that these variables proxy for the global economic state affecting the

ALSI. The insinuation here is that the effects of these factors may be fully captured by the inclusion of an ALSI variable in the industry index models, and hence at least some of these international indices are likely to be redundant. It is perhaps unsurprising given the local optimisation that occurs in the formation of decision trees that a certain degree of variable redundancy occurs.

Given the paucity in comparable training period metrics for neural networks it is difficult to definitively say whether a particular index is better modelled by a linear or a GRNN model. It is perhaps even more difficult to say if the GRNN models are well specified since there is no neural network equivalent to the coefficients in a linear regression. However, the high correlations of the fitted and actual values in the GRNN models points to the models having an appropriate fit. The analysis of model fit through a Sign test, distributional tests and the scrutiny of residuals provides few instances in which the model fit can be rejected outright. This intimates insufficient evidence to reject *Hypothesis 1.2,* on balance. Test period analysis may shed light on the robustness of the linear and GRNN models and thus point to their respective viability.

5.3 TEST PERIOD PERFORMANCE

The models estimated over the training sets of data are used to predict the test period returns. The values for the returns of the indices provided by the models represent point predictions since they assume the actual values are unknown, but because they are dependent on the contemporaneous data from the explanatory variables (which is known) they could equally be called estimates. They are referred to as predictions in this section to differentiate them from the training period estimates. The test period performance of the models is evaluated by the traditional prediction measures: Mean Absolute Percentage Error (MAPE), Mean Absolute Error, (MAE) and Root Mean Squared Error (RMSE). These measures evaluate the errors of point predictions of the response variables. The precision with which MAPE, MAE and RMSE values represent the prediction error is dependent on the number of predictions. As the test periods comprised a limited number of point predictions (eighteen) some variance in the accuracy of these measures is expected. MAPE in particular may suffer from scaling issues when the denominator is very small – which may be the

case for scaled monthly equity index returns. Nevertheless, collectively the prediction error measures may give a satisfactory overview of the accuracy of the predictions. In addition, to the traditional accuracy measures, Pearson correlation coefficients are computed to assess the correlation between the actual test period values and predicted test period values. A Sign test is also conducted to evaluate if the median difference between the actual and predicted values is zero. The test period in the full sample is the same as the test period in the second sub-sample allowing inferences on the effect of the length of the training period on prediction accuracy. The results from the test period prediction tests are presented in Table 5.14.

Table 5.14 Test period performance metrics for the ALSI and industry indices models.

Test period performance for the ALSI and Industry Indices Models																
		Full-period					Sub-period 1					Sub-period 2				
Index		MAPE	MAE	RMSE	CORR	SIGN	MAPE	MAE	RMSE	CORR	SIGN	MAPE	MAE	RMSE	CORR	SIGN
ALSI	LR	141.5	0.3661	0.5107	0.5528	0.7222	175.2	0.5163	0.5844	0.8227	0.3889	133.0	0.4539	0.5949	0.6429	0.6111
	NN	149.8	0.3914	0.5422	0.3876	0.4444	132.0	0.5403	0.7235	0.7488	0.1667	118.8	0.5377	0.7168	0.2925	0.5556
J500	LR	118.6	0.3994	0.5053	0.7520	0.7222	145.1	0.4663	0.6104	0.8358	0.6111	102.5	0.4295	0.5669	0.7069	0.5556
	NN	128.0	0.4842	0.6014	0.2947	0.4444	116.8	0.5141	0.6565	0.8252	0.6667	112.7	0.5215	0.7616	0.2993	0.5000
J510	LR	109.7	0.4975	0.6255	0.7737	0.5556	146.9	0.5172	0.6662	0.8873	0.2778	87.3	0.5422	0.6409	0.7942	0.6111
	NN	99.5	0.6025	0.6994	0.7021	0.4444	231.4	0.7521	1.1110	0.3907	0.4444	95.9	0.7083	0.8382	0.5763	0.5000
J520	LR	1282.4	0.5990	0.6874	0.7720	0.0556	220.2	0.4907	0.6214	0.7181	0.3333	603.8	0.7397	0.8595	0.6837	0.0556
	NN	1039.0	0.4448	0.5206	0.6674	0.2778	244.2	0.5233	0.6626	0.6870	0.2222	348.5	0.5010	0.6117	0.7717	0.2778
J530	LR	277.2	0.3434	0.4254	0.7970	0.2222	493.2	0.4808	0.6140	0.7890	0.3889	366.6	0.4549	0.5628	0.8515	0.2222
	NN	151.0	0.3752	0.4995	0.7347	0.2222	277.7	0.3800	0.4207	0.8542	0.4444	152.2	0.5050	0.6819	0.5633	0.3333
J540	LR	129.0	0.5451	0.7370	0.4507	0.2222	139.1	0.8286	1.0531	0.2513	0.5000	202.0	0.6678	0.9255	0.2967	0.2222
	NN	95.2	0.4198	0.5631	0.4025	0.6667	120.1	0.8320	0.9830	0.4768	0.2222	152.2	0.4859	0.7495	0.2838	0.3889
J550	LR	121.7	0.4557	0.5607	0.6534	0.5000	168.0	0.6441	0.8317	0.4825	0.3333	76.4	0.5349	0.6763	0.7297	0.5000
	NN	104.9	0.5132	0.6272	0.5969	0.6111	225.3	0.6327	0.7597	0.6509	0.2222	94.3	0.6695	0.8481	0.4157	0.6111
J560	LR	208.0	0.3685	0.4447	0.5413	0.4444	263.3	0.5427	0.6133	0.6204	0.6111	1743.0	0.8857	1.0925	0.0364	0.6667
	NN	190.3	0.4695	0.6504	0.3180	0.7778	199.5	0.5041	0.6114	0.5898	0.4444	1035.2	0.7088	0.8705	0.2464	0.6667
J580	LR	113.2	0.3646	0.5010	0.4095	0.4444	235.5	0.6472	0.8441	0.4548	0.4444	697.0	0.4683	0.5899	0.6317	0.3333
	NN	127.1	0.3297	0.4121	0.6767	0.4444	328.3	0.7756	0.9994	0.2900	0.5000	1252.5	0.4042	0.4994	0.7693	0.2778
J590	LR	84.8	0.1751	0.2330	0.8228	0.4444	176.3	0.6351	0.7273	0.4205	0.4444	204.4	0.8909	1.0299	0.5802	0.0556
	NN	142.2	0.3143	0.3708	0.5001	0.4444	119.9	0.4461	0.5243	0.5515	0.3889	103.3	0.5092	0.5688	0.6535	0.4444

Note: LR – linear model, NN- GRNN model.

For the ALSI, test period predictions using the linear model appear to be better than those of the neural network over the full period (Table 5.14). The values for MAPE, MAE and RMSE are

clearly lower in the linear model, but the differences in the values for the linear model and GRNN are marginal. The correlation and the ratio of correctly predicted signs are also higher for the linear model in the full period. The correlation between the actual values and the linear predictions of the ALSI are relatively high which points to fairly robust test period performance. This is especially true in the first sub-period. In the first and second sub-periods the linear predictions for the ALSI also appear to be better than the neural network predictions. In both sub-periods the MAPE is higher for the linear predictions than the neural network predictions, but the MAE, RMSE, correlation and sign ratio are all in favour of the linear predictions. Overall it can be deemed that the ALSI linear models appear to provide the better test period performance than the GRNN models.

The industry indices offer a more ambiguous picture of the linear and GRNN test period performance. In the full period the linear model predictions of the J500 are superior in all the reported measures (Table 5.14). In the sub-periods the linear model for the J500 also appears to offer predictions superior to the GRNN. Similarly, the J510 linear models also appear to better predictors of J510 returns than their GRNN equivalents (Table 5.14). Over the full period the MAPE for the linear model predictions is higher than that of the GRNN model, but the other metrics are in favour of the linear model. The linear full period J520 model does not appear to offer suitable predictions of the J520 returns. The prediction error metrics are better for the GRNN (Table 5.14), but the correlation of the predicted and actual values for the GRNN full period model is lower than that of the linear model. The prediction metrics in the J520 sub-periods offer a mixed picture with the GRNN providing better predictions in the second sub-period and the linear model providing better predictions in the first. It is difficult to determine which model provides better full period predictions from the metrics in Table 5.14, but the low RMSE and MAE, and high correlation of the predictions with the actual values implies that both full period models provide relatively good predictions. In the first sub-period the neural network model provides the better predicitons but in the second subperiod the linear model predicitons have better metrics.

The predictions of the J540 do not appear to be particularly good in any of the periods under study. The highest correlation between the actual and predicted returns in any of the J540 models is only

0.48. This is perhaps unsurprising given that the training period fit of the J540 models is unsatisfactory in most cases. If the models are not a good fit in the training period it is illogical for them to provide good test period estimates. It is also evident in the second sub-period that the linear model for the J550 reports prediction metrics that are consistently superior to those of the GRNN. This includes the lowest MAPE in the study but the MAE and RMSE are both lower for the linear model predictions, and the correlation between the actual and the predicted values is higher (Table 5.14). In the first sub-period the differences in the prediction metrics for the J560 models are too small for meaningful deductions to be made, and in the second sub-period the GRNN gives better predictions but the error rates are so high as to indicate that both sets of predictions are deficient. Indeed, the correlations between the actual and predicted values for the linear model and GRNN are only 0.04 and 0.25, respectively. The J580 full period model prediction error metrics point to the GRNN model providing a better test period prediction than the linear model for this industry index. The GRNN gives lower MAE and RMSE scores and has the higher correlation between the predicted and actual values. The J580 sub-period prediction error metrics are relatively high for both models. In the first sub-period the linear model appears to provide better predictions and in the second sub-period the GRNN appears to provide better predictions. The full period J590 linear model predictions have the lowest RMSE and MAE of any of the predictions in this study. This linear model shows a lower MAPE score than its GRNN counterpart, and the predictions show a high correlation with the actual values (0.82). In contrast to the J590 full period models, the metrics for the first and second sub-period models show the GRNN predictions to be uniformly superior by all the metrics considered.

In general, the prediction error metrics in Table 5.14 appear to be considerably high. In many cases, the correlations of the predicted values with the actual values are also high. Hence, it is possible that the high values for the error metrics could be an artefact of the small test period over which they are computed. Nevertheless, the overall impression is that both the linear models and the GRNN models do not provide robust test period estimations. There is also evidence that the accuracy of the test period predictions is period dependent. In the main, the predictions are better in the first sub-period and the linear models appear have better test period performance than the GRNN models. For the GRNN models the smoothing of the predictions by smoothing parameter,

σ, appears to result in a reduction of variance which in many cases makes the prediction of volatile returns over a short period difficult, but may lead to better longer run results.

