Department of Econometrics

Faculty of Business and Economics

Monash University

STATISTICAL CLUSTERING OF U.S. STOCK DATA VIA THE GENERALISED STYLE CLASSIFICATION ALGORITHM

By Woon Weng Wong

Joint Honours in Econometrics and Economics

Bachelor of Commerce (Econometrics)

Bachelor of Economics (Economics)

Supervisors: **Associate Professor Paul Lajbcygier, PhD**
Doctor Lee Gordon-Brown, PhD

A PhD thesis submitted
in fulfilment of
the requirements for the
Doctor of Philosophy

August 2011

Copyright Notices

Notice 1

Notice 2

Contents

Acknowledgments

First and foremost, I would like to thank my family: Mom, Dad, Su Leng, David and Lucy for your unconditional love and support. To my Ah Yee, Sar Yee, Bee Yee and Ee Vonn, I thank you as well for your encouragement (and delicious food!). To my Tai Pak, I hope I have made you proud. This achievement is for our family.

To my very dear friends: Jeremy, Alex, Leong, Tuan and Vincent who have endured hours of my incessant babbling. Thank you for your encouragement throughout the numerous lows; and your celebrations throughout the many small victories. I believe you can measure a man's wealth by his friends. Thank you for making me the richest man alive.

To Rohan Fletcher, Steve Gardner and Vincent Chau, thank you for your professional dedication. This research would not have been possible without your generous assistance.

I would also like to thank Professor Richard Heaney and Professor Robert Faff for your comments and invaluable feedback.

Lastly but by no means least, I would like to extend my heartfelt gratitude to my supervisors: Associate Professor Paul Lajbcygier and Doctor Lee Gordon-Brown. Thank you for always keeping your door open and sharing your wisdom and experience with me. I am eternally grateful for your guidance through these years. I will miss our free and liberal exchange of ideas, which is the true joy of research. Thank you not only for your professional advice but more importantly your personal friendship. While the PhD is an immense intellectual challenge, it is an even greater emotional and psychological one. Thank you for being a friend first and a supervisor second. I look forward to many more years of successful collaboration.

Statement of Authorship

Except where reference is made in the text of the PhD Thesis, this thesis contains no materials published elsewhere or extracted in whole or in part from a thesis or report presented by me for another degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any other degree or diploma in any other tertiary institution.

Woon Weng Wong

Abstract

This study explores the creation of homogeneous groups of stock based on returns. Currently no such classification scheme exists and industry classification schemes are used instead. These schemes do not make groupings based on return and so there is a fundamental mismatch between the way these groupings are made and their ultimate use in the literature. Such homogeneous returns groupings can be used to create a returns based classification scheme, which have the potential to improve various applications such as the identification of control firms for benchmarking purposes; and can lead to improved industry cost of capital estimates. To create these homogenous return groups, an innovative statistical clustering method known as the Generalised Style Classification (GSC) algorithm and an objective method for determining the optimal number of clusters known as the Gap statistic test is used.

The results indicate that the GSC can successfully create a returns based industry classification scheme; and that these GSC industry clusters are superior to current industry classification schemes at explaining the cross section of stock returns both in and out of sample. Further tests indicate that the GSC is superior at partitioning risky assets into separate risk classes while minimising returns variation within each risk class, which are the conditions necessary for improving industry cost of capital estimates.

Ideologically, this research has wider implications for the theory of asset pricing. The current dominant paradigm suggests that returns can be explained by exposure to generic risk factors however such studies rely on arbitrary partitioning of the data and this practice may lead to a number of econometric issues including truncation and selection bias, loss in power of statistical tests and data snooping bias. Contrary and less widely accepted studies have suggested that returns can be explained by industry factors. This study finds evidence of the latter. This indicates that the impact of industry effects on the returns generating process must be reconsidered. Methodology wise, the approach used to arrive at the results in this thesis, does not rely on data partitioning thereby making it immune to the aforementioned econometric issues.

The GSC represents an exciting addition to the finance literature. This research demonstrates how it may be applied to stock returns with great success. A vast number of applications in the finance literature will benefit from the homogeneous returns groupings created via the

GSC and researchers wishing to adopt a data driven, objective means of creating such groupings must consider the use of the GSC.

Chapter 1

1 Introduction

In finance, a security is a negotiable instrument representing financial value. Securities may take the form of stocks, which represent ownership of an entity; bonds, which represent debt agreements with a lender; or derivatives, which represent rights to ownership of an entity. Companies issue securities as a means to raise capital (financial wealth) for expansion. Potential investors may choose to hold securities such as stock for a number of reasons. One common reason is to make a capital gain, which occurs when an asset is purchased at a relatively low price and resold in subsequent periods at a higher price. The purchase and sale of stocks takes place at a stock exchange which provides facilities for such transactions.

The largest stock exchange in the world is the New York Stock Exchange (NYSE) which has a market capitalisation of USD11.838 trillion, followed by the Tokyo Stock Exchange (USD3.306 trillion) and the NASDAQ (USD3.239 trillion) as at 31 December 2009[1].

Given the immense amount of financial wealth stored in the stock market, academics and practitioners have sought to better understand the mechanism(s) by which a stock derives its value. One common approach would be to compare the stock performance of a sample of firms which operate under certain market conditions or possess some characteristic of interest against a sample of matched firms (often referred to as control firms) which do not operate under those conditions or possess the characteristic of interest but are otherwise identical or homogeneous to the original sample. Any differences in stock performance may therefore be attributable to the phenomenon under investigation.

A more direct approach would be to estimate the relationship between a stock's return[2] and variables that are hypothesized to have an impact. In finance theory, this relationship is expressed as the risk-return tradeoff. This principle states that in order for an investment such as a stock to earn higher returns, investors must be willing to accept higher risk. The exact source(s) of this risk however remains a heavily debated area in the literature. King (1966) saw this risk as coming from a firm's industry of operation. Using factor analysis, he found that stocks within an industry tended to

[1] Source: World Federation of Exchanges – Statistics/Monthly, http://www.world-exchanges.org/statistics/ytd-monthly, accessed 21-07-2010
[2] Calculated as the growth or decline in asset prices relative to a base period

have greater co-movement in returns[3] than stocks between industries indicating that industry effects are correlated to individual stock returns.

To perform such analyses however, researchers must first find a way to create homogenous groups of stock. Why? Consider for example, the exercise of identifying control firms. In this scenario, the sample of firms under investigation (the 'treatment' group) must be matched to the control group such that both groups are similar or homogeneous with the exception of the phenomenon under study.

In King's study, stocks are divided into industries with the expectation that doing so will create homogeneous groups. But in what way are these groupings meant to be homogeneous? And how can such homogeneous groups be identified?

One common approach is to use industry classification schemes. Industry classification schemes divide firms into coarse industrial groupings that describe their basis of operation. The intuition is that firms that operate in the same industry are homogeneous in some way. In some cases, this may be an accurate assessment as different industries face opportunities and challenges specific to their sector of operation. However, in terms of stock returns, which are the relevant metric in the aforementioned applications, there is no reason to believe industrial groups formed via common industry classification schemes exhibit homogeneity. Consider for example the case of Yahoo Incorporated and the Escala Group Incorporated. Under the Standardised Industry Classification (SIC) scheme, both companies are allocated to the same industrial group, which is *Business Services*. While Yahoo is a technology services company, Escala are traders of precious metals and other collectables such as coins, stamps and wine. Figure 1-1 shows a comparison of stock returns patterns between these two companies between 1997 and 2006. The correlation coefficient between the two returns series is 0.2367 indicating a low degree of correlation.

[3] King used monthly changes in closing price, which is not technically a return since returns are calculated as a proportionate change in prices relative to some base period.

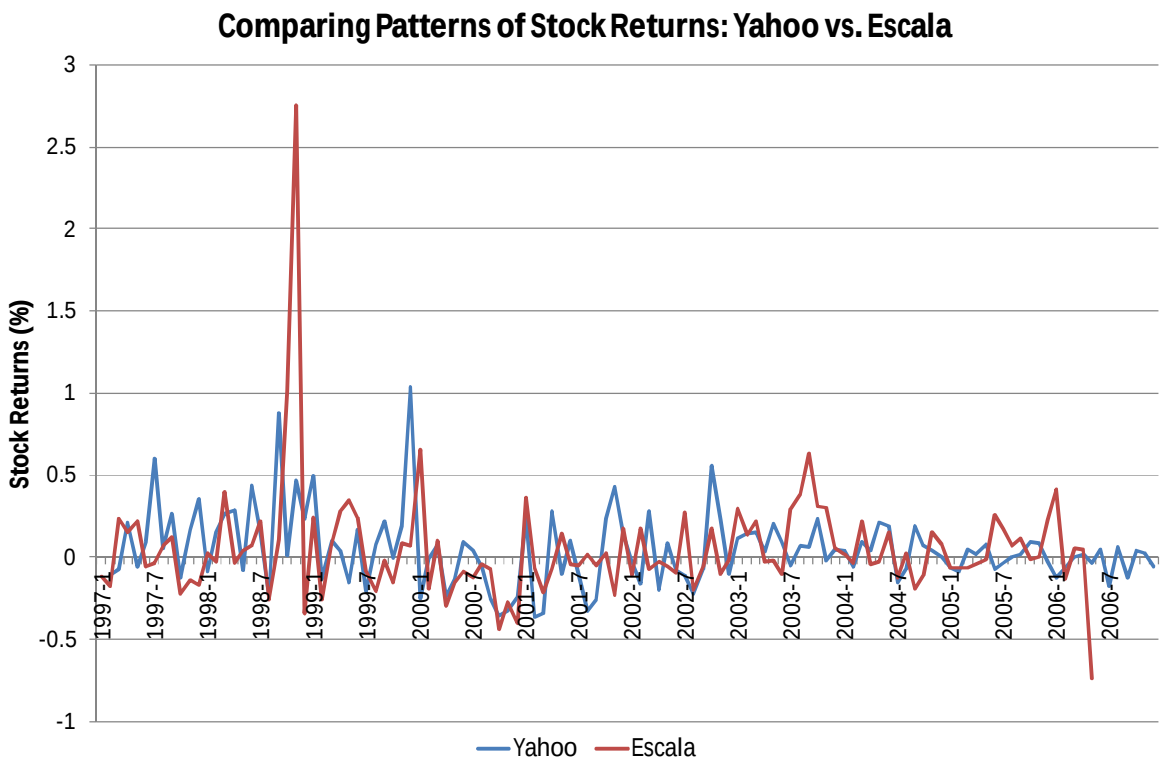


Figure 1-1 A comparison of stock returns patterns between Yahoo Incorporated and the Escala Group Incorporated between 1997 and 2006. The correlation coefficient between these two returns series is 0.2367.

Cleary, these two companies do not exhibit homogeneity in returns however under the SIC scheme, they are allocated to the same industrial group.

In the U.S., several industry classification schemes are widely available[4]. These include the:

- Standardised Industry Classification (SIC)
- North American Industry Classification System (NAICS)
- Global Industry Classification Standard (GICS)

More recently, the academic community have also developed a system of industry classification (Fama and French, 1997). For the remainder of this document, this will be referred to as the Fama French (FF) industry classification scheme.

The problem with these classification schemes is that the criteria used to determine a firm's industrial membership are unclear and at times conflicting. Furthermore, firms are not grouped on

[4] SIC, NAICS and GICS codes are available for each firm/stock in the CRSP/COMPUSTAT database which is the primary data source of this study.

the basis of returns but rather economic activity which is less useful for the aforementioned financial applications.

Therefore, what is lacking in the literature is a classification scheme that creates homogeneous groups of stocks based on similarity in returns. The objective of this research is to explore possible ways to create such homogeneous groups, which are based on similarity in returns.

Such homogeneous groupings have the potential to improve studies in areas such as the identification of control firms for benchmarking purposes; and estimation of the cost of capital.

Methodology wise, such homogeneous groupings will be formed via an innovative form of cluster analysis. Cluster analysis is a procedure that groups elements into 'clusters'. Elements within a cluster are similar to other elements in the same cluster while elements between clusters are dissimilar to each other. Since cluster analysis is applied to stock returns, the clusters formed contain stocks that have a similar pattern of returns and these patterns will be different to those of other clusters. In other words, cluster analysis has the ability to form homogeneous groups of stocks based on similarity in returns, which is exactly what the primary objective of this study is.

Pursuant to this overarching objective, there are several sub-objectives of this research. These are:

I. To explore how statistical clustering may be used to form a new industry classification scheme based on returns
II. To explicate how such clustering can lead to better industry cost of capital estimates

The findings of this research also have wider implications for asset pricing theory in general and the ideological paradigms that underpin these theories. These issues will also be discussed in later chapters.

1.1 Innovative clustering of stocks by industrial sector

Recent surveys of top financial journals indicate a number of applications where industry classification schemes have been used to form homogeneous groups of stocks. These include: identifying control firms (over half of the studies), describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate.

Consider for example, the exercise of identifying control firms. This is commonly done for benchmarking purposes. In these studies, researchers aim to isolate and estimate the effect of a particular phenomenon of interest on stock returns. To do so, researchers first identify a set of control firms, which do not possess the phenomenon of interest and then compare its stock performance against the firms that do (Ritter, 1991; Spiess and Affleck-Graves, 1994; Hendricks and Singhal, 2001). The intuition behind such studies is that if a difference exists, then the phenomenon has an effect on stock returns and this effect can be measured and estimated.

To identify control firms, these studies typically rely on an industry classification scheme such as SIC. That is, if two firms (or two groups of firms) operate in the same industry (as indicated by the industry classification scheme) and one possesses the characteristic under study while the other does not, then the latter will be used as a control firm.

The implicit assumption however is that the control firm(s) is in fact identical to the firm(s) being studied except that one possesses the characteristic of interest while the other does not. In theory, if two firms belong to the same industry, then they face the same operational environment and are thus homogeneous.

In practice however, a number of problems exist. Firstly, it is unclear how industry classification schemes make such groupings. In some cases, groupings appear to be made on the basis of similarity in production method and in others, on the basis of similarity in product use (Clarke, 1989). In either case, these groupings are not particularly useful especially when the objective of the study is to identify control firms to detect systematic differences in stock returns. There is no guarantee that if two firms belong to the same 'industry' (in terms of say production method) that they would exhibit similarity in stock returns.

The clustering procedure used in this study, known as the Generalised Style Classification (GSC) algorithm is unambiguous in the way it groups stocks. The GSC is specifically designed to group stocks on the basis of similarity in returns. Therefore, when two firms are allocated to the same cluster under GSC, they will in fact belong to the same returns class.

Chapter 6 explores how the GSC can be applied to the data to derive such clusters. Furthermore, these clusters can be profiled via existing industry classification schemes to develop an industry interpretation (and validation) of the clusters. These GSC based industry clusters may then be used as a replacement for existing industry classification schemes for applications such as identifying control firms for benchmarking purposes. The advantage of the GSC based industry clusters over existing classification schemes, is that such industry clusters would be able to more reliably identify control firms since stocks within the same cluster belong to the same returns class.

Furthermore, the GSC based industry clusters provide a new taxonomy for industries within an economy. While traditional classification schemes rely on similarities in economic or business activity to identify industrial sectors, the GSC based industry clusters identifies industrial sectors based on principles of risk and return. For example, one of the sectors identified is the **primary resources** sector. Stocks within this sector typically generate moderate levels of return but experience moderate to high levels of risk. This is expected since firms operating in this sector are exposed to a number of sources of risk including exploration which require large amounts of capital with an uncertain payoff. Other sources of risk include fluctuations in commodity prices, external risk and exchange rate risk as well as changes to the political landscape. By contrast, sectors such as the **utilities** sector generate low to moderate levels of return but experience low levels of risk. The proximity of this sector to the Government may have a stabilising effect on stock prices which results in lower returns volatility and less overall risk. Thus the GSC is able to form industry sectors within a risk/return framework as opposed to existing classification schemes which are based on economic factors.

The implications of this technology for the finance literature are profound. Researchers now have a powerful method for making industrial groupings based on risk and return. In fact, the GSC is flexible enough that it allows researchers to select the data input that best suits the objective of their research. In addition, the GSC can easily be applied to various financial ratios such as price to earnings ratios, price to book value as well as measures of operational characteristics such as scaled research and development expense (Bhojraj, Lee and Oler, 2003). It has already been successfully

applied to study mutual funds styles (Brown and Goetzman, 1996) and now is applied here to individual stock returns at the firm level. There is no reason to believe the GSC would not perform equally well when adapted to various other financial variables of interest. Such technology is revolutionary for the finance literature as researchers now have control over how homogeneous groupings of stock are formed where they previously relied on industry classification schemes to perform this function.

1.2 Towards better industry cost of capital estimates

Another contentious issue in the literature is the estimation of industry costs of capital. Firms require capital for expansion and growth. This capital is raised on one of two ways: by issuing equity or assuming debt. If the firm raises capital through assuming debt, then the cost of this capital is the interest repayments on the outstanding debt. If on the other hand, the firm raises capital by issuing equity, then the cost of this capital is the return to the investor. In general, the cost of capital is the cost to the firm of raising this capital. Industry costs of capital therefore represent the cost of capital for firms in a given industry.

Unfortunately, previous attempts at estimating industry costs of capital have led to "unavoidably imprecise" estimates (Fama and French, 1997, p.153). Once again, the reason for this imprecision is related to the industry classification scheme used to identify industries. Common approaches to estimating industry costs of capital require stocks to be first divided into distinct industrial groups. It is within these groups then, that costs of capital are estimated. However, in order for such a method to be effective, the stocks within the industry groups must be homogeneous in risk (and implicitly return). Miller and Modigliani (1966) expressed this as an equivalent risk class assumption. If stocks within an industry group are heterogeneous in risk and return, then any cost of capital estimate (which is supposed to be representative of the industry group) will be estimated with standard errors that are too large to be of practical use.

As section 1.1 explains, current industry classification schemes make groupings based on economic or business activity and not stock returns. Therefore there is no certainty that current industry classification schemes will be able to divide stocks into homogeneous risk/return groups or 'equivalent risk classes' as envisaged by Miller and Modigliani.

The Generalised Style Classification (GSC) clustering algorithm however does divide stocks into groups that are homogeneous in risk/return thereby satisfying the equivalent risk class assumption. The evidence presented in Chapter 7 conclusively proves that the GSC is superior to other industry classification schemes at forming these homogeneous groups.

Furthermore, the GSC is superior to other industry classification schemes at explaining the cross section of returns both in and out of sample. In a regression between individual returns and cluster/industry grouped means, the GSC outperforms other classification schemes by a factor of 2 to 3. In fact, the GSC explains up to 60 percent of the cross sectional variation in stock returns. This result is further evidence of the GSC's ability to form the homogeneous groupings that are missing in the literature.

This has some ideological implications as it calls into question the validity of industry cost of capital estimates in favour of grouped cost of capital where groupings (or clusters) are formed on the basis of returns which are more relevant when cost of capital is concerned.

1.3 Comparison to other cost of equity studies

The key challenge in estimating the cost of capital is estimating the cost of equity. The cost of capital is comprised of two components: the cost of debt and the cost of equity. Compared to the cost of equity, the cost of debt is relatively easy to estimate. The cost of debt is derived by taking the *risk free rate* (from a security whose duration matches the term structure of corporate debt) and adding a *default premium*. Estimating the cost of equity on the other hand is relatively more challenging as it is not clear what drives returns.

Conventional theory dictates that a relationship exists between risk and return – the so called risk-return tradeoff. That is, in order for an investor to assume higher levels of risk, they must be compensated with higher levels of return.