5.3.1 DIEBOLD-MARIANO TESTS

Diebold-Mariano tests assess the loss functions of the model predictions and give a measure of the statistical significance of any differences between the linear and GRNN models. The standard Diebold-Mariano test is conservative when used on small samples. As there are only eighteen point predictions conducted here it is safe to assume that the DM statistics reported in this study are conservative. The results of Diebold-Mariano tests conducted on the linear and GRNN models are presented in Table 5.15. The table also presents Diebold-Mariano tests which evaluate whether the linear and GRNN predictions are superior to those of a naïve model.

Table 5.15 Diebold-Mariano tests for testing period performance of the indices linear and neural network models.

Diebold and Mariano (DM) Tests							
		Full period		Sub-period 1		Sub-period 2	
Index	Models	*DM* statistics	p- value	*DM* statistics	p- value	*DM* statistics	p- value
ALSI	LR vs NN	-0.6766	0.5077	-1.0880	0.2918	-1.6116	0.1255
	LR vs Naïve	-1.9435	0.0344	-2.2201	0.0202	-2.0403	0.0286
	NN vs Naïve	-1.9607	0.0333	-2.1204	0.0245	-1.8847	0.0383
J500	LR vs NN	-0.8513	0.4064	-0.0518	0.9593	-1.1734	0.2568
	LR vs Naïve	-2.0374	0.0287	-2.0077	0.0304	-2.2171	0.0203
	NN vs Naïve	-2.1578	0.0228	-2.3794	0.0147	-2.1890	0.0214
J510	LR vs NN	-0.5458	0.5923	-1.7109	0.1053	-1.8805	0.0773
	LR vs Naïve	-2.0664	0.0272	-3.5543	0.0012	-2.1054	0.0252
	NN vs Naïve	-1.9749	0.0324	-2.3086	0.0169	-1.9874	0.0316
J520	LR vs NN	1.6893	0.1094	-1.3196	0.2045	2.8383	0.0114
	LR vs Naïve	-0.8802	0.1955	-1.8473	0.0411	-0.9138	0.1868
	NN vs Naïve	-1.9420	0.0345	-1.8016	0.0447	-2.8725	0.0053
J530	LR vs NN	-0.2957	0.7711	0.1324	0.8962	-1.0238	0.3203
	LR vs Naïve	-2.4967	0.0116	-2.0447	0.0284	-2.5466	0.0104
	NN vs Naïve	-2.7737	0.0065	-2.0521	0.0280	-2.4344	0.0131
J540	LR vs NN	1.4492	0.1655	0.4973	0.6254	2.0793	0.0530
	LR vs Naïve	-1.2103	0.1214	-2.1871	0.0215	-1.0295	0.1588
	NN vs Naïve	-1.8304	0.0424	-2.7008	0.0076	-1.7216	0.0516
J550	LR vs NN	-1.0703	0.2995	0.7517	0.4625	-2.6317	0.0175
	LR vs Naïve	-3.0387	0.0037	-2.1158	0.0247	-3.0748	0.0034
	NN vs Naïve	-2.8192	0.0059	-2.2681	0.0183	-2.5688	0.0100
J560	LR vs NN	-1.7336	0.1011	0.3836	0.7060	2.3778	0.0294
	LR vs Naïve	-3.2844	0.0022	-2.5494	0.0104	-1.6635	0.0599
	NN vs Naïve	-3.4723	0.0015	-2.6306	0.0088	-2.7112	0.0074
J580	LR vs NN	1.6141	0.1249	0.1661	0.8700	1.1652	0.2600
	LR vs Naïve	-1.5207	0.0734	-2.6021	0.0093	-1.7735	0.0470
	NN vs Naïve	-1.8278	0.0426	-2.3082	0.0169	-1.9449	0.0343
J590	LR vs NN	-2.7108	0.0148	1.8803	0.0773	3.2939	0.0043
	LR vs Naïve	-3.5249	0.0013	-1.4222	0.0865	0.0792	0.5311
	NN vs Naïve	-2.5556	0.0102	-2.5984	0.0094	-2.6907	0.0077

Note: LR – linear model NN- GRNN model.

The null hypothesis for the linear and neural network prediction comparisons is that the loss functions of the linear and GRNN predictions are indistinguishable from each other. Overall, there are few instances in which this null hypothesis can be rejected. Consequently, the majority of differences in the testing period performance of the linear and GRNN models are insignificant. Exceptions occur almost exclusively in the second sub-period – which incidentally shares a testing period with the full sample. In Section 5.3 the metrics show that the GRNN gives better predictions in the second sub-period. The DM statistic for the comparison of the J520 linear and GRNN models in the second sub-period rejects the null hypothesis with a p-value of 0.01. This indicates that the GRNN predictions for the J520 in this period are significantly superior to those of the linear model. Similarly, the prediction comparisons for the J550 in the second sub-period reject the null hypothesis. In this case the prediction metrics indicate that the linear model provides better predictions. In combination, this suggests that the linear model has significantly better predictions in this period. The GRNN predictions are significantly better than the linear model predictions for the J560 in the second sub-period but the metrics indicate that both are unlikely to be good predictions. The J590 models provide the only full period prediction comparison of the linear and GRNN models for which the null hypothesis is rejected. The J590 linear model leads to significantly better predictions than the GRNN model. In the second sub-period it is the J590 GRNN model that provides statistically better predictions. These exceptions are indeed atypical. In general, there appears to be insignificant evidence to reject *Hypothesis 2*.

The linear and GRNN model predictions are generally better than naïve predictions. Table 5.13 shows that all of the GRNN model predictions and the majority of linear model predictions are better than naïve predictions at a 5% significance level. Exceptions include the linear J520 and J540 predictions for the full period and second sub-period. In the latter case, this result is expected as the training period adjusted-R^2 for these models is considerably low (0.39 in the second sub-period). The fact that the model predictions perform better than naïve predictions implies that the models hold some information that is gerneralisable to test periods.

5.4 GENERAL REGRESSION NEURAL NETWORKS WITH LINEAR THE FEATURES

The joint hypothesis proposed in *Hypothesis* 2 assesses both the appropriateness of features selected by the feature selection techniques (LARS and Random Forests), and the respective linear and GRNN techniques used to model the features. This section presents a way of (at least partially) separating out this hypothesis by building a GRNN model based on the linear feature sets selected by the LARS algorithm. The variables in each model are therefore exactly the same as those used in the OLS-regressions in Section 5.2.1. The only difference is the modelling technique employed. Table 5.16 shows the training period model fit diagnostics for the GRNN models of the ALSI.

Table 5.16 Training period diagnostics for GRNN with linear feature set.

Training period performance of GRNN with linear set						
		Model fit			Residuals	
Index	Period	CORR	SIGN	KS-test	LB(5) test	Shapiro
ALSI	f	0.8961	0.4451	0.0989	7.037	0.6836^
			(0.16)	(0.34)	(0.22)	(<0.01)
ALSI	s1	0.8819	0.4636	0.1545	3.4641	0.8515^
			(0.5)	(0.194)	(0.22)	(<0.01)
ALSI	s2	0.8594	0.4636	0.136	15.35^	0.99
			(0.5)	(0.26)	(<0.01)	(0.63)

Modelling the LARS selected variables with the GRNN as opposed to linear regression appears to lead to a marginally better training period fit. Over the full period the GRNN (with LARS selected variables) fitted values have a correlation of 0.90 with the actual values. This compares to a correlation of 0.81 for the linear (OLS-regression) model (Table 5.4). The correlations for the first and second sub-periods are also higher in the GRNN model than they are in the linear model. Consistent with the GRNN converging to the underlying regression, serial correlation evident in the second sub-period linear model is also evident in the second sub-period GRNN model in Table 5.16. The training period fit of the GRNN models with the linear variables set appear comparable to the training period fit of the GRNN with variables selected by the Random Forests. This suggests that the variables selected by LARS (which have only linear relationships with returns) are at least as good as variables selected by Random Forests (which include nonlinear interactions), when modelled using a GRNN. The test period prediction metrics in Table 5.17 confirm this general

observation. The metrics for the GRNN with LARS variables in Table 5.17, fit almost monotonically in a range with upper and lower bounds determined by the metrics from the standard linear predictions and standard GRNN predictions (with Random Forests variables) in Table 5.14.

Table 5.17 Test period performance metrics for GRNN with linear feature set.

Test period performance of GRNN with linear set						
		Full-period				
Index		MAPE	MAE	RMSE	CORR	SIGN
ALSI	NN(L)	170.9	0.4170	0.5190	0.4304	0.4444
		Sub-period 1				
ALSI	NN(L)	137.7	0.5252	0.6807	0.7701	0.1667
		Sub-period 2				
ALSI	NN(L)	111.6	0.4677	0.6108	0.5565	0.5000

Diebold-Mariano tests show that the predictions of the GRNN with LARS selected variables is indistinguishable from the linear predictions in any of the periods (Table 5.18). Together with the analysis of training period fit there is insufficient evidence to reject *Hypothesis 2.1*. Given that the only difference between the models is the modelling technique, this suggests that the modelling techniques result in models that are roughly equivalent. This is not unexpected as the GRNN converges to the underlying regression surface, and if the variables have mainly linear relationships with the ALSI, then the GRNN will approximate a linear regression. Tests in Table 5.18 also show that the GRNN model predictions (with the LARS variables) are indistinguishable from the standard GRNN predictions (with the variables obtained from Random Forests). Considered in combination with the similarities of the training period fits there is insufficient evidence to reject *Hypothesis 2.2*.

Table 5.18 Diebold-Mariano tests on the forecasts of the GRNN models with the linear feature sets.

Diebold and Mariano (DM) Tests for GRNN with linear feature set							
		Full period		Sub-period 1		Sub-period 2	
Index	Models	*DM* statistics	p- value	*DM* statistics	p- value	*DM* statistics	p- value
ALSI	LR vs NN (L)	-0.1397	0.8905	-0.8680	0.3975	-0.3322	0.7438
	NN vs NN(L)	0.4921	0.6289	0.6720	0.5106	1.5806	0.1324
	NN(L) vs Naïve	-2.0018	0.0307	-2.1926	0.0213	-2.1047	0.0253

5.5 SUMMARY OF RESULTS AND REMARKS

The primary goal of this study is to demonstrate the ability of feature selection techniques, namely LARS and Random Forests, to select appropriate variables for time-series modelling in an APT-type framework. The feature selection techniques, LARS and Random Forests, appear to select viable subsets of variables for the construction of the linear and GRNN models of the ALSI and industry indices. The fact that *Hypothesis 1* is not rejected suggests that the feature selection techniques provide, a solution to the uncertainty about the number and identity of relevant variables. However, evidence of variation in the selected variable subsets, and the fact that model performance is period dependent – leading to the rejection of *Hypothesis 3*, suggests that the solution provided by the techniques used in this study is not unequivocal. Indeed, a look at Tables 5.4 through to 5.13 is instructive. There are few common variables selected across industry indices and certainly few common variables selected across sub-periods. This observation coupled with evidence that suggests that the predictive ability of models is also period dependent, may be symptomatic of an adaptive framework for returns such as the one proposed by Lo (2004).