In the 1960s, Sharpe and Linter among others, articulated this relationship in the form of the Capital Asset Pricing model (CAPM) for which Sharpe, Markowitz and Merton Miller later won the Nobel Memorial Prize in Economics. In asset pricing, a common approach is to identify 'risk factors' and estimate the 'exposure' of an individual security to these risk factors. Under the CAPM, one risk factor is identified – the 'market'. The degree of exposure is measured by the 'market beta'. The

ideological paradigm implied by the CAPM is that an individual stock's return is affected only by the stock market as a whole. That is, the only way for a stock to perform well (systematically) is if the stock market as a whole performs well. The identification of the market as the only risk factor is one common ideological criticism of the CAPM. Put simply, there are other factors outside the market that determine the performance of a stock.

In 1966, King identified industry effects as an additional risk factor in the pricing of stocks. Using factor analysis, King concluded that up to 20 percent of the variation in a stock's return may be affected by such industry effects. However, given the technological and data constraints at the time, King limited his research to a sample of 60 to 70 stocks from 1927 to 1960. In addition, only 6 industry groups were studied. He concluded that further inquiry over a larger data set was required for better exploration of industry effects. Despite these limitations, the important ideological paradigm advanced by King's study is that a stock's return is affected by its industry of operation. Risk as a concept is multifaceted. The various sources of risk and its effects are captured by the firm's operational environment, i.e. its industry. Consider for example, firms operating in the high technology sector. Such firms are characterised by heavy investment in research and development into novel technology. The risk to the investor is that large amounts of capital are required, which may yield a potentially greater yet uncertain payoff. This type of risk is different in nature to say that of the export sector, which faces risk from a number of external sources such as exchange rate risk, changes in political and economic conditions both domestic and foreign. In this context, considering the 'market' as the only source of risk is an oversimplification as it fails to capture the various intricacies of 'idiosyncratic risk'[5] inherent to a particular industry. In some sense, King was trying to identify and estimate the effects of these industry specific risks on a stock's return.

In 1992, Fama and French refined the CAPM by introducing two additional risk factors: 'size' and 'value' risk. The *size* effect is supposed to capture the risk to an investor for investing in a small firm[6]. Here, small firms are assumed to be inherently more risky than large firms and investors must hence be paid a premium for bearing the additional risk. The *value* effect is supposed to capture the risk to an investor for investing in value stocks[7] over growth stocks. Value stocks are perceived to be underpriced as opposed to growth stocks which exhibit a history of strong earnings and high

[5] In finance theory, risk can be regarded as systematic or idiosyncratic. Systematic risk is the component of risk that is inherent to all stocks in the 'market'. Idiosyncratic risk on the other hand is the component of risk that inherent to the firm or industry. Portfolio theory suggests that idiosyncratic risk can be minimised through diversification while systematic risk cannot.
[6] Measured by its market capitalisation.
[7] Characterised by high book-to-market equity ratio

profitability. Again value stocks are assumed to be inherently more risky than growth stocks and so investors must be compensated with a premium.

There is however a number of potential econometric issues with the Fama French model. Most of these are due to the arbitrary and potentially excessive partitioning of data. In performing the Fama French regressions, stocks are partitioned into 25 portfolios ranked on *size* and *value*. The partitions are entirely arbitrary, which the authors acknowledge.

Good econometric practice involves utilising as much of the data as possible in its original form as this allows for better exploration of the true underlying structures within the data and its relationship to the explanatory variables. Data partitioning is commonly done to reduce 'noise' however when done excessively may lead to criticisms of selection bias, loss in power and data snooping. This may lead to an inflation of R^2 and findings of statistical significance where in fact none may exist (or at least to a lesser extent than that suggested by the results).

By contrast, the GSC based approach implemented in this study is immune to these econometric problems precisely because it does not rely on partitioning of the data (arbitrary or otherwise) in estimating the cross section of returns. Although in a head to head comparison of adjusted R^2, the GSC fails to outperform the impressive results achieved by Fama French, the advantage of the GSC lies in the fact that it is entirely data driven thus making it a truly objective approach that does not rely on elaborate theoretical constructs.

The GSC also possesses a number of innovative features such as a GLS correction for heteroskedasticity which reduces the distorting effects of extreme observations or periods of high volatility (such as during abnormal market events). Furthermore, the GSC clusters are easy to interpret. They simply represent the risk to a firm inherent to its industry of operation, which as previously explained can be varied and multifaceted.

As the GSC clusters can be interpreted on the basis of industry, ideologically, the GSC based approach represents a return to industry effects paradigm originated by King. The reason why King's industry effects may have failed to explain a larger proportion of returns variation is once again due to the industry classification scheme used to identify different industrial sectors. In identifying industrial sectors, King relied on the *Directory of Companies Filing Annual Reports with the Securities and Exchange Commission*, published by the Securities and Exchange Commission in 1961. Again, it

is unclear how a firm's industrial affiliation is determined under this scheme. Therefore, when dividing stocks into industrial sectors, King may not have been forming industrial groups on the basis of returns making the patterns of correlation between industry and individual returns less clear.

The GSC however does not suffer from this problem as the GSC industry clusters are formed on the basis of returns. The result is that the GSC based approach explains up to 60 percent of the cross sectional variation of a stock's return, which is a vast improvement on the 20 percent explanatory power identified by King.

In addition, the approach implemented in this study allows stocks to be reclassified dynamically between industries thereby allowing greater flexibility in how industry group average returns are matched to their respective individual returns whereas King's methodology did not allow for such reclassification.

1.4 Summary

In summary the GSC represents an exciting addition to the financial literature. One of the key objectives of this research is to explore how the GSC may be used to form homogeneous groups of stocks based on returns. By forming homogeneous groups based on returns, it provides the solution to many outstanding issues in the literature. It has the potential to improve many financial applications such as the identification of control firms for benchmarking purposes, provides taxonomy of the various industrial sectors from a risk and return perspective and improve industry cost of capital estimates.

The creation of homogeneous groups of stock has been a longstanding issue in the financial literature (see Section 2.2). Common approaches typically rely on the use of industry classification schemes however problems arise as the criteria used to form such homogeneous groupings is unclear. The GSC solves this problem as it is transparent in the way it makes groupings. Groupings via the GSC are based on returns. Furthermore, this is done entirely objectively as the algorithm iteratively makes cluster allocations (and reallocations) based on minimising the total sum of squares. Such a data driven approach completely removes subjective judgements from the grouping procedure providing researchers with a truly objective way to form homogeneous groups of stocks. In addition the GSC possesses a number of innovative features such as a GLS correction to mitigate the effects of extreme observations.

When applied to the universe of stocks, the GSC creates a new taxonomy for the various sectors of the economy that is based on a risk/return interpretation rather than systems of economic or business activity. Such taxonomy is useful for a number of financial applications such as the identification of control firms for benchmarking purposes. The use of common industry classification schemes to identify such control firms may be problematic if the objective of the study is to detect abnormalities in stock performance between control and 'treatment' groups. This is because common industry classifications schemes are not derived using the criteria that researchers require. Specifically, the industrial groupings should be formed on returns but under current industry classification schemes are formed on economic/business activity. This creates a fundamental mismatch between the way the homogeneous groupings are derived and their eventual application in the literature. The GSC solves this problem by forming industry groupings based on returns, which is exactly what is required but lacking in the literature.

The GSC can also be applied to improve industry cost of capital estimates. To accurately estimate industry costs of capital, firms belonging to the same industry must be homogeneous in risk (and implicitly return). Once again, current industry classification schemes fail to create such homogeneous groups. The GSC algorithm however succeeds where current industry classification schemes fail. By creating industry clusters which are in fact homogeneous in risk and return, industry costs of capital derived from such a classification scheme are estimated with lower error thereby making these estimates more accurate.

When compared to other approaches to asset pricing, the GSC based approach signals a return to the industry effect paradigm of King (1966). In recent times, the Fama French model has gained popularity however the model has been criticised because it makes arbitrary partitions of the data both for modelling and in constructing the explanatory variables. Despite being novel in its contribution to the literature, King's study failed to achieve a greater impact because of the limitations in the data but more importantly because of the limitations in the industry classification scheme used to divide stocks into industries, specifically in that it does not make divisions based on returns but other factors such as economic or business activity. The GSC based approach used in this study overcomes this issue since it makes industrial divisions based on returns, which is precisely what is required whether it be for specific applications such as estimating industry costs of capital or more generally in the area of asset pricing.

Such technology has the potential to contribute to key research areas in the finance literature. The evidence presented in this research illustrates how this technology can be implemented in several key areas to address specific shortcomings in the literature.

The remainder of this study is structured as follows: Chapter 2 provides a review of the literature exploring the deeper issues and key shortcomings in current approaches in a number of related areas of research. Chapter 3 describes the data used as well as the filtering conditions implemented. Chapter 4 outlines the methodology employed in this study including the overall modelling strategies, detailed description of the GSC algorithm as well as the GAP statistic test, which form the key driving technology behind this research. Chapter 5 deals with issues relating to data selection such as determining the appropriate data period and modelling intervals. Chapter 6 shows how the GSC algorithm is used to derive an industry classification scheme based on returns as opposed to current industry classification schemes which are based on economic or business activity. Chapter 7 examines the ability of the GSC based approach to explain the cross section of stock returns both in and out of sample, which is a key step in estimating the cost of capital. Furthermore, evidence is also presented to indicate the GSC's ability to form risk homogeneous groupings of stock which are necessary in estimating industry costs of capital. Chapter 8 discusses the significance and implications of these results and how they contribute to the advancement of the literature and provides conclusions to the research.

Chapter 2

2 Literature Review

2.1 Introduction

The introductory chapter outlined a number of key themes in this study. The first is the application of cluster analysis to stock returns to determine if natural groupings of stocks exist. The second is the use of cluster analysis to explain the cross section of stock returns – in essence to estimate the cost of equity. Better cost of equity estimates lead to better cost of capital estimates, which have a number of useful applications in finance such as the discount rate in NPV calculations.

This chapter explores the background of these themes and their development and current state of understanding in the literature.