The models show a large degree of variation in test period performance but in general the error metrics for the test period performance of the models suggest that they are not particularly robust. Diebold-Mariano tests show that in most cases the predictions of the linear models cannot be differentiated from those of the GRNN models. Nevertheless, the test period predictions of the linear and GRNN models are generally better than those of a naïve model. Consistent with the idea that a GRNN converges to the underlying regression, an ALSI GRNN model trained with the LARS variable sets is shown to have a good fit in the training period and perform better than a naïve model in predicting ALSI returns in the test period. A comparison of GRNN ALSI models trained with the LARS and Random Forests respectively, shows the models to have similar training period fit and test period predictions indistinguishable from each other.

Of the modelling ideals highlighted in Chapter 1, model appropriateness (as measured by fit) is emphasised as the most important. On this count, the models based on the variables selected by

LARS and Random Forests perform relatively well. Unlike techniques such as Principal Component Analysis, using the LARS algorithm allows the variables to retain their economic meaning. The resultant linear models are readily interpretable. This is also true of the Random Forests but unfortunately nonlinear models, and in particular neural networks, are rather more opaque modelling methods, providing no observable relationship metrics like coefficients. The robustness of the models is tested in a test period independent of the period in which the models are trained, and while the error metrics are anecdotally high, both the linear and nonlinear models generally performed better than naïve models over this period. Parsimony is a relative term in effect. The modelling of asset returns is undoubtedly a complex problem requiring a complex solution. Nonetheless, other empirical studies on macroeconomic multifactor models in South Africa have pointed to the ability of a two-factor model to adequately model returns (van Rensburg, 2000; van Rensburg & Slaney, 1997). By those standards the models presented in this study are certainly not parsimonious.

6. CONCLUSION

The literature on asset pricing models is constantly growing with models incorporating a plethora of effects and pricing considerations. Of the many asset pricing theories proposed, the APT largely endures. In part, this is because of its lack of assumptions and open ended approach to asset pricing. The very components which have ensured its endurance also hamper its adoption. Several questions mire the practical use of such models. Chief amongst them is that of variable selection. The APT is conspicuously silent on what variables should be included for its proper implementation, and whilst the extent of this problem is unique to the APT, the problem itself is emblematic of asset pricing. Even when a satisfactory theory can be devised some form of variable selection is required to identify suitable proxies for the hypothesised effect. A traditional approach to variable selection suggests a version of forward stepwise regression. However, as the number of possible variables increases stepwise regressions become tedious and impractical. As more data becomes easily available and more potential pricing variables emerge, a trial and error approach increasingly becomes a constraint.

Statistical techniques that harness the growing computational power available to researchers have advanced to the degree that they comprise a genre, known as data mining. From this broad spectrum of tools, a field has emerged. Data science embeds data mining tools in a framework for solving complex (often high dimensional) problems such as variables selection for modelling. This study demonstrates the effectiveness of feature selection techniques in selecting variable subsets for multifactor models of returns of the ALSI and industry indices. The variables selected by LARS and Random Forests are shown to be period and industry dependent but overall the subsets lead to viable models of returns. Test period assessment of linear models and the GRNN models suggest that in most cases there is no meaningful difference between the predictive performance of the two. The models are also shown to have significantly greater predictive performance than naïve models.

The results in this study corroborate the appropriateness of the ALSI as a pricing factor for equity returns. If one considers that industry indices are akin to industry specific weighted portfolios, then it is intuitive to understand why the CAPM is often widely used since market return is the only constant in all the industry models. However, as evidenced in the linear models, the other explanatory variables clearly provide additional explanatory power. This points to the economic and fundamental state having important effects on returns. The period dependence of the variable subsets suggests that the best descriptors of the economic state of concern to investors change over time which is consistent with an adaptive framework for returns. The Adaptive Market Hypothesis suggests one such framework. One of the few testable implications of the AMH is the presence of time-varying return predictability. This study demonstrates evidence for time-varying return predictability in JSE equity indices, and also demonstrates that the relevant variables for modelling the return generating process are time and industry dependent.

The data science tools employed in this study have shown a capacity to perform variable selection as evidenced in the literature. However, like many tools they do not entirely eliminate the problem of possible spurious relationships. Even if variable subsets can be drawn computationally the process of modelling ultimately requires one to carefully interrogate the nature of possible relationships. As the goal of this study is merely to demonstrate the usefulness of feature selection algorithms in solving the variable selection problem in asset pricing, the analysis has been somewhat cavalier in the assessment of the actual variables in the resultant models. As a result, no formal inferences are made on the exact nature of the underlying factors that macroeconomic and financial variables proxy for. Future research in this area could go farther in the applications for variables selection techniques and the analysis of the resultant models. The number of feature selection techniques available suggests that a study that purely assessed a set of feature selection techniques for financial applications may be beneficial. Regardless, the techniques used here appear to be sufficiently robust for general application. A conventional method of determining the robustness of asset pricing models is to test them on portfolios sorted by various known pricing effects (e.g. size, value, momentum etc.). Cross-sectional analysis represents the most obvious extension to this study.

This study avoids much of the theoretical debate on the origin of the relationship between pricing factors and returns. This is useful for purely objective modelling, but it falls short of explaining what truly drives returns – that is, it fails to identify why any particular factors are relevant for a particular model. The temptation to follow a purely objective modelling approach with little review can be great but the perils are made clear by the redundancy observed in many of the models – especially those selected by Random Forests. Indeed, the novelist Thomas Hardy best articulates this dilemma when he suggests demise as the inevitable fate that attends "philosophers who follow out a train of reasoning to its logical conclusion, and attempt perfectly consistent conduct in a world made up so largely of compromise" (Hardy, 1993, p. 74). As such, a judicious practitioner may wish to combine the feature selection techniques outlined in this study with a more traditional theoretical explanation. A researcher with a particular conviction and theoretical explanation for a possible asset pricing effect (e.g. mean-reversion) can construct a set of possible proxies and then use the variable selection techniques outlined in this study to hone in on the best proxy or subset of proxies. This study illustrates a general outline for the inclusion of feature selection techniques as a tool in asset pricing, but the possible applications of these data mining techniques are numerous. As the field of finance continues to adopt computational tools for modelling, data mining techniques will increasingly warrant more attention.

7. REFERENCES

Abreu, D., & Brunnermeier, M. K. (2002). Synchronization Risk and Delayed Arbitrage. *Journal of Financial Economics, 66*(2), 341-360.

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural computation, 9*(7), 1545-1588.

Armano, G., Marchesi, M., & Murru, A. (2005). A Hybrid Genetic-Neural Architecture for Stock Indexes Forecasting. *Information Sciences, 170*(1), 3-33.

Asprem, M. (1989). Stock Prices, Asset Portfolios and Macroeconomic Variables in ten European Countries. *Journal of Banking & Finance, 13*(4), 589-612.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying Stock Market Forecasting Techniques–Part II: Soft Computing Methods. *Expert Systems with Applications, 36*(3), 5932-5941.

Auret, C., & Sinclaire, R. (2006). Book-to-Market Ratio and Returns on the JSE. *Investment Analysts Journal, 63*(1), 31-38.

Banz, R. W. (1981). The Relationship between Return and Market Value of Common Stocks. *Journal of Financial Economics, 9*(1), 3-18.

Basiewicz, P., & Auret, C. (2010). Feasibility of the Fama and French Three Factor Model in Explaining Returns on the JSE. *Investment Analysts Journal, 71*(1), 13-25.

Bekaert, G., Hodrick, R. J., & Zhang, X. (2009). International Stock Return Comovements. *The Journal of Finance, 64*(6), 2591-2626.

Berry, M. A., Burmeister, E., & McElroy, M. B. (1988). Sorting Out Risks Using Known APT Factors. *Financial Analysts Journal, 44*(2), 29-42.

Bilson, C. M., Brailsford, T. J., & Hooper, V. J. (2001). Selecting Macroeconomic Variables as Explanatory Factors of Emerging Stock Market Returns. *Pacific-Basin Finance Journal, 9*(4), 401-426.

Black, F. (1986). Noise. *The Journal of Finance, 41*(3), 529-543.

Bodie, Z. (1976). Common Stocks as a Hedge Against Inflation. *The Journal of Finance, 31*(2), 459-470. doi: 10.2307/2326617

Bonilla, C. A., Romero-Meza, R., & Hinich, M. J. (2006). Episodic Nonlinearity in Latin American Stock Market Indices. *Applied Economics Letters, 13*(3), 195-199.

Breiman, L. (1996). Bagging Predictors. *Machine learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, L., & Stone, J. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Breslow, L. A., & Aha, D. W. (1997). Simplifying Decision Trees: A Survey. *The Knowledge Engineering Review, 12*(1), 1-40.

Cakmakli, C., & van Dijk, D. J. (2010). Getting the most out of Macroeconomic Information for Predicting Stock Returns and Volatility. *Tinbergen Institute Discussion Paper, 2010-115/4*.

Callen, J. L., Kwan, C. C., Yip, P. C., & Yuan, Y. (1996). Neural Network Forecasting of Quarterly Accounting Earnings. *International Journal of Forecasting, 12*(4), 475-482.

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A Comparison between Fama and French's Model and Artificial Neural Networks in Predicting the Chinese Stock Market. *Computers & Operations Research, 32*(10), 2499-2512.

Cao, Q., Parry, M. E., & Leggio, K. B. (2011). The Three-Factor Model and Artificial Neural Networks: Predicting Stock Price Movement in China. *Annals of Operations Research, 185*(1), 25-44.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance, 52*(1), 57-82.

Chasset, P.-O. (2013). GRNN: General Regression Neural Network for the Statistical Software R. Nancy, France: Independant Scientist. Retrieved from http://flow.chasset.net/r-grnn/

Chen, C. (1994). *Neural Networks for Financial Market Prediction.* Paper presented at the IEEE World Congress on Computational Intelligence, Neural Networks, 1994.

Chen, N.-F., Roll, R., & Ross, S. A. (1986). Economic Forces and the Stock Market. *Journal of Business, 59*(3), 383-403.

Chen, S.-J., & Jordan, B. D. (1993). Some Empirical Tests in the Arbitrage Pricing Theory: Macro Variables vs. Derived Factors. *Journal of Banking & Finance, 17*(1), 65-89.

Cheung, Y.-W., & Ng, L. K. (1998). International Evidence on the Stock Market and Aggregate Economic Activity. *Journal of Empirical Finance, 5*(3), 281-296.

Clare, A. D., & Thomas, S. H. (1994). Macroeconomic Factors, the APT and the UK Stockmarket. *Journal of Business Finance & Accounting, 21*(3), 309-330.

Cleveland, W. S. (2001). Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. *International statistical review, 69*(1), 21-26.

Connor, G., & Korajczyk, R. A. (1988). Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics, 21*(2), 255-289.

Coval, J. D., & Shumway, T. (2005). Do Behavioral Biases Affect Prices? *The Journal of Finance, 60*(1), 1-34.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of control, signals and systems, 2*(4), 303-314.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *The Journal of Finance, 53*(6), 1839-1885.

Dase, R., & Pawar, D. (2010). Application of Artificial Neural Network for Stock Market Predictions: A Review of Literature. *International Journal of Machine Intelligence, 2*(2), 14-17.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise Trader Risk in Financial Markets. *Journal of political Economy, 98*(4), 703-738.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1991). The Survival of Noise Traders in Financial Markets. *Journal of Business, 64*(1), 1-19.

DeBondt, W. F., & Thaler, R. (1985). Does the Stock Market Overreact? *The Journal of Finance, 40*(3), 793-805.

Deng, H., & Runger, G. (2012). *Feature Selection via Regularized Trees.* Paper presented at the The 2012 International Joint Conference on Neural Networks (IJCNN).

Deng, H., & Runger, G. (2013). Gene Selection with Guided Regularized Random Forest. *Pattern Recognition, 46*(12), 3483-3489.

Dhrymes, P. J., Friend, I., & Gultekin, N. B. (1984). A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory. *The Journal of Finance, 39*(2), 323-346. doi: 10.2307/2327863

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics, 13*(3), 253-263.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of statistics, 32*(2), 407-499.

Eilers, P. H., & Marx, B. D. (2010). Splines, Knots, and Penalties. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(6), 637-653.

Enke, D., & Thawornwong, S. (2005). The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns. *Expert Systems with Applications, 29*(4), 927-940.

Epperson, J. F. (1987). On the Runge Example. *American Mathematics Monthly, 94*(4), 329-341.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*(2), 383-417.

Fama, E. F. (1981). Stock Returns, Real Activity, Inflation, and Money. *The American economic review, 71*(4), 545-565.

Fama, E. F. (1998). Market Efficiency, Long-Term Returns, and Behavioral Finance. *Journal of Financial Economics, 49*(3), 283-306.

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance, 47*(2), 427-465.

Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics, 33*(1), 3-56.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI magazine, 17*(3), 37.

Flannery, M. J., & Protopapadakis, A. A. (2002). Macroeconomic Factors do influence Aggregate Stock Returns. *Review of Financial Studies, 15*(3), 751-782.

Forster, M. R. (2004). *Occam's Razor and the Relational Nature of Evidence*. Unpublished Manuscript.

Forsythe, G. E., Moler, C. B., & Malcolm, M. A. (1977). *Computer Methods for Mathematical Computations*. Englewood Cliffs, NJ: Prentice-Hall.

French, K. R., & Roll, R. (1986). Stock Return Variances: The Arrival of Information and the Reaction of Traders. *Journal of Financial Economics, 17*(1), 5-26.

Ground Rules for the FTSE/JSE Africa Index Series (2015).

Gehr, A. (1978). Some tests of Arbitrage Pricing Theory *Journal of Midwest Finance Association, 7*(4), 91-106.

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of Incremental Concept Formation. *Artificial intelligence, 40*(1), 11-61.

Gibbons, M. R. (1982). Multivariate Tests of Financial Models: A new approach. *Journal of Financial Economics, 10*(1), 3-27.

Gonzalo, J., & Granger, C. (1995). Estimation of Common long-Memory Components in Cointegrated Systems. *Journal of Business & Economic Statistics, 13*(1), 27-35.

Graham, B., & Dodd, D. L. F. (1934). *Security Analysis*. New York, NY, USA: McGraw-Hill

Graham, M., & Uliana, E. (2001). Evidence of a Value-Growth Phenomenon on the Johannesburg Stock Exchange. *Investment Analysts Journal, 53*(1), 7-18.

Granger, C. W., & Newbold, P. (1974). Spurious Regressions in Econometrics. *Journal of econometrics, 2*(2), 111-120.

Gray, J. B., & Fan, G. (2005). Regression Tree Analysis using TARGET. *Journal of Computational and Graphical Statistics, 14*(1), 206-218.

Grossman, S. J., & Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American economic review, 70*(1), 393-408.

Grubinger, T., Zeileis, A., & Pfeiffer, K.-P. (2011). evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R: Working Papers in Economics and Statistics.

Guyon, I., & Elisseeff, A. (2003). *The Journal of Machine Learning Research, 3*, 1157-1182.

Hamilton, J. D. (1994). *Time Series Analysis* (Vol. 2). Princeton, NJ: Princeton University Press Princeton.

Hardy, T. (1993). *Far from the Madding Crowd*. Ware, Hertfordshire: Wordsworth Editions.

Hargreaves, C., & Hao, Y. (2013). Prediction of Stock Performance Using Analytical Techniques. *Journal of Emerging Technologies in Web Intelligence, 5*(2), 136-142.

Hasbrouck, J. (1995). One Security, many Markets: Determining the Contributions to Price Discovery. *The Journal of Finance, 50*(4), 1175-1199.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Hirshleifer, D. (2001). Investor Psychology and Asset Pricing. *The Journal of Finance, 56*(4), 1533-1597.

Ho, C., & Hung, C.-H. (2009). Investor Sentiment as Conditioning Information in Asset Pricing. *Journal of Banking & Finance, 33*(5), 892-903.

Hong, H., & Stein, J. C. (1999). A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *The Journal of Finance, 54*(6), 2143-2184.

Hsieh, H.-h., & Hodnett, K. (2011). Tests of the Overreaction Hypothesis and the Timing of Mean Reversals on the JSE Securities Exchange (JSE): The Case of South Africa. *Journal of Applied Finance & Banking, 1*(1), 107-130.

Humpe, A., & Macmillan, P. (2009). Can Macroeconomic Variables Explain Long-Term Stock Market Movements? A Comparison of the US and Japan. *Applied Financial Economics, 19*(2), 111-119.

Hurtford, A. (2012). Why Parsimony? Retrieved from https://theartofmodelling.wordpress.com/2012/03/14/why-parsimony/

Jefferis, K., & Smith, G. (2005). The Changing Efficiency of African Stock Markets. *South African Journal of Economics, 73*(1), 54-67.

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance, 48*(1), 65-91.

Jegadeesh, N., & Titman, S. (2001). Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *The Journal of Finance, 56*(2), 699-720.

John, G. H., Kohavi, R., & Pfleger, K. (1994). *Irrelevant Features and the Subset Selection Problem.* Paper presented at the Machine Learning: Proceedings of the Eleventh International Conference.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica: Journal of the Econometric Society, 47*(1), 263-291.

Keerthi, S. S. (2005). *Generalized LARS as an Effective Feature Selection Tool for Text Classification with SVMs*. Paper presented at the Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany.

Kim, H.-j., & Shin, K.-s. (2007). A Hybrid Approach based on Neural Networks and Genetic Algorithms for detecting Temporal Patterns in Stock Markets. *Applied Soft Computing, 7*(2), 569-576.

Kim, J. H., Shamsuddin, A., & Lim, K.-P. (2011). Stock Return Predictability and the Adaptive Markets Hypothesis: Evidence from Century-Long US Data. *Journal of Empirical Finance, 18*(5), 868-879.

Kim, K.-j., & Han, I. (2000). Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index. *Expert Systems with Applications, 19*(2), 125-132.

Kohavi, R., & John, G. H. (1997). Wrappers for Feature Subset Selection. *Artificial intelligence, 97*(1), 273-324.

Kryzanowski, L., Galler, M., & Wright, D. W. (1993). Using Artificial Neural Networks to Pick Stocks. *Financial Analysts Journal, 49*(4), 21-27.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*: Springer.

Kulkarni, S. G., Chaudhary, A. K., Nandi, S., Tambe, S. S., & Kulkarni, B. D. (2004). Modeling and Monitoring of Batch Processes using Principal Component analysis (PCA) assisted Generalized Regression Neural Networks (GRNN). *Biochemical Engineering Journal, 18*(3), 193-210.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–A System for Feature Selection. *Fundamenta Informaticae, 101*(4), 271-285.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 1-13.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the alternative of a Unit Root: How sure are we that Economic Time Series have a Unit Root? *Journal of econometrics, 54*(1), 159-178.

Kwon, C. S., & Shin, T. S. (1999). Cointegration and causality between macroeconomic variables and stock market returns. *Global Finance Journal, 10*(1), 71-81. doi: http://dx.doi.org/10.1016/S1044-0283(99)00006-X

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian Investment, Extrapolation, and Risk. *The Journal of Finance, 49*(5), 1541-1578.

Lehmann, B. N., & Modest, D. M. (1988). The Empirical Foundations of the Arbitrage Pricing Theory. *Journal of Financial Economics, 21*(2), 213-254.

Leigh, W., Paz, M., & Purvis, R. (2002). An Analysis of a Hybrid Neural Network and Pattern Recognition Technique for Predicting Short-Term Increases in the NYSE Composite Index. *Omega, 30*(2), 69-76.

Lettau, M., & Ludvigson, S. (2001). Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance, 56*(3), 815-849.

Li, C., Bovik, A. C., & Wu, X. (2011). Blind Image Quality Assessment using a General Regression Neural Network. *IEEE Transactions on Neural Networks, 22*(5), 793-799.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics, 47*(1), 13–37.

Liu, H., & Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining* (Vol. 454). U.S.: Springer Science & Business Media.

Ljung, G. M., & Box, G. E. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika, 65*(2), 297-303.

Lo, A. W. (2004). The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *Journal of Portfolio Management, 30*(5), 15-29.

Looney, C. G. (1996). *Advances in Feedforward Neural Networks Demystifying Knowledge Acquiring Black Boxes*. Paper presented at the IEEE Transactions on Knowledge and Data Engineering.

Mallows, C. L. (1973). Some Comments on Cp. *Technometrics, 15*(4), 661-675.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance, 7*(1), 77-91.

McCann, L., & Welsch, R. E. (2007). Robust Variable Selection using Least Angle Regression and Elemental Set Sampling. *Computational Statistics & Data Analysis, 52*(1), 249-257.

Mei, A. K.-C. (2010). Applying the General Regression Neural Network to Forecast Stock Closing Price. *Journal of Statistics and Management Systems, 13*(3), 639-649.

Moolman, E., & Du Toit, C. (2005). An Econometric Model of the South African Stock Market: Economics. *South African Journal of Economic and Management Sciences, 8*(1), 77-91.

Nasseh, A., & Strauss, J. (2000). Stock Prices and Domestic and International Macroeconomic Activity: a Cointegration Approach. *The Quarterly Review of Economics and Finance, 40*(2), 229-245.

Neal, R. M. (1995). *Bayesian Learning for Neural Networks.* Unpublished doctoral dissertation. University of Toronto, Toronto, Canada.

Page, M. (1986). Empirical Testing of the Arbitrage Pricing Theory using data from the Johannesburg Stock Exchange. *South African Journal of Business Management, 17*(1), 38-42.

Page, M., & Way, C. (1992). Stock Market Overreaction: the South African Evidence. *Investment Analysts Journal, 36*(4), 35-49.