2.2 Homogeneous return groups

Capital market research often requires firms to be divided into homogeneous groups. Currently, researchers rely on industry classification schemes such as SIC, NAICS and GICS[8] to produce such groupings.

Kahle and Walking (1996) find that at least 81 articles published in top financial journals[9] between 1992 and 1995 employ some form of industrial classification. Furthermore, in a survey of seven major accounting and finance journals from 2000 to 2001, Bhojraj, Lee and Oler (2003) find over 116 studies use some sort of industry classification in their research design. Among these studies, the most common uses for forming industrial groupings include: identifying control firms (over half), describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate.

The effectiveness of these studies however depends largely on the ability of the industry classification schemes to create homogeneous groups of firms. But in what way are these firms

[8] SIC is an abbreviation for the Standard Industry Classification scheme
NAICS is an abbreviation for the North American Industry Classification System
GICS is an abbreviation for the Global Industry Classification Standard
[9] These include: the Journal of Financial Economics, the Journal of Finance, the Journal of Financial and Quantitative Analysis, the Review of Financial Studies and the Journal of Business

expected to be homogeneous? One perspective would be to group firms based on similarity in production methods. Another would be to group firms based on similarity in output. In fact, the SIC scheme does both. Consider for example, *Mobile Homes* (SIC: 2451) and *Prefabricated Wood Buildings* (SIC: 2452) which are similar in product use. Both are allocated to the same SIC Division (*Manufacturing*) and Major Group (*Lumber and Wood Products*). Firms belonging to this group are therefore considered to be 'homogeneous'. In terms of product use, they are. Consider on the other hand, *Elevators and Moving Stairways* (SIC: 3534) and *Conveyors and Conveying Equipment* (SIC: 3535). Again, both are allocated to the same SIC Division (*Manufacturing*) and Major Group (*Industrial and Commercial Machinery and Computer Equipment*). And once again, firms belonging to this group are therefore considered to be 'homogeneous'. However, in terms of product use, they are not – one is used in commercial multistorey buildings, the other in industrial manufacturing. They are however homogeneous in production method.

The criteria therefore for making SIC groupings is not clear. In some cases, groupings appear to be made on consideration of production methods and in some others on product uses. Therefore, when two firms are placed into the same SIC industrial group, it is not immediately clear what characteristics if any, they share in common. The problem with these industry classifications is that they are constructed on the basis of similarity in economic activity and "intended to aid economic and marketing analysis and do not directly address the concerns of investors" (Chan, Lakonishok and Swaminathan, 2007, p.58). There is evidence in the literature to suggest that current industry classification systems do not even group firms according to economic activity well. (Clarke, 1989; Fertuck, 1975).

One measure of economic relatedness between firms that is useful to financial researchers is the extent to which returns are correlated. From a financial perspective, it matters less that two firms utilise similar production methods or have similar product uses but rather whether they share the same returns pattern. Consider for example, the exercise of identifying control firms. This is commonly done for benchmarking purposes to evaluate the performance of a firm against a group of its peers. In Ritter (1991) for example, the author compares the underpricing of IPOs against a sample of matched firms in the same industry across time. In Spiess and Affleck-Graves (1994), the authors find that firms making seasoned equity offerings substantially underperformed a sample of matched firms from the same industry and of similar size that did not issue equity. Furthermore, in Hendricks and Singhal (2001) the authors compare stock performance for firms implementing TQM policies against matched firms in the same industry. In all such studies, SIC codes are used to identify

control firms which are then matched against particular firms within the same industry, which exhibit some phenomenon of interest.

However, as previously mentioned, SIC codes create groupings based on similarity in production methods or product use. There is no guarantee (or even expectation) that two firms which share similar production methods or product use would have similar returns patterns so under the SIC scheme a reliable benchmark cannot be obtained. However, as Bhojraj, Lee and Oler (2003) point out, "the popularity of SIC codes is attributable to the absence of a superior, widely available alternative".

It is in this context that cluster analysis (in particular the GSC clustering algorithm) has a role to play. By grouping firms based on similarity in returns patterns, it provides the required 'superior alternative'. Clusters are formed in order to maximise within cluster homogeneity (in returns) while maximising between cluster heterogeneity. In this way, better benchmarks will be identified since firms that belong in the same cluster will have a similar returns profile and this profile will be distinct to firms from other clusters. Once a suitable benchmark has been identified, individual stock performance may then be evaluated against its peers (comprised of other stocks in the same cluster, or simply the overall cluster performance). Such an application is not that different to that of Brown and Goetzmann (1997) who originated the GSC algorithm in order to describe mutual fund styles and obtain better benchmarks to evaluate fund performance.

Brown and Goetzmann (1997) were motivated in part by the problem of mutual fund style self misclassification. It is suggested that such misclassification occurs as fund managers (who are compensated based on performance) are evaluated against other funds in the same style category. A favourable evaluation may be achieved by switching the fund's self reported style (which may be different to its actual style) to alter its benchmarks thus maximising *ex post* relative performance. Through the GSC, Brown et al were able to identify 8 distinct mutual funds styles from the myriad of self reported styles, which were subsequently used for benchmarking and find evidence of extensive self misclassification by fund managers.

A natural extension of this research would be to apply the GSC in same way to the universe of stocks and determine if distinct categories (or 'styles' in the Brown et al context) of stocks can be identified. Given the work of King (1966), it may also be natural to characterise these categories according to

industry. Such categorisations would form the basis for improving future research for example in the area of identifying control firms for benchmarking or describing industrial structure.

2.2.1 Cluster analysis

Data summarisation techniques such as cluster and factor analysis are commonly applied across a broad range of business related disciplines such as economics, finance and marketing (Hair, Black, Babin and Anderson, 2010). In finance, such techniques have been used to summarise the universe of stocks/firms into a more manageable and interpretable set of assets/entities (see Dimson and Mussavian (1999) for a detailed history). As Jensen (1971) explains:

"*Whenever a vast amount of data must be evaluated, there is a need to categorise the data for ease of comprehension ... In business, we see numerous classifications of business firms (or operating units within firms) into industries, regions, risk classes, etc.*"

Gupta and Huefner (1972) for example examine financial ratios at a macro level for broad industry classes and find a correspondence between the accounting data and broad industry classes. To do so, they first use cluster analysis to group firms based on financial ratios before performing correspondence tests between their clusters and similar groups formed along industry lines.

Jensen (1971) examines several clustering strategies to obtain better (objective) classification of securities and presents these as alternatives to subjective approaches of classification based on 'technical' or 'fundamentalist' processes. Jensen (1971) suggests that such objective classification have a number of useful applications in particular as they apply to investment decisions.

In more closely related research, Elton and Gruber (1971) argue that industry classifications are not ideal when forming homogenous firm risk classes. They question whether industry based groupings "employed by almost all authors" are valid and claim to possess a "host of evidence that grouping by industries is not particularly suitable for most of purposes for which it is employed" (p.434). In response to this problem, they propose an alternative method of forming homogenous groups using cluster analysis. Consistent with the notion of Modigliani and Miller's (1958) risk class, Elton and Gruber (1971) cluster firms together into pseudo-industries (statistically homogenous groups) using a combination of 23 financial ratios with the aim of forecasting the change in earnings over one year and test whether these forecasts are more accurate than those using standard industry groupings.

In more recent work, Ahn et al. (2005) apply a clustering strategy on returns on to form *basis assets* – the group of portfolios that dominate the *ex ante* opportunity set represented by individual assets. Hierarchical agglomerative clustering is performed using Ward's minimum variance method, which seeks to minimise the increase in the total sum of squares from the incremental combination of clusters and/or elements.

By contrast, the GSC algorithm is used in this research to form our version of basis assets (using the terminology of Ahn et al., 2005). However, unlike Ahn et al. (2005), the GSC algorithm does not require estimation of the covariance matrix of returns. Furthermore, this method has a theoretical link to the work of Modigliani and Miller (1958). Originally developed by Brown and Goetzmann (1997) to group mutual funds into styles, here it is extended to firms to identify basis assets which correspond to Modigliani and Miller's (1958) risk-equivalent classes.

Methodology wise, the various studies that utilise cluster analysis differ based on the distance measure used. In cluster analysis, clusters are formed on the basis of minimising the total sum of squares, which itself is calculated by summing the squared deviations of individual elements within a cluster from their cluster centroid across all clusters. The deviations themselves are computed using a distance measure. It is here that the GSC algorithm distinguishes itself from its peers. Under the GSC, deviations (from the cluster mean) are computed using a cluster mean which has been adjusted for heteroskedaticity via a GLS correction. If a stock (or group of stocks) exhibit heteroskedasticity (a common problem in financial time series), the GSC minimises this effect by allocating a smaller weight to these observations thus reducing the effect of outliers in the classification, thus improving the clustering. It does not appear that any other clustering methodology employed in the literature possesses such an innovative feature.

2.3 Industry Costs of Capital

The cost of capital is an integral component in many financial applications. For example, it is commonly used as a discount rate in Net Present Value (NPV) calculations, which is used in *discounted cash flow* analysis, computing the *time value of money* to appraise long term projects and capital budgeting in general. In essence, NPV calculations take a stream of future cash flows and discount them back to its present value. The Net Present Value is therefore computed as:

$$NPV = I_0 + \frac{I_1}{1+r} + \frac{I_2}{(1+r)^2} + \cdots + \frac{I_t}{(1+r)^t}$$

Where: I_0 = initial cost of project

I_t = income in period t

r = discount rate/interest rate

As a general rule, all projects which generate a positive NPV should be undertaken since the values of future cash flows exceed the immediate cost (Copeland, 2004). However, from the equation above, it is clear that the choice of discount rate plays a major role in NPV calculations as it the rate used to discount future cash flows to the present value. Note that future cash flows must be discounted since the value of a dollar tomorrow is less than the value of a dollar today as a dollar today may be invested to earn interest today while a dollar tomorrow cannot. An improperly specified discount rate can lead to misleading NPV calculations and subsequent misallocation of funds which can result in financial loss; or missed profit opportunities.

A commonly used risk adjusted, firm specific discount rate is the Weighted Average Cost of Capital (WACC) which takes into account the capital structure of the firm.

The cost of capital represents the cost to a firm of raising capital. Firms raise capital in one of two ways: through issuing stock (equity) or by assuming debt. The cost of raising this capital is the return to stockholders (cost of equity) in the case of equity or interest payments in the case of debt.

The WACC takes into account the cost of both forms of raising capital. It is weighted to take into account the relative proportions of equity or debt that a firm raises. For example, if a firm raises 90 percent of its capital through stock, then the WACC will be heavily weighted toward the cost of equity.

While the cost of debt is relatively easy to observe and measure, the same is not true for the cost of equity. Investors buy equity from firms with no guarantee of return. The cost of this equity must therefore be estimated. Estimating the cost of equity however is far more complex. If successful, it has the potential to improve WACC calculations, which in turn can improve NPV calculations (as it is used as a discount rate); and this leads to better allocation of funds and ultimately profit maximisation. This research contributes to this area by generating more accurate industry cost of capital estimates, for which no clear method currently exists.

Industry costs of capital represent the cost of capital to firms operating within an industry. As different industries are exposed to different sources of risk, their cost of capital must vary. For example, a mining firm is exposed to greater risk than say a utilities firm. Therefore, two cost of capital estimates are required – one for the mining industry and one for the utilities sector

Driven either by early works such as Miller and Modigliani or perhaps for lack of a better option, financial information vendors such as Valueline, Morningstar/Ibbotson and Bloomberg regularly publish estimates of industry cost of capital using SIC codes as delineators of industry groups despite the many documented problems in the literature associated with this industry classification scheme.

Early work by Miller and Modigliani (1966) estimate the cost of capital for the electric utility industry. However, Litzenberger and Rao (1972) argue that such estimates are only valid for industries comprised of firms which share similar risk profiles. If intraindustry risk heterogeneity is present, then estimating industry cost of capital is inappropriate for the simple fact that returns and hence costs of capital vary substantially within an industry. Statistically, such estimates would be useless since they generate unacceptably large standard errors[10]. In recent work, Fama and French (1997) attempt to estimate industry costs of equity but find these estimates to be "unavoidably imprecise"[11] (p.153). This imprecision is likely due to the choice of industry classification scheme. Specifically the industry classification scheme used by Fama and French does not create partitions useful for estimating cost of capital[12] and the estimates are therefore trying to summarise what is essentially a vastly disparate group of elements.

As Litzenberger and Rao (1972) note, for "industries where interfirm differences in operating risk are greater, the equivalent risk class assumption would be less appropriate. That is, each firm in a heterogeneous industry may be considered in a unique risk class and the concept of an industry cost of capital would be suspect." In addition, Boness and Frankfurter (1977) even find evidence of non-

[10] Meaning that one could obtain a point estimate for the cost of capital for a given industry but this point estimate would be useless since the standard error surrounding this estimate would be so large that the true cost of capital estimate could be anything from a very small to very large number thus making it practically useless for budget allocation purposes and NPV calculations.

[11] They are referring to the standard errors which exceed 3.0% per year. As Fama and French explain: "The annualized average market premium for 1963-1994 is 5.16% per year. The standard error of the premium, 2.71% implies that the one-standard-error bounds on the CAPM Cost of Equity of a project known to have a true beta equal to 1.0 are 2.45% to 7.87% per year. The two-standard-error bounds are -0.26% to 10.58%." (p.175)

[12] Since they are essentially a reclassification of SIC codes which are not particularly useful for grouping firms into returns homogeneous groups and/or explaining the cross section of expected returns in the first place.

homogeneity within the "believed-to-be most homogeneous of industries" (p.775), the electric utilities industry used by Miller Modigliani (1966). Using a randomised coefficient regression, they test for consistency in the Miller Modigliani equation and find the coefficient estimates to be inconsistent in randomised drawings from the population of electric utility firms. They warn against the "arbitrary categorization of industries" and argue that in order to obtain industry cost of capital estimates, data must first be partitioned into "homogeneous group(s)". As Boness et al (p. 786) explain: "researchers would be well advised if they first turned to methods which would provide homogeneous groupings by the evidence suggested by the data. Attempts have been made to create such homogenous groupings (Fama French, 1997) but these have so far been unsuccessful.

Therefore, it seems reasonable to argue that the arbitrary categorization of "industries" should be put to the test before it is accepted in academic research". That is, the industry classification system must be able to group firms into homogeneous risk/return classes.

Therefore in order for industry cost of capital estimates to be meaningful, industry groups must be comprised of firms that are homogeneous in risk/return. Current industry classification systems do not perform this function well. There is evidence to indicate substantial variation in firm returns both within and across industries as defined by 'traditional' classification systems such as SICs (Chan et al, 2007[13]; Rapach, Strauss, Tu and Zhou, 2010; Asness, Porter and Stevens, 2000). Their results indicate that industries such as communications, internet services, some resources and certain speciality trades exhibit much higher variability in returns than industries such as utilities, transportation, food and beverage manufacturing and certain essential services. The reason being that the latter industries represent necessities or subsistence goods and therefore less exposed to stock market fluctuations while the former industries such as resources face higher risks inherent to their operations. The resources sector for example experiences risk from a number of sources including fluctuations in demand and commodity prices as well as exploration operations. Other industries such as the internet sector were in their infancy during the modelling period which saw the founding and subsequent failure of numerous 'dot-coms'.

[13] Although the authors eventually reject a statistical clustering based approach in favour of existing industry classification schemes, our contention is that the clustering algorithm used was overly simplistic. The distance measure used was $1 - \rho_{ij}$ (where ρ_{ij} represents the returns correlation between any two firms). This approach, unlike ours, ignores (or rather condenses) much of the dynamic time varying patterns and does not have any correction for heteroskedasticity which is common in financial time series.

Therefore any attempt to estimate industry costs of capital using current industry classification systems will result in imprecise estimates. What is needed is a classification system that partitions data into homogeneous risk/returns groups. The GSC classification algorithm described in Chapter 4 does exactly that.

2.4 Cost of Equity

As Section 2.3 explains, firms raise capital in one of two ways: by issuing equity or assuming debt. However, firms cannot raise this capital at zero cost. It must pay its investors in exchange for the capital. It therefore faces costs when raising this capital.

To understand the relationship between cost of equity and stock returns, consider the following: Suppose a firm wishes to raise capital to fund its operations. It thus issues (sells) equity. In essence it is offering ownership in exchange for funds. In order for an investor to purchase this equity, they must be paid some kind of return in order to compensate them for their investment. One way for the investor to earn this return is to purchase the stock at a low price and then sell it at a higher price thus realising a capital gain. This return is thus calculated as the difference between the price at some future point and the price as some previous point expressed as percentage. For example, if a stock is worth $1 today and $1.10 tomorrow, the return is 10%. This return to the investor is the cost that the firm must 'pay' to raise the capital from selling its equity. Investors must be able to earn this return otherwise they will invest elsewhere and the firm will not be able to raise the required capital. Therefore, the cost of equity *is* the return that investors must earn on their investment.

What drives this return however is unclear. Some stocks earn a high return while others earn a low return. In finance theory, risk is hypothesized to be the key driver of return. But how does one measure risk? Or more specifically how does one measure the relationship between risk and return?

Those familiar with the finance literature will recognise the contribution of Sharpe, Linter, Markowitz and Merton Miller in the 1960s. Known as the Capital Asset Pricing Model (CAPM), this approach asserts that a stock's return is a direct function of the overall market return. In asset pricing parlance, the market is identified as a risk factor. Mathematically, the CAPM is articulated thus:

$$E(R_i) = R_f + \beta_i\left(E(R_m) - R_f\right)$$

Where $E(R_i)$ = Expected return on capital asset *i*

$E(R_m)$ = Expected market return

R_f = Risk free rate

$\beta_i = \frac{cov(R_i, R_m)}{Var(R_m)}$ = Market 'beta'

Note: the expectations are usually constructed as the geometric average of historical returns

The intuition behind the CAPM is simple. Investors need to be compensated in two ways: the first is for the time value of money and the second is for bearing the additional risk of holding the security. The time value of money is represented by the risk free rate. Since the market is identified as the only source of risk, the risk to the investor is therefore represented by excess market returns over and above the risk free rate. Of vital importance is the market beta which represents the degree of 'exposure' of the security to the market.

Consider the following example: Suppose the risk free rate is 3% and the market generates an average return of 10% over the period. For a security with a beta = 1.5, the security must generate at least $3\% + 1.5(10\% - 3\%) = 13.5\%$ otherwise it fails to generate the required return and should be avoided.

More importantly, the paradigm advanced by the CAPM is that the only risk factor relevant to the pricing of capital assets is the overall market itself. That is, the only way for an asset to earn a higher return systematically is if the market as a whole performs well. The sensitivity of a given asset to the market is measured by the market beta. The greater the beta, the more sensitive the asset is to movements in the market while the opposite is true for lower values of beta.

King (1966) originated the hypothesis that "the movement of a group of security price changes can be broken down into market and industry components". In other words, that a stock's return is affected by its industry. To test this hypothesis, factor analysis was applied to the covariance matrix of monthly changes in closing prices and the results indicate that stocks from the same industry tended to have greater co-movement than stocks from different industries. As King explains: "Few persons who are interested in the behaviour of the stock market will quarrel with the idea that the prices of stocks moving together. The very fact that we have averages of industrials, rails and

utilities, not to mention indexes founded on narrower classifications of securities implies that many investors think of stocks as falling into groups based on similarity of performance" (p. 139). While King confirmed that the majority of a stock's monthly changes in closing prices could be explained by the overall market performance, he also determined that a substantial proportion (up to 20 percent) could be attributed to industry specific effects with the remainder being driven by unexplained factors specific to the firm itself.