Park, J., & Sandberg, I. W. (1993). Approximation and Radial-Basis-Function Networks. *Neural computation, 5*(2), 305-316.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics, 33*(3), 1065-1076. doi: 10.2307/2237880

Patel, P. B., & Marwala, T. (2006). *Forecasting Closing Price Indices using Neural Networks.* Paper presented at the International Conference on Systems, Man and Cybernetics, 2006. SMC'06. IEEE

Popper, K. R. (1963). *Science as Falsification* . In Conjectures and Refutations. London: Routledge and Kegan Paul.

Powell, M. J. D. (1977). Restart Procedures for the Conjugate Gradient Method. *Mathematical programming, 12*(1), 241-254.

Qi, M. (1999). Nonlinear Predictability of Stock Returns using Financial and Economic Variables. *Journal of Business & Economic Statistics, 17*(4), 419-429.

Quinlan, J. R. (1993). *C4. 5: Programs for Machine Learning* (Vol. 1). San Mateo, CA: Morgan Kaufmann.

Racine, J. (2001). On the Nonlinear Predictability of Stock Returns using Financial and Economic Variables. *Journal of Business & Economic Statistics, 19*(3), 380-382.

Refenes, A., Azema-Barac, M., & Zapranis, A. (1993). *Stock Ranking: Neural Networks vs Multiple Linear Regression.* Paper presented at the IEEE International Conference on Neural Networks, 1993.

Refenes, A. N., Zapranis, A., & Francis, G. (1994). Stock Performance Modeling using Neural Networks: A Comparative Study with Regression Models. *Neural networks, 7*(2), 375-388.

Reinganum, M. R. (1981). The Arbitrage Pricing Theory: Some Empirical Results. *The Journal of Finance, 36*(2), 313-321. doi: 10.2307/2327013

Reinsch, C. H. (1967). Smoothing by Spline Functions. *Numerische mathematik, 10*(3), 177-183.

Roll, R., & Ross, S. A. (1980). An Empirical Investigation of the Arbitrage Pricing Theory. *The Journal of Finance, 35*(5), 1073-1103.

Rosenberg, B., Reid, K., & Lanstein, R. (1985). Persuasive Evidence of Market Inefficiency. *The Journal of Portfolio Management, 11*(3), 9-16.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information storage and Organization in the Brain. *Psychological review, 65*(6), 386.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of economic theory, 13*(3), 341-360.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). *Learning Representations by Back-Propagating Errors*. Cambridge, MA: The MIT Press.

Schalkoff, R. J. (1997). *Artificial Neural Networks*: McGraw-Hill Higher Education.

Schwert, G. W. (1990). Stock Returns and Real Activity: A Century of Evidence. *The Journal of Finance, 45*(4), 1237-1257.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance test for Normality (complete samples). *Biometrika, 52*(3/4), 591-611.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance, 19*(3), 425-442.

Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of economic perspectives, 17*(1), 83-104.

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics, 19*(2), 279-281.

Smith, G., Jefferis, K., & Ryoo, H.-J. (2002). African Stock Markets: Multiple Variance Ratio Tests of Random Walks. *Applied Financial Economics, 12*(7), 475-484.

Specht, D. (1990a). A General Regression Neural Network. *IEEE Transactions on Neural Networks, 2*(6), 568-576.

Specht, D. F. (1990b). Probabilistic Neural Networks. *Neural networks, 3*(1), 109-118.

Stein, J. C. (2009). Presidential address: Sophisticated Investors and Market Efficiency. *The Journal of Finance, 64*(4), 1517-1548.

Steiner, M., & Wittkemper, H.-G. (1997). Portfolio Optimization with a Neural Network Implementation of the Coherent Market Hypothesis. *European journal of operational research, 100*(1), 27-40.

Thawornwong, S., & Enke, D. (2004). The Adaptive Selection of Financial and Economic Variables for use with Artificial Neural Networks. *Neurocomputing, 56*, 205-232.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological, 58*(1), 267-288.

Tkacz, G. (2001). Neural Network Forecasting of Canadian GDP Growth. *International Journal of Forecasting, 17*(1), 57-69.

Todea, A., Ulici, M., & Silaghi, S. (2009). Adaptive Markets Hypothesis: Evidence from Asia-Pacific Financial Markets. *The Review of Finance and Banking, 1*(1), 7-13.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. doi: 10.2307/2237638

Vaga, T. (1990). The Coherent Market Hypothesis. *Financial Analysts Journal, 46*(6), 36-49.

van Rensburg, P. (1995). Macroeconomic Variables and the Johannesburg Stock Exchange: a Multi-Factor Approach. *De Ratione, 9*(2), 45-63.

van Rensburg, P. (1996). Macroeconomic Identification of the Priced APT Factors on the Johannesburg Stock Exchange. *South African Journal of Business Management, 27*(4), 104-121.

van Rensburg, P. (2000). Macroeconomic Variables and the Cross-Section of Johannesburg Stock Exchange Returns. *South African Journal of Business Management, 31*(1), 31-43.

van Rensburg, P., & Robertson, M. (2003). Style Characteristics and the Cross-Section of JSE Returns. *Investment Analysts Journal*(57), 7-15.

van Rensburg, P., & Slaney, K. (1997). Market segmentation on the Johannesburg Stock Exchange. *Journal for Studies in Economics and Econometrics, 21*(3), 1-23.

Wang, Z., & Tan, S. (2009). Identifying Idiosyncratic Stock Return Indicators from Large Financial Factor Set via Least Angle Regression. *Expert Systems with Applications, 36*(4), 8350-8355.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* (Ph.D dissertation), Harvard University.

White, H. (1988). *Economic Prediction using Neural Networks: The Case of IBM Daily Stock Returns.* Paper presented at the IEEE International Conference on Neural Networks, 1988.

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *The Journal of Machine Learning Research, 5*, 1205-1224.

Yuan, M., & Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49-67.

APPENDICES

APPENDIX 1

A: R CODE FOR SPLINE

The cubic spline interpolation technique used in this study aimed to construct a smoothed series avoiding oscillations between the defined data points. The **spline** function is now a core function in R needing no specific package for implementation. The following code details the spline technique used in this paper:

```
spliner<- function(name,n){
  #initialise object i for determining the columns to final data frame
  i<- 2;
  #use the length of the 'name' as the limits of i
  r<- length(colnames(name));
  #Create new object which specifies the 'DATE' column as a recognised date in r
  Date<-as.Date(name[["DATE"]],format="%m/%d/%Y");
  #Create new date column abd skip n period for which there are NA
  NewDate <- seq(Date[n],tail(Date,1),by="1 month");
  df <- data.frame(DATE=NewDate);
  #while there are columns left in name make a mini data frame
  #and spline out the interpolation using NewDate as x values
  while(i<=r){
    DF<- data.frame(DATE= as.Date(Date),VALUE=name[[i]]);
    #Give y values (output) from spline of the column and add it to growing df
    df<- data.frame(df, i=spline(DF, method="fmm", xout=NewDate)$y);
    i<-i+1;
  }
  colnames(df)<- colnames(name)
  return(df)
}
```

The Forsythe-Malcolm-Moler "fmm" method of computing the interpolation was used here. Readers interested in the technical differences between the different spline methods may find experimentation with the methods useful. In general it was found that the differences were slight aside from the use of hyman which is targeted at performing a monotonic implantation of the Forsythe-Malcolm-Moler technique.

B: R CODE FOR FEATURE SELECTION USING RANDOM FORESTS

As stated in Chapter 4, feature selection for the nonlinear models was computed using the Boruta algorithm which is essentially a wrapper for Random Forests. The Random Forest R

implementation **randomForest** is the underlying package. The Boruta algorithm is implemented through the **Boruta** package using the following function:

```
require(Boruta);

borutarer<- function(index){
  cat(paste(index, " Random Forest FEATURE SET\n",sep=""));
  df<-get(paste("t",index,sep="."))
  RF2<- Boruta(x=as.matrix(df[-1]),
y=as.vector(df[,1]),nodesize=7, ntrees=500);
  print("Full list")
  print(getSelectedAttributes(RF2,withTentative=T));
  print("List without tentatives");
  print(getSelectedAttributes(RF2,withTentative=F));
  return(getSelectedAttributes(RF2,withTentative=F));
}
```

The function computes cross-validation and feature selection simultaneously. The function can take values for arguments native to **randomForest** such as *nodesize* and *ntrees* (number of trees). Obviously the smaller the terminal *nodesize* the less computational time required. Conversely the larger the *ntrees* the longer the computational time required. It was observed that for a set of ~40 variables each with ~200 data points the function took about 5 minutes to run[19]. As can be expected the same ~40 variables with ~900 data points each took considerably longer. In this analysis the emphasis here was on resolving 'tentatives' and thus a large number of iterations was employed. Readers investigating a smaller number of variables or less complex relationship may likely get away with fewer iterations. Rather than building a single Random Forest further research in this area may consider running ~10-20 iterations of a single forest and then repeating to allow averaging over the a number of Random Forests.

C: R CODE FOR THE CONSTRUCTION OF THE GRNN

Many thanks to WenSui Lui (https://**statcompute**.wordpress.com) for first presenting a robust implementation of the GRNN algorithm in R and then kindly allowing the adaptation of it for the construction of GRNN's in this study. The R GRNN package **grnn** is written by Chasset (2013). The version currently available does not allow batch (dataframe) processing but rather fits the grnn cluster by cluster. This requires one to write their own function to iterate through a dataframe.

[19] Processing time is obviously CPU dependent.