However, King limited his study to 6 industries (Tobacco products, Petroleum products, Metals, Railroads, Utilities and Retail stores) with approximately 10 to 11 firms within each industry resulting in a total of 63 securities being studied between 1927 and 1963; and ultimately concluded that a broader study incorporating more securities would be beneficial. This notwithstanding, the important ideological paradigm advanced by King is that industry effects can influence the pricing of securities and these effects will vary depending on the industry in question. More specifically, King hypothesized that the market itself is subject to a steady flow of information which through various mechanisms have an effect on a security's price. Some of this information is broad ranging in its effect and some others are focused only on a particular group of stocks such as those operating in given industry. As King explains:

"*The stock market is subject to a steady flow of information, much of which will have an effect on the set of anticipations that determine the price of security j. This does not mean, however that all of the information affecting j affects j only. In fact, it is intuitively appealing to think of incoming information as falling into various classes according to the scope of its effect on the market. There is some news of a monetary nature, for example, which is bound to have a market-wide impact on security price. The magnitude of impact need not, however be the same for all stocks. Other information may affect only certain subgroups of stocks, such as the news of a change in defense policy and its influence on the aircraft industry*" (p.140)

In other words, risk as a concept is multi-faceted. Some components of risk are generic in nature and have a wide impact on all securities but some are industry specific. By considering only generic factors such as the market, one ignores much of the industry specific effects that also exert an influence on an asset's price.

This study expands on the pioneering work of King in several ways. Firstly, the number of stocks (and their corresponding industries) under investigation is expanded to include the entire universe of

stocks (subject to the filtering conditions outlined in Chapter 3) and the data period is updated to include recent periods.

Methodology wise, this study implements an innovative form of cluster analysis in combination with a versatile yet highly objective way of determining the optimal number of clusters, *K*. More importantly, while King relied on an external source of industry classification, namely the SEC's *Directory of Companies Filing Annual Reports with the Securities and Exchange Commission*, this study creates its own classification scheme based on returns. As such the industry clusters used to explain the cross section of returns will be a better representation of those individual securities as they will belong to the same (homogeneous) returns class, which will itself be distinct to other classes. In other words, the clustering procedure used in this study will better create the "various classes" that King refers to.

2.4.1 Industry factors and the cross section of stock returns

In subsequent research, both Meyers (1973) and Livingston (1977) continued the work of King (1966). Both authors cite several technical shortcomings in King's study which may have led to an overstatement of industry effects. They still nonetheless arrive at the same conclusion - namely that industry effects are present and exert an influence on the co-movement of stock returns. Meyers (1973) updated his sample to include the original 60 companies studied by King as well as an additional 60 companies and extended the original data period to 1967. Similarly, Livingston's sample included 50 companies in 10 industry groups between January 1966 and June 1970. Livingston found approximately 18 percent of residual variance of returns co-movements could be attributed to industry effects (as opposed to 20 percent in King's study).

In a much larger study, Sharpe (1982) whose sample included 2,197 stocks between 1931 and 1979 found that model fit could be improved by including dividend yield, company size and bond beta in addition to a market index. Likewise Pari and Chen (1984) expanded their model to include factors such as general market indices, price volatility and interest rate risk to explain the cross section of returns. Chen, Roll and Ross (1986) include other factors such as the spread between long and short term interest rates, expected and unexpected inflation, industrial production and the spread between high and low grade bonds as part of their asset pricing tests.

Chen (1991) provided an improved framework for analysing stock returns in a wider macroeconomic context. Factors relating to macroeconomic state such as lagged production growth rates, the default risk premium, the term premium, the short term interest rate and the market dividend-price ratio were found to be important predictors of economic growth, which in turn were negatively correlated with market excess return.

Although the methodologies and findings differ somewhat in these studies, one common theme is that industry factors have been found to explain anywhere between 10 and 20 percent of the cross section of stock returns once various factors have been controlled for. Since the 1980s, researchers began applying the CAPM or variants thereof in their asset pricing tests. Some of these studies focused on various industries and found that industry effects do have a significant effect in explaining the cross section of stock returns. For example, Fama and French (1997) apply both the CAPM and their three-factor model to the cross section of stock returns across different industries and find evidence that both the market beta estimate as well as coefficient estimates of their SMB and HML factors vary substantially across industries suggesting that firms in different industries have varying degrees of exposure to various risk factors.

2.4.2 Fama French model

In 1992 and 1993 Fama and French (FF) refined the CAPM by introducing two additional risk factors: 'size' and 'value' risk. The *size* effect is meant to capture the risk to an investor for investing in a small firm[14]. Here, small firms are assumed to be inherently more risky than large firms and investors must hence be paid a premium for bearing the additional risk. The *value* effect is supposed to capture the risk to an investor for investing in value stocks over growth stocks. Value stocks are regarded to be underpriced as opposed to growth stocks which exhibit a history of strong earnings and high profitability. Again value stocks are assumed to be inherently more risky than growth stocks and so investors must be compensated with a premium.

FF suggest that the book to market equity ratio and size can be used as indicators of distress and distressed firms are more vulnerable to adverse conditions in the business environment, like changes in credit conditions thereby making them more risky.

[14] Measured by its market capitalisation.

Lakonishok, Shleifer and Vishny (1994) also suggest that low book to market equity (or growth) stocks appear more glamorous than value stocks which attracts naïve investors who push up prices and lower the expected returns of these stocks.

To represent the *size* and *value* risk, Fama-French develop the *SMB* and *HML* factors meant to mimic these dimensions of risk. To construct the *SMB* and *HML* factors, the universe of stock is first partitioned into 6 (2 x 3) portfolios along two dimensions. The first is *size* and the second is *value*. The size partition is made by dividing the stocks into small and big stocks. Stocks are ranked on their market capitalisation (measured by Market Equity, ME). Stocks below the 50th percentile are considered to be small while stocks above the 50th percentile are considered to be big.

The value partition is made by dividing the stocks into low, medium and high stocks. Stocks are ranked on their book-to-market equity, which is calculated as BE/ME. For details on how ME and BE are calculated, refer to Chapter 3.

Stocks are then ranked on their Book-to-market equity (BE/ME) ratio. Stocks falling in the bottom 30th percentile are considered to be *low*, while stocks falling in between the 30th and 70th percentile are considered to be *medium* and lastly stocks falling above the 70th percentile are considered to be *high*.

Thus 6 portfolios are formed at the intersection between the *size* and *value* breakpoints. The following diagram represents the formation of the six portfolios:

		Book to Market (BE/ME) ratio		
Size (ME)		Low (<30%)	Medium (30 – 70%)	High (>70%)
	Small (<50%)	*SL*	*SM*	*SH*
	Big (>50%)	*BL*	*BM*	*BH*

Table 2-1 Stocks are divided into 6 portfolios: SL, SM, SH, BL, BM, BH which are then used to construct the *SMB* and *HML* factors which are mimicking portfolios meant to represent *size* and *value* risk

6 portfolios are formed at the intersection of the size and value sorts. These are small-low (SL), small-medium (SM), small-high (SH), big-low (BL), big-medium (BM) and big-high (BH). Stocks are divided into one of these six portfolios. For example, stocks which are small (those below the 50th percentile of *ME*) and have low value (those below the 30th percentile of BE/ME) will be placed into the SL portfolio. Stocks which are small and medium are placed into the SM portfolio and so forth.

Value weighted returns within each portfolio are calculated as:

$$Ret^*_{ij} = w_{ij} Ret_{ij}$$

$$Ret^*_{j} = \sum_{i}^{I} w_{ij} Ret_{ij}$$

$$w_{ij} = \frac{ME_{ij}}{\sum_{i=1}^{n} ME_{ij}} \qquad \forall\, j = 1 \ldots 6$$

where Ret^*_{ij} = Value weighted return of stock *i* in portfolio *j*

Ret_{ij} = Return of stock *i* in portfolio *j*

w_{ij} = weighted market equity of stock *i* in portfolio *j*

Ret^*_{j} = Weighted average return of portfolio *j*

To construct the *SMB* factor, Fama and French state the following: "our portfolio *SMB* (small minus big), meant to mimic the risk factor in returns related to size, is the difference, each month between the simple average of the returns on the three small stock portfolios (*SL*, *SM*, *SH*) and the simple average of the returns on the three high stock portfolios (*BL*, *BM*, *BH*)."

In other words, the *SMB* factor is calculated thus:

$$SMB = \frac{(SL + SM + SH) - (BL + BM + BH)}{3}$$

To construct the *HML* factor, Fama and French state the following: "the portfolio *HML* (high minus low), meant to mimic the risk factor in returns related to book to market equity, is defined similarly. *HML* is the difference each month, between the simple average of the returns on the two high-*BE/ME* portfolios (*SH* and *BH*) and the average of the returns on the two low low-*BE/ME* portfolios (*SL* and *BL*)."