Cross validation to identify the optimal value of σ also has to be computed manually. For robustness the rows in the training set are randomised and with roughly 25% of the sample being allocated randomly to cross validation and the remaining 65% left for estimation of the model. The following function was used to perform GRNN modelling:

```
require(doParallel);
require(foreach);
require(grnn);

GRNNer<- function(x,x2,x3,main){
  set.seed(100)
  rows <- sample(1:nrow(x), nrow(x) - as.integer(0.25*(nrow(x)+nrow(x2))));
  set1 <- x[rows, ];
  set2 <- x[-rows, ];

  pred_grnn <- function(x, nn){
    xlst <- split(x, 1:nrow(x))
    pred <- foreach(i = xlst, .combine = rbind) %dopar% {
      data.frame(pred = guess(nn, as.matrix(i)), i, row.names = NULL)
    }
  }

  # Find optimal sigma value for the grnn
  cv <- foreach(s = seq(0.3, 1, 0.01), .combine = rbind) %dopar% {
    grnn <- smooth(learn(set1, variable.column = 1), sigma = s)
    pred <- pred_grnn(set2[, -1], grnn)
    test.sse <- sum((set2[, 1] - pred$pred)^2)
    data.frame(s, sse = test.sse)
  }
  print(cv)
  print(best.s <- cv[cv$sse == min(cv$sse), 1])

  # Check how GRNN fits the data
  final_grnn <- smooth(learn(set1, variable.column = 1), sigma = best.s);
  pred_all <- pred_grnn(x[, -1], final_grnn);
  plot(pred_all$pred, x[,1],xlab="Model", ylab=colnames(x[1]));
  inresid<- x[,1]- pred_all$pred

  print("WAYNE'S Tests")
  plot(pred_all$pred,inresid,type='p',main=paste("Residual Plot
for",colnames(x[1]),"Linear Model")
       ,ylab="Residuals", xlab="Fitted Values");
  k<- ks.test(x[,1],pred_all$pred,alternative = "two.sided")
  print(paste("KS test stat:",k$statistic))
  print(paste("KS p-value:",k$p.value))
  print(paste("The IN Correlation:",cor(pred_all$pred,x[,1])))
  print(SIGN.test(inresid, md=0, alternative = "two.sided", conf.level = 0.95))

  print(paste("Ljung test:",Box.test(inresid, lag = 5, type = "Ljung-Box")$statistic,
          + "p-value:",Box.test(inresid, lag = 5, type = "Ljung-Box")$p.value))
  print(paste("Shapiro test:",shapiro.test(inresid)$statistic,"p-
value:",shapiro.test(inresid)$p.value))

  # Predict for forecasting period
  pred_for<- pred_grnn(x2[,-1],final_grnn);
  print(accuracy(pred_for$pred,x2[,1]));

  outresid<- x2[,1] - pred_for$pred
  print(paste("The OUt Correlation:",cor(x2[,1],pred_for$pred)))
  print(SIGN.test(outresid, md=0, alternative = "two.sided", conf.level = 0.95))

  return(resid<-x2[,1]-pred_for$pred);
}
```

APPENDIX 2

A: HISTOGRAMS AND PLOTS OF INDEX RETURNS

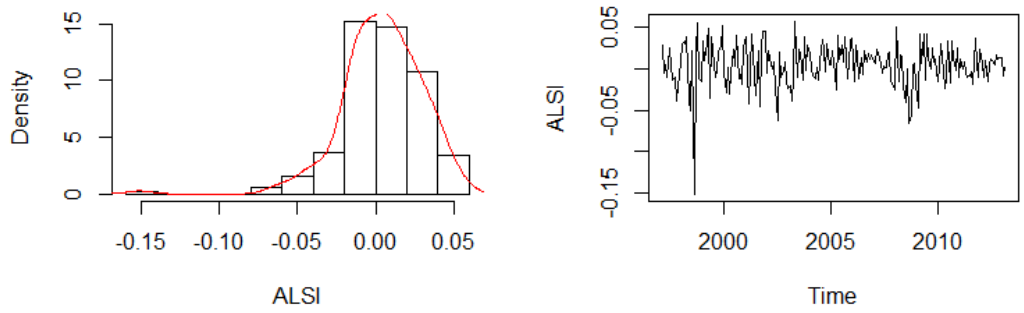


Figure 2.A1 Histogram and plot of the ALSI mean monthly return over the sample period.

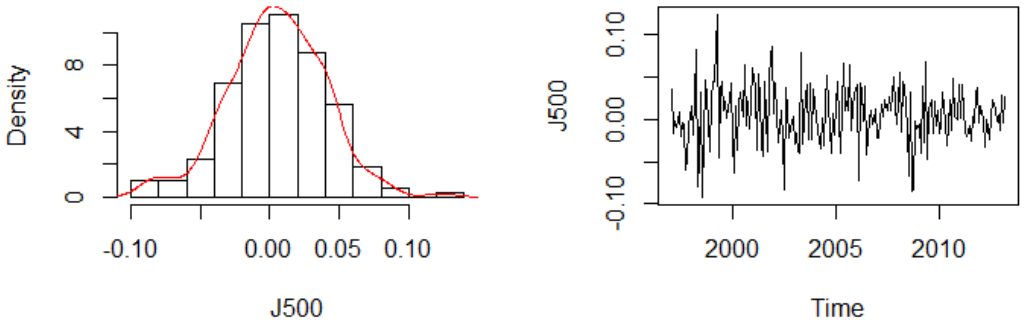


Figure 2.A2 Histogram and plot of the J500 mean monthly return over the sample period.

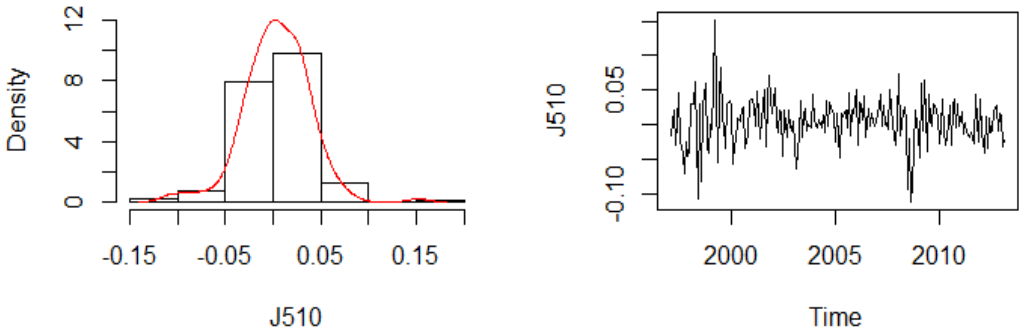


Figure 2.A3 Histogram and plot of the J510 mean monthly return over the sample period.

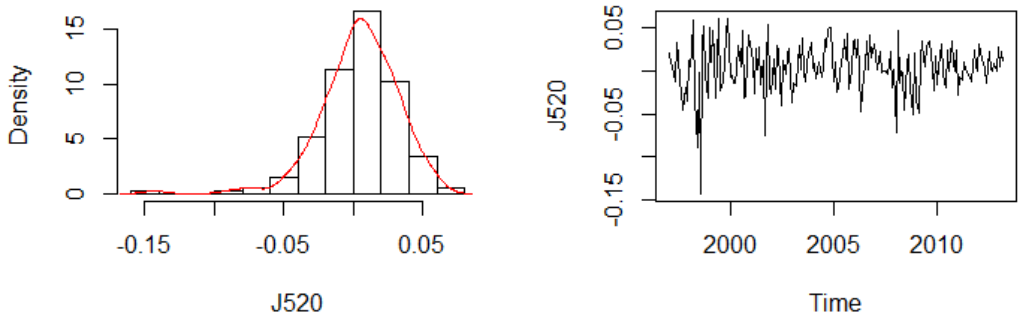


Figure 2.A4 Histogram and plot of the J520 mean monthly return over the sample period.

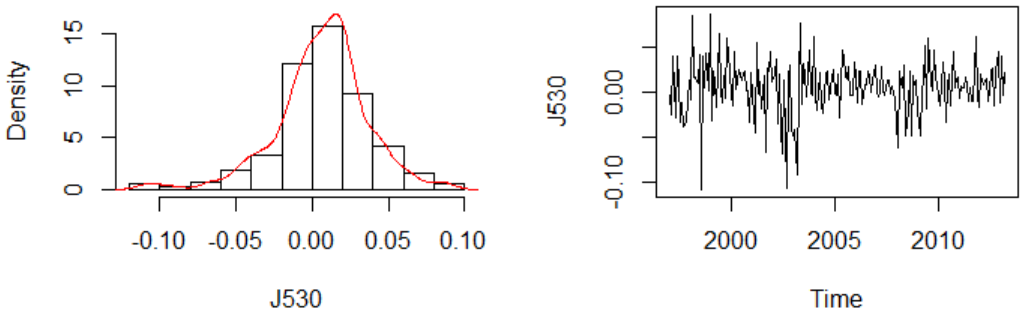


Figure 2.A5 Histogram and plot of the J530 mean monthly return over the sample period.

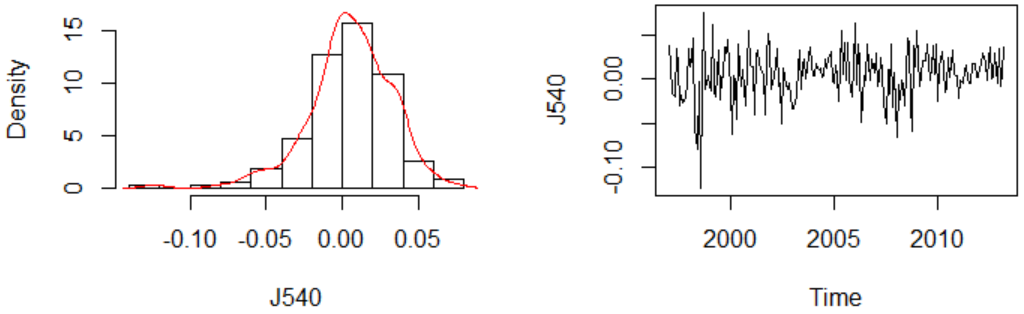


Figure 2.A6 Histogram and plot of the J540 mean monthly return over the sample period.

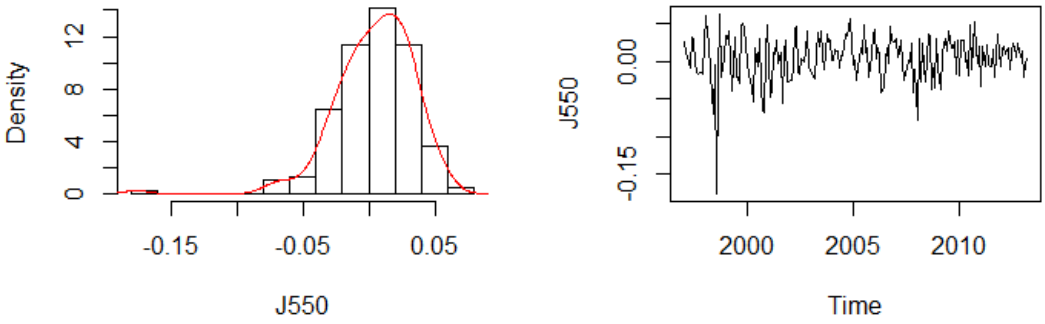


Figure 2.A7 Histogram and plot of the J550 mean monthly return over the sample period.

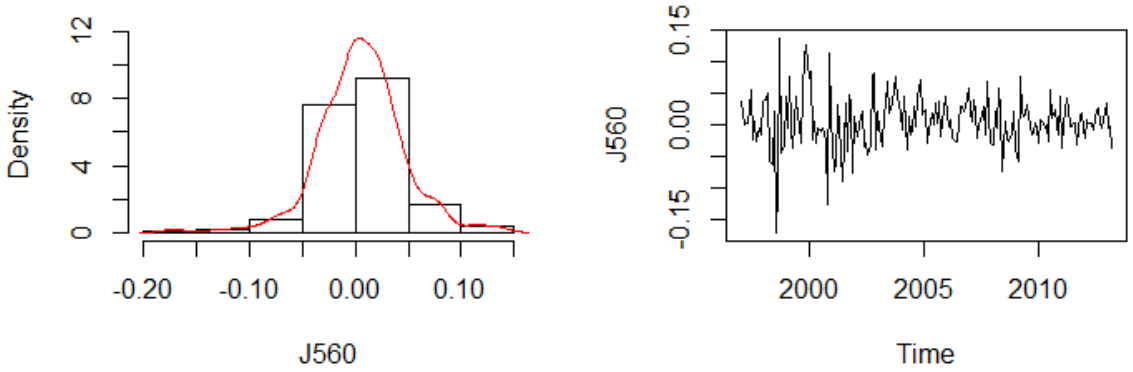


Figure 2.A8 Histogram and plot of the J560 mean monthly return over the sample period.