In other words, the *HML* factor is calculated thus:

$$HML = \frac{(SH + BH) - (SL + BL)}{2}$$

To test the explanatory power of the *SMB* and *HML* factors, FF further construct an additional 25 portfolios into which the returns data are placed. Again, the returns are ranked on *ME* and *BE/ME*. The stocks are divided into 5 *ME* categories (using 20^{th}, 40^{th}, 60^{th}, 80^{th} percentiles) and 5 *BE/ME* categories (using 20^{th}, 40^{th}, 60^{th}, 80^{th} percentiles). These partitions however are entirely arbitrary and the authors acknowledge this. As Fama French (1993) state: "the splits are arbitrary, however, and we have not searched over alternatives. The hope is that the tests here and in Fama French (1992) are not sensitive to these choices. We see no reason to argue that they are" (p.9)

Nevertheless, Table 2-2 indicates how the stocks are partitioned based on these 'arbitrary splits':

Size (ME)	Book to Market (BE/ME) ratio				
	<20%	20 – 40%	40 – 60%	60 – 80%	>80%
<20%	*Portfolio(1,1)*	*Portfolio(1,2)*	*Portfolio(1,3)*	*Portfolio(1,4)*	*Portfolio(1,5)*
20 – 40%	*Portfolio(2,1)*	*Portfolio(2,2)*	*Portfolio(2,3)*	*Portfolio(2,4)*	*Portfolio(2,5)*
40 – 60%	*Portfolio(3,1)*	*Portfolio(3,2)*	*Portfolio(3,3)*	*Portfolio(3,4)*	*Portfolio(3,5)*
60 – 80%	*Portfolio(4,1)*	*Portfolio(4,2)*	*Portfolio(4,3)*	*Portfolio(4,4)*	*Portfolio(4,5)*
>80%	*Portfolio(5,1)*	*Portfolio(5,2)*	*Portfolio(5,3)*	*Portfolio(5,4)*	*Portfolio(5,5)*

Table 2-2 This table represents the 5 x 5 portfolios which the data is partitioned into. Monthly average value weighted returns are calculated within each portfolio and these are modelled against the FF factors, *SMB* and *HML*

Value weighted returns are calculated within each portfolio. The following equation is estimated within each portfolio[15]:

$$R(t) - RF(t) = a + b[RM(t) - RF(t)] + sSMB(t) + hHML(t) + \varepsilon(t)$$

Where $R(t)$ = value weighted return

$RF(t)$ = risk free rate

Table 2-3 is a reproduction of the Fama French results:

Size (ME)	Book to Market (BE/ME) ratio				
	<20%	20 – 40%	40 – 60%	60 – 80%	>80%
<20%	0.94	0.96	0.97	0.97	0.96
20 – 40%	0.95	0.96	0.95	0.95	0.96
40 – 60%	0.95	0.94	0.93	0.93	0.93
60 – 80%	0.94	0.93	0.91	0.89	0.89
>80%	0.94	0.92	0.88	0.90	0.83

Table 2-3 Reproduced table of R^2 from the Fama French (1993) model.

From this, it is clear that Fama French are able to achieve incredibly high model fit however the method is not without criticism.

[15] Using the notation of Fama French (1992, 1993)

2.4.3 Interpreting the Fama French factors

Understanding the relationship between a firm's *size* and Market Equity (ME) is relatively straightforward. ME is calculated as *price* x *number of shares outstanding*. It represents the total dollar value of the firm's equity in the stock market. 'Small' firms, which have issued fewer shares will have less shares outstanding. Hence the ME for a small firm is low. By contrast, 'big' firms, which have issued more shares will have more shares outstanding. Hence the ME for a big firm is high. Small firms are inherently more risky than big firms (such as so called 'blue chips') because big firms are able to operate more profitably under varying economic conditions than small firms.

To understand the relationship between a firm's *value* and Book-to-Market Equity (BE/ME) ratio is relatively more complex. The Book Equity is the value of a firm's assets minus its liabilities. In some sense, it represents the value of a firm from an accounting perspective. Market Equity on the other hand has to do with the stock market's valuation of the firm. If the firm is expected to perform well, this will be reflected by a high share price and hence ME will be high.

Therefore, when a firm has a high BE/ME ratio, this implies that ME < BE. In other words, the market has a lower valuation of this firm's value compared to its accounting value hence the firm is undervalued, which is referred to as a *value* stock. As the firm is undervalued, investing in such value stocks represents a risk to the investor and so the investor should be compensated with greater returns via a greater risk premium.

The opposite is true when a firm has a low BE/ME ratio, which implies that ME > BE. In other words, the market has a greater valuation of this firm's value compared to its accounting value hence the firm is overvalued, which is referred to as a *growth* stock. Investing in such growth stocks represents less risk to the investor hence there is less risk premium associated with such an investment.

The SMB factor measures the spread in returns between 'small' and 'big' firms and is meant to represent the magnitude of the *size* risk. The Fama French model performs regressions between the average excess value weighted return of various portfolios ranked on size and value and the explanatory 'risk' factors which include excess market returns, SMB and HML. Small portfolios are expected to earn higher returns via higher risk premiums and this is reflected by higher coefficient estimates (meant to measure the degree of 'risk exposure'). For example, the 'small' stocks have consistently higher coefficient estimates on the SMB factor than the 'big' stocks indicating greater exposure to *size* risk.

The HML factor measures the spread in returns between the 'high' and 'low' value firms and is meant to represent the magnitude of the *value* risk. Again, high value firms (value stocks) are expected to earn higher returns via higher risk premiums and this is reflected by higher coefficient estimates. For example, the 'high value' stocks have consistently higher coefficient estimates on the HML factor than the 'low value' stocks indicating greater exposure to *value* risk (See Fama French, 1997, Table 6, p. 24).

2.4.4 Criticisms of the Fama French model

Daniel and Titman (1997) reject the FF model on the basis that there is no return premium associated with the three factors but rather that a firm's characteristics such as behavioural biases or liquidity determine its returns.

Their approach, referred to as the characteristics based model (CBA) directly links an asset's returns to its characteristics rather than a dimension of risk. Specifically, the CBA makes 3 partitions based on Size and 3 partitions based on Book to market equity forming 9 groups. A further 5 partitions are made per group based on HML factor loadings making a total of 45 partitions.

Daniel and Titman (1997) argue that if the FF model is theoretically valid, then stocks which have a low loading on *HML* are less risky and should therefore earn lower expected returns than stocks which have a high loading on *HML*. They find that within each of the nine groups, there is no discernable difference in expected returns between the five sub-groups. They cite this as evidence to reject the theoretical basis of the FF model, namely that there is no return premium associated with risk (at least those supposedly proxied by *HML*) once firm characteristics such as size and book to market equity are 'controlled'[16] for.

Furthermore, Chou Chou and Wang (2004) reproduces the Fama French model and applies it to data outside the original modelling period presented in FF's seminal work. They find that the predictive ability of the size and value effect diminishes once the model is applied to data outside FF's original data period. However, they find that size effects persist in January months while value effects persist in non-January months but are indistinguishable from zero when considered together and applied across all periods.

[16] Via arbitrary partitioning of the data

Schwert (2002) cites the insignificance of size and value effects outside of the original modelling period as evidence to suggest that market participants actively incorporate academic findings into their trading strategies thus exhausting profit opportunities and that such research findings actually make the market more efficient. In either case, there is evidence to suggest that the Fama French model no longer holds when applied to data outside of the original modelling period indicating a lack of robustness and a need for a new asset pricing paradigm.

On the other hand, Barber and Lyon (1997) adopt a different approach and find evidence that validates the FF model. The authors assert that since FF exclude financial firms from their original data that this creates a natural out of sample set of test data also referred to as a holdout sample. By applying the FF multifactor asset pricing model to financial firms, they find evidence of a relationship between returns and size and value effects thus validating the FF multifactor asset pricing model. They cite this finding as evidence to suggest that the original results presented by FF are not the result of data snooping and selection biases.

Furthermore, Davis (1997) also finds evidence that validates the FF model. Using data between 1943 and 1963 which is outside the original FF study, Davis still finds evidence of a value premium. Fama French (1998) cite this as evidence of out of sample validation of their approach.

Regardless of the out of sample validity of the FF model, the main shortcoming of the FF model is the arbitrary division of data. In principle, good econometric practice includes as much variability in the data as possible as this allows the researcher to explore the relationships within the data. In asset pricing models, partitioning is done to reduce noise but this practice may lead to a range of estimation problems including truncation and selection bias, loss in power and data snooping bias.

Truncation and selection bias is the introduction of error due to systematic differences in the characteristics between those selected and those not selected for a given study. In the Fama French context, selection bias arises because of the arbitrary partitioning of the data. Returns data are divided into 25 portfolios and the regressions are performed within each portfolio. Therefore, each portfolio is in effect isolated from the others resulting in an incomplete representation of the returns data within the portfolios. (Black, Jensen and Scholes, 1972)[17]

[17] They are referring to grouping stocks based on ranked market beta in a CAPM context but the argument of selection bias still applies.

Loss in power occurs when there is insufficient dispersion in the data to determine if a statistically significant effect exists. Again, when the data is arbitrarily partitioned and each portfolio only represents a subset of the population, there may be insufficient variation for statistical tests to have power.

Data snooping is the process of 'over fitting' the model as the data is visited more than once. In the current context, the partitioning of the data means that the same returns or more specifically the returns that share similar characteristics are used for estimation which may lead to an artificially significant result and misleading relationships. Some have likened it to fitting 'noise' to data. However, with enough mining of the data, even such noise can appear to be statistically significant.

The clustering procedure used in this study however, circumvents all these shortcomings as it does not rely on any partitioning – arbitrary or otherwise of the data. Since there is no partitioning, the clustering procedure suffers from none of the econometric problems that arise from such partitioning.

Furthermore, the clustering approach can be applied to individual stock returns at the firm level. The FF model cannot. Under the FF model, stocks are partitioned into portfolios making analysis possible only at the portfolio level. This is another key advantage of the clustering approach over the FF model. In fact, the clustering approach is flexible enough that it can be applied to data at virtually any level of aggregation. In this study, it is successfully applied to stocks at the individual level while in other studies such as Brown and Goetzman (1996), it is applied to mutual funds at the portfolio level making it an extremely versatile technique that can adapted to suit various research needs.

2.4.5 Size and value matching

In other related research, Barber and Lyon (1997) examine the empirical power and specification of test statistics in event studies to detect abnormal stock returns. They find that traditional approaches that rely on test statistics calculated using reference portfolios are misspecified and that this misspecification is due to several biases including: new listing bias, rebalancing bias and skewness bias. To correct for these biases, sample firms were matched to control firms based on size (as measured by Market Equity) and value (as measured by Book-to-Market ratio). Matching is done by selecting control firms which are simply closest to the sample firm in terms of firm characteristics.

For example, when matching on value, sample firms are simply matched to control firms which are closest in book-to-market ratio.