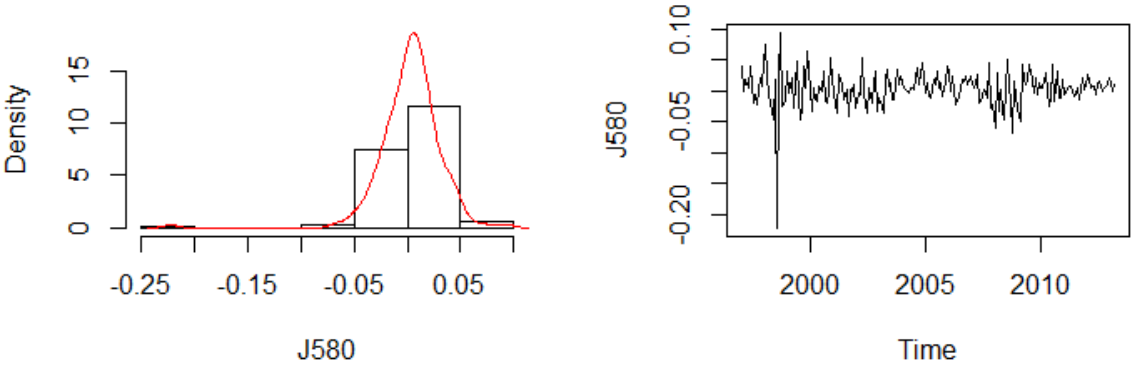


Figure 2.A9 Histogram and plot of the J580 mean monthly return over the sample period.

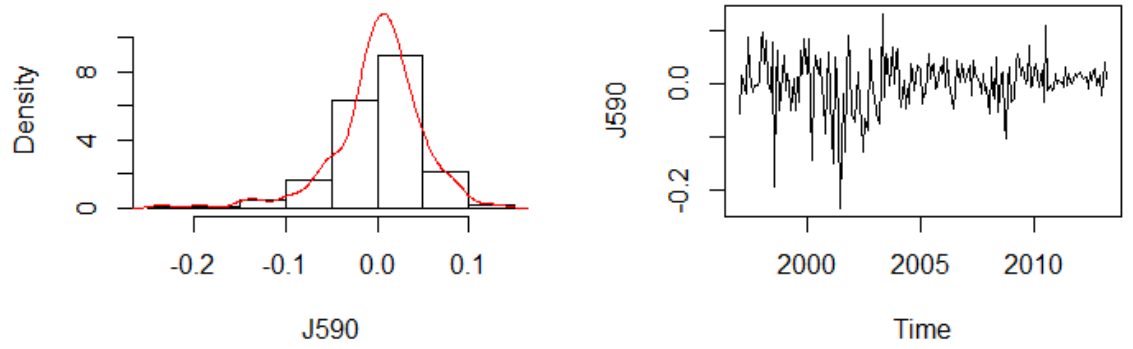


Figure 2.A10 Histogram and plot of the J590 mean monthly return over the sample period.

B: CORRELATION MATRICES FOR THE INDUSTRY INDICES VARIABLE SETS.

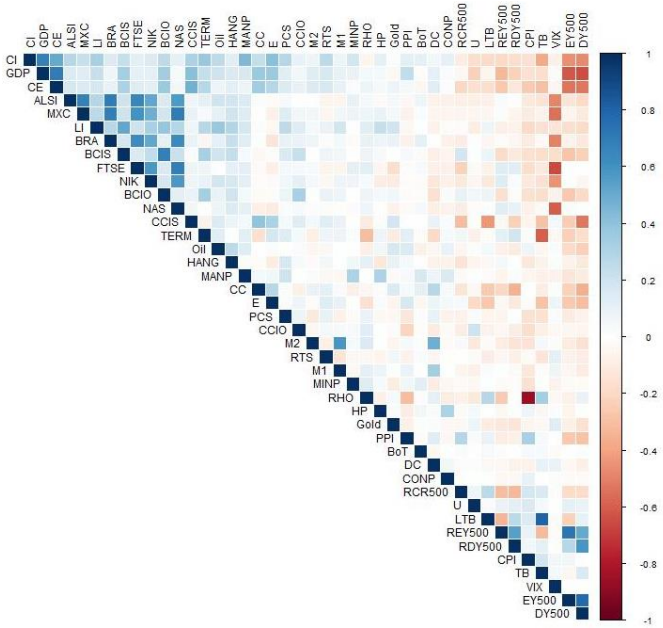


Figure 2.B1 The correlation-matrix for the J500 explanatory variables.

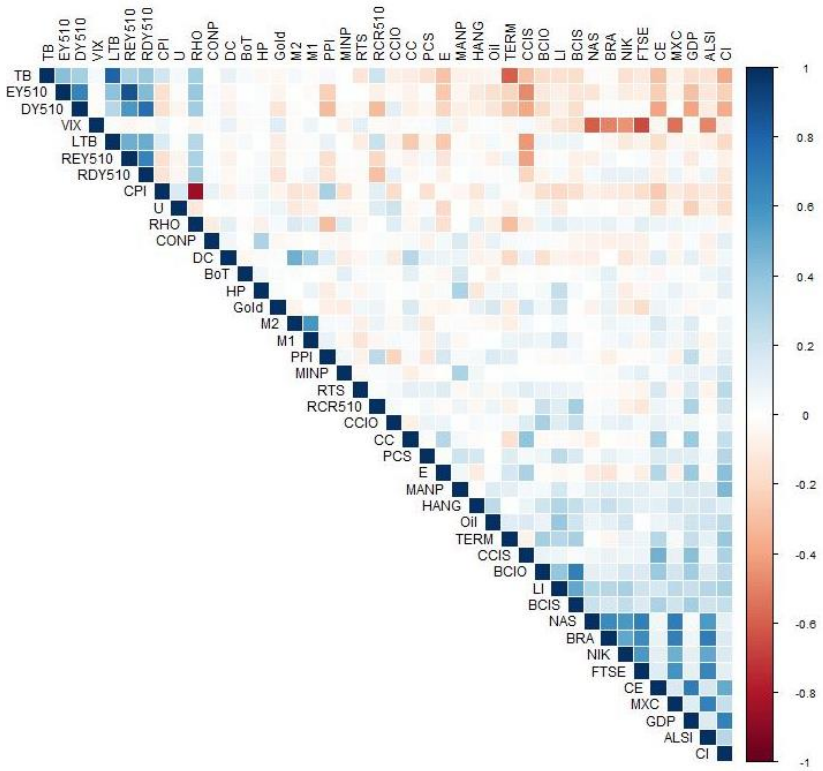


Figure 2.B2 The correlation-matrix for the J510 explanatory variables.

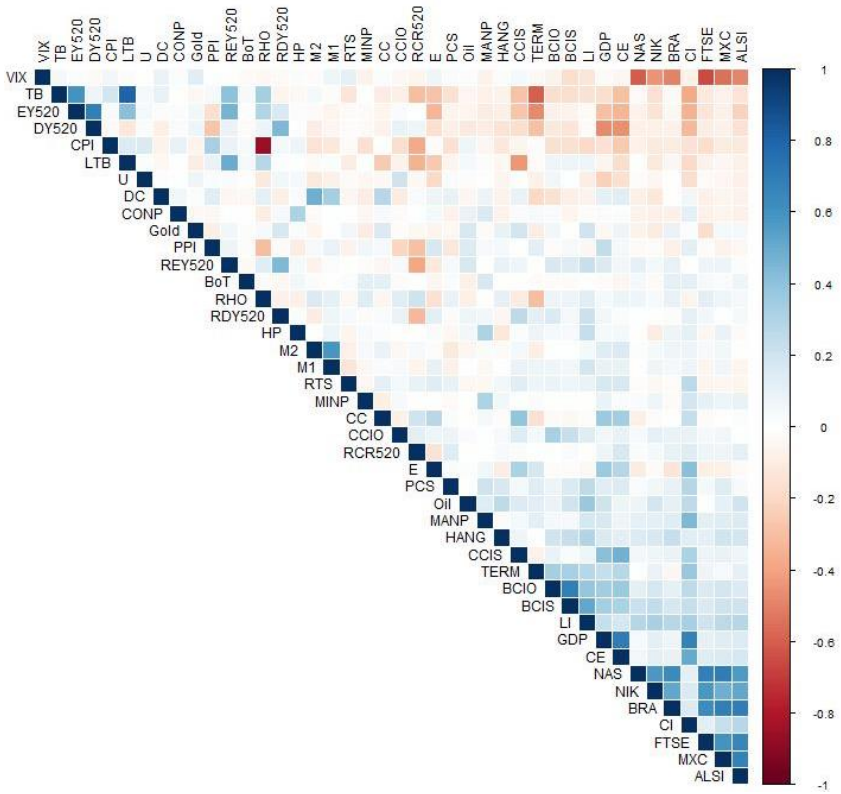


Figure 2.B3 The correlation-matrix for the J520 explanatory variables.

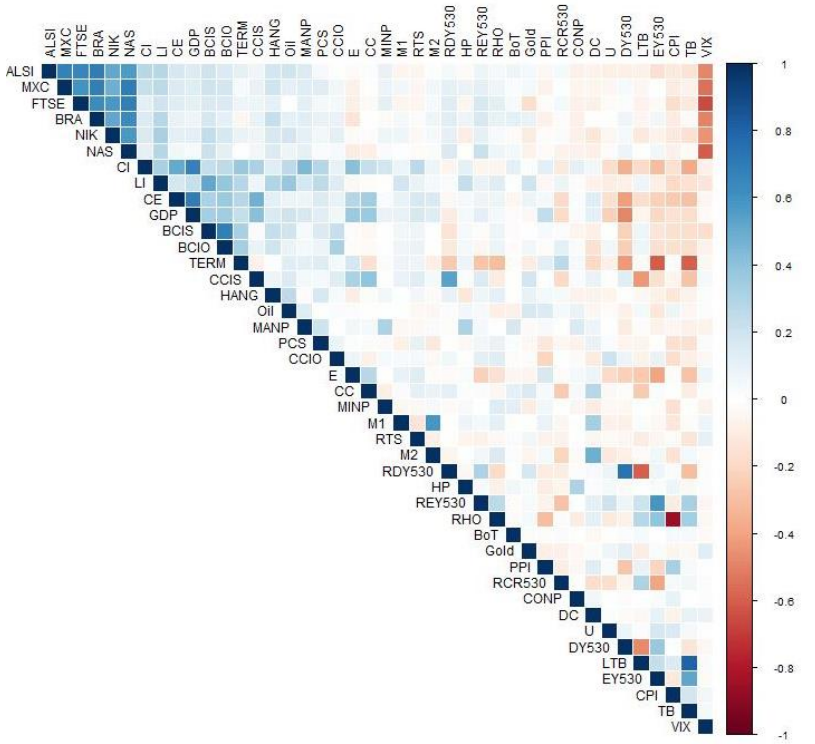


Figure 2.B4 The correlation-matrix for the J530 explanatory variables.

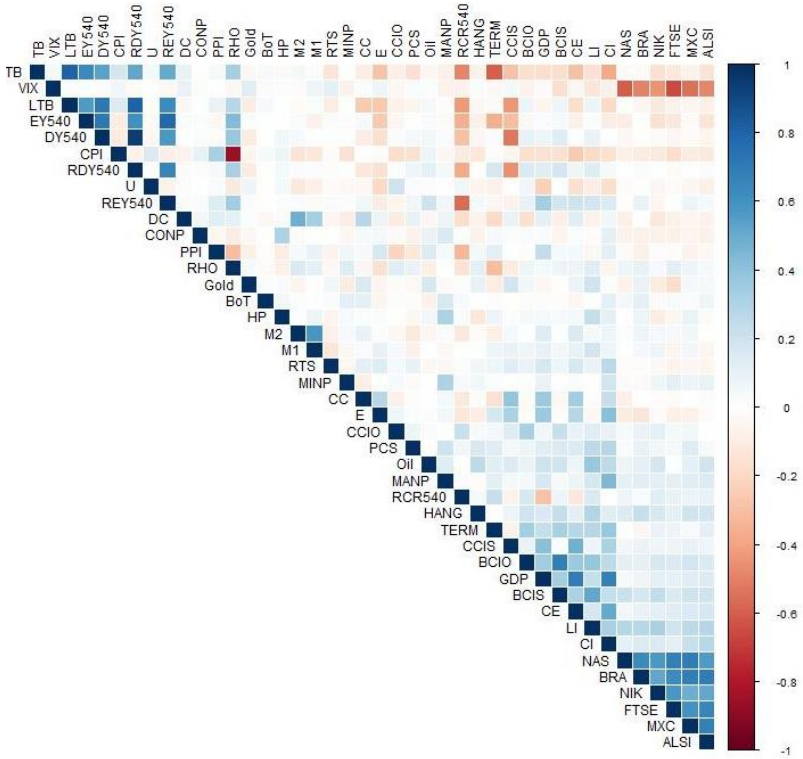


Figure 2.B5 The correlation-matrix for the J540 explanatory variables.

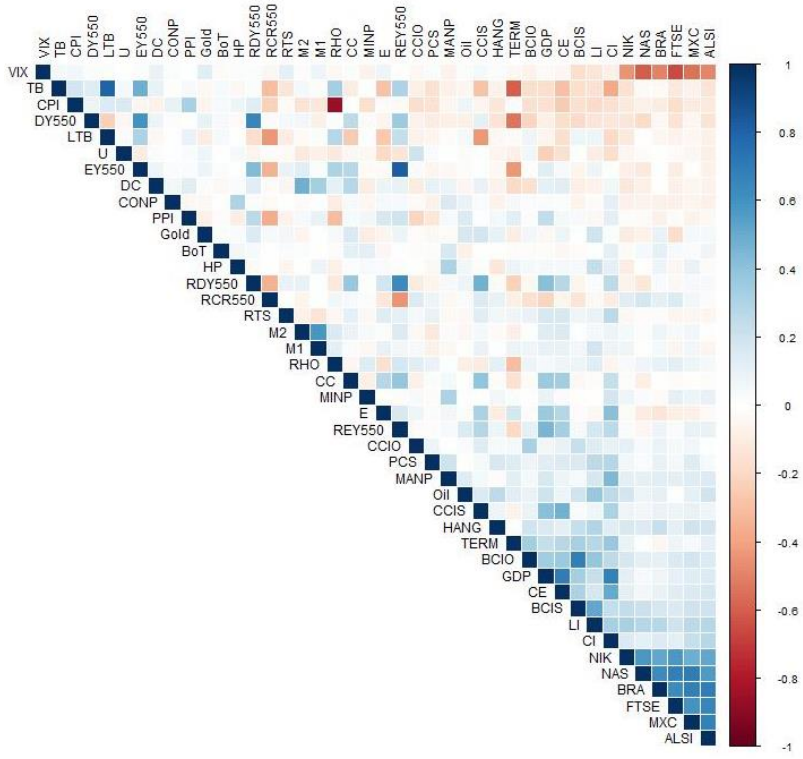


Figure 2.B6 The correlation-matrix for the J550 explanatory variables.

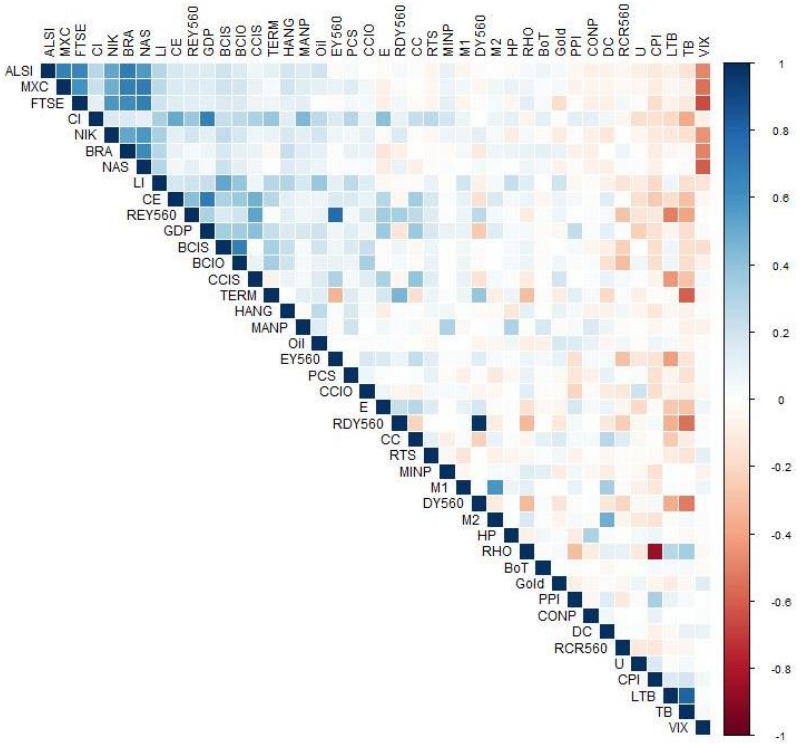


Figure 2.B7 The correlation-matrix for the J560 explanatory variables.

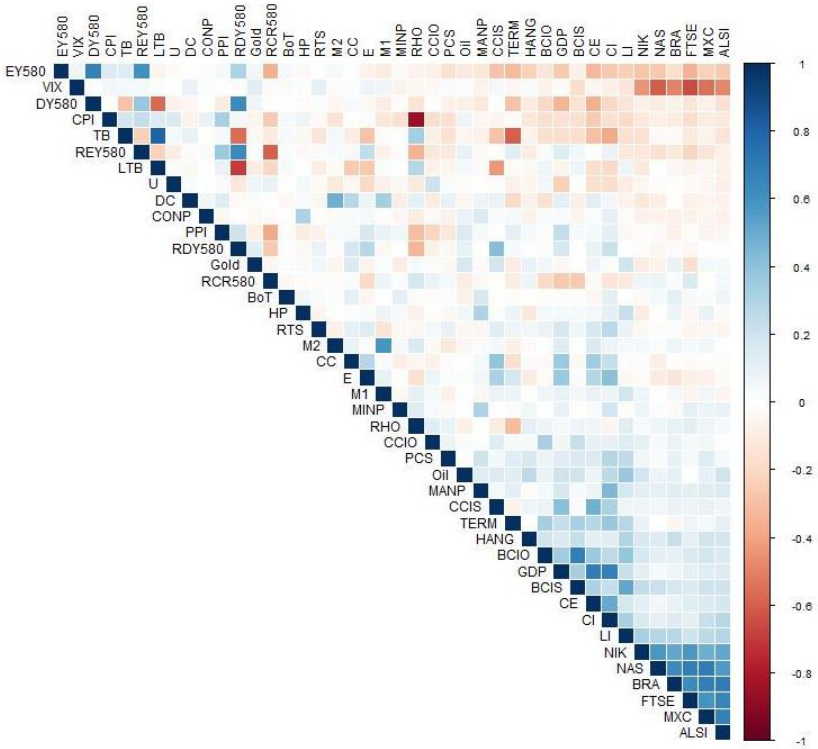


Figure 2.B8 The correlation-matrix for the J580 explanatory variables.

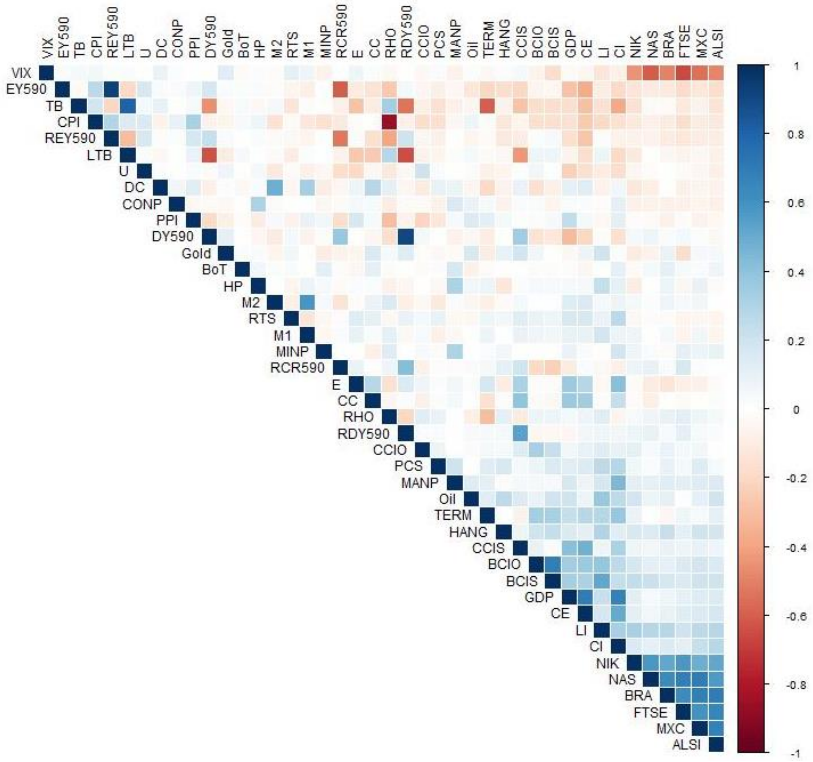


Figure 2.B9 The correlation-matrix for the J590 explanatory variables.