Despite its simplicity, the matching approach works remarkably well in circumventing most sources of bias with the exception of measurement bias. As Barber and Lyon explain:

"*The control firm approach eliminates the new listing bias (since both the sample and control firm must be listed in the identified event month), the rebalancing bias (since both the sample and control firm returns are calculated without rebalancing), and the skewness problem (since the sample and control firms are equally likely to experience large positive returns). When cumulative abnormal returns are used to detect long-run abnormal stock returns however, the measurement bias remains when the control firm approach is used.*" (p. 354)

This study does not consider the use of size or value in clustering however the GSC algorithm may be modified to use firm characteristics such as market equity or book-to-market ratio as opposed to returns to create clusters which are homogenous in size and value instead of returns. Such homogeneous size and value clusters may then be used for the purposes of matching sample to control firms or to compute benchmarks. This is outside the scope of this thesis but should be investigated in future research.

2.5 Conclusion

Throughout this research, a number of themes are explored such as clustering of stock returns and estimating the cost of capital. Implicit to the latter is estimating the cost of equity. Clustering of stock returns is performed to form homogeneous groupings of stocks. Forming such homogeneous groups has a number of applications in the literature such as: identifying control firms, describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate.

However, to make such homogeneous groupings, researchers have typically relied on industry classification schemes such as SIC, NAICS and GICS. Unfortunately, these industry classification schemes do not allocate stocks into homogeneous groups in a way that is useful for financial research. These classification schemes make groupings based on economic considerations such as similarity in production method or product use. Thus, there is a need in the literature for a

classification scheme that can make groupings in a way that is useful for financial research, for example by making groupings based on similarity in returns.

This research also explores issues relating to estimating the cost of capital – in particular industry costs of capital. The cost of capital represents the cost to a firm of raising capital. Industry costs of capital represent the cost to a firm of raising capital in a given industry. The cost of capital has two components, the cost of equity and the cost of debt. The key challenge in estimating the cost of capital however is estimating the cost of equity. Many attempts have been made in the literature to estimate industry cost of capital however these often fail because the mechanism used to delineate industries, i.e. industry classification schemes such as SIC, NAICS and GICS do not form homogeneous groups useful for estimating the cost of equity resulting in "unavoidably imprecise" estimates (Fama French, 1997, p.153). These homogeneous groups must be able to partition stocks into separate risk and return classes before meaningful industry cost of capital estimates can be made. Currently, no such risk/returns based classification system exists in the literature.

Implicit to estimating the cost of capital is estimating the cost of equity. King (1966) found initial evidence that industry effects explain up to 20 percent of the variation in stock returns. More recently however, the Fama French approach has gained popularity. This approach relies on arbitrary partitioning of the data and such partitioning may result in a range of econometric problems such as truncation and selection bias, loss in power and data snooping. Therefore, in order to derive a true representation of the returns generating process, the data from which the model is estimated should avoid any such partitioning.

These issues are all relevant in the finance literature and within each; there exists a number of opportunities for improvement. The innovative clustering technology used in this research resolves many of these problems and has the ability to advance the current state of understanding providing a way forward for better research in this area.

Chapter 3

3 Data

3.1 Introduction

This chapter outlines the data used in this study. Accounting and financial data for companies listed on three major U.S. stock exchanges are used in this study. The three major stock exchanges are: AMEX, NASDAQ and NYSE. Financial data is extracted from the merged CRSP/COMPUSTAT database while accounting data is extracted from the COMPUSTAT database. Financial data is available at monthly frequency while accounting data is available at annual frequency. Although data is available from the 1960s onwards, data particularly in the earlier periods is scarce. The final dataset utilises data between 1983 and 2006.

Additionally this study utilises the same filtering conditions as Fama French (1992, 1993). The choice to implement these conditions was based on several considerations. Firstly, a number of the filtering conditions are useful to ensure data integrity. Secondly, as some parallels are drawn between the Fama French model and the methodology employed in this study, it is important to ensure that the same data source is used to facilitate fair comparison.

The remainder of this chapter is structured as follows: Section 3.2 outlines the data items used and describes the source and purpose of those items. Section 3.3 outlines the filtering conditions applied to the data and Section 3.4 concludes.

3.2 Data items

Financial data (available in monthly frequency) is extracted from the merged CRSP/COMPUSTAT database, *prices, dividends and earnings file*. Accounting data (available in annual frequency) is extracted from the COMPUSTAT database, *monthly stock file* and *Industrial annual file*. In total, there are 14 data items which come from 3 separate database files.

All data is available at the firm level. Individual PERMNOs[18] are used to link the observations across the various data sources. Table 3-1 lists these items and specifies the CRSP/COMPUSTAT code, source file and purpose of the data item. Refer to Table 10-1 for a full description of the data items.

	Data Item	CRSP/COMPUSTAT code	CRSP/COMPUSTAT source file	Purpose
1	Standardised Industry Classification (SIC)	DNUM	Prices, dividends and earnings file	Modelling
2	Global Industry Classification	GIC	↓	Modelling
3	North American Industry Classification System	NAICS		Modelling
4	Closing Price	PRCC		Data filtering
5	Common Shares Outstanding	CSHOQ		Data filtering
6	Share Code	SHRCD		Data filtering
7	Exchange Code	EXCHD		Data filtering
8	Holding Period Return	RET	Monthly stock file	Modelling
9	Liquidation value	Data10	Industrial annual file	To calculate Book Equity
10	Investment Tax Credit (ITC)	Data51		
11	Redemption value	Data56	↓	↓
12	Book value (BV)	Data60		
13	Balance Sheet Deferred Taxes (BSDT)	Data74		
14	Par value (PV)	Data130		

Table 3-1 List of data items, CRSP/COMPUSTAT code, source file and purpose of the data item used in this study

The first three items contain the industry classification codes. Item 8 contains returns at the individual security level. Items 4 to 7 are used for data filtering. Filtering conditions are described in the following pages. Items 9 to 14 are used to calculate Book Equity which is also used for data filtering[19].

In addition to the three industry classification codes extracted from CRSP/COMPUSTAT, Fama-French (FF) industry classification (48 portfolios) is also extracted. In contrast to SIC, GIC and NAICS, the FF industry classification is developed by Fama and French (1997) and provides an alternative classification structure based on common risk characteristics.

[18] Which is a unique permanent security identification number assigned by CRSP. According to CRSP: "Unlike the CUSIP, Ticker Symbol, and Company Name, the PERMNO neither changes during an issue's trading history, nor is it reassigned after an issue ceases trading. The user may track a security through its entire trading history in CRSP's files with one PERMNO, regardless of name or capital structure changes", CRSP documentation [online] Available at: <http://www.crsp.com/documentation/product/stkind/definitions/PERMNO.html >, [accessed July 2011]

[19] For details on how Book Equity is calculated, refer to the appendices

In addition to the financial and accounting data items, three additional data items were constructed. These include: Market Equity (ME), Book Equity (BE) and Book to Market Equity (BE/ME).

Market Equity is simply the market capitalization of a stock. It is meant to proxy the 'size' of an equity. It is calculated as:

$$ME = Price \times Common\ Shares\ Outstanding$$

Book Equity on the other hand is a highly constructed variable. In the context of fundamentals analysis, BE is meant to represent the difference between a firm's assets and its liabilities.

According to Fama and French (1992; 1993), BE is defined as "the COMPUSTAT book value of stockholders' equity plus balance sheet deferred taxes and investment tax credit (if available), minus the book value of preferred stock. Depending on the availability, we use the redemption, liquidation or par value (in that order) to estimate the value of preferred stock."

To maintain consistency with Fama and French, BE was calculated as:

$$BE = BV + BSDT + ITC - Value\ of\ PS$$

Where BE = Book Equity
BV = Book Value
BSDT = Balance Sheet Deferred Taxes
ITC = Investment Tax Credit
Value of PS = Value of Preferred Stock

Value of Preferred Stock is constructed via the following:

If Redemption $\neq 0$, **then** Value of PS = Redemption; **else**

If Liquidity $\neq 0$, **then** Value of PS = Liquidity; **else**

If Par Value $\neq 0$, **then** Value of PS = Par Value; **else**

Value of PS = 0

Here, BV, BSDT and ITC represent a firm's assets while Value of PS represents its liabilities.

Lastly, the Book to Market Equity (BE/ME) is simply calculated as the ratio of Book Equity for the fiscal year ending in calendar year $t - 1$ to Market Equity at the end of December $t - 1$.

3.3 Filtering conditions

Figure 3-1 indicates the number of observations per year from 1965 to 2006. Approximately 20,000 to 25,000 annual observations were available in the 1960s. This grew sharply to approximately 50,000 in the 1970s with the founding of the NASDAQ. The number of annual observations grew steadily to 70,000 in the 1980s and over 100,000 in the late 1990s. The number of annual observations peaked at 106,596 in 1997 but declined in the following years.

The decision to implement Fama-French filtering conditions is twofold. Firstly, we agree with the authors in that a number of these filtering conditions are necessary to preserve data quality and integrity. For example the requirement for a company to have more than two years of historical data on CRSP/COMPUSTAT helps minimise the effect of survivorship bias. Secondly, as several important theoretical comparisons are made between the methodology in this study and that of Fama French (1992, 1993), it is important to ensure that the data same data source is used.

As with Fama and French (1992; 1993), several filtering conditions were applied to the data. In total, there are 7 distinct conditions:

1. Financial firms are excluded
2. Only stocks listed on NYSE, AMEX and NASDAQ are included
3. A stock must have CRSP monthly prices for December of year $t - 1$ and June of year t
4. A stock must have appeared on COMPUSTAT for more than 2 years
5. A stock must have BE for year $t - 1$
6. Only common shares are included
7. Shares with negative book equity are excluded

According to Fama and French (1992; 1993):

"We use all nonfinancial firms in the intersection of (a) the NYSE, AMEX and NASDAQ return files from CRSP and (b) the merged COMPUSTAT annual industrial files of income statement and balance sheet data. We exclude financial firms because the high leverage that is normal for these firms probably does not have the same meaning as for nonfinancial firms"

From this, we obtain conditions 1 and 2. Furthermore:

"To ensure that the accounting variables are known before the returns they are used to explain, we match the accounting data for all fiscal yearends in calendar year $t - 1$ (1962 – 1989) with the returns for July of year t to June $t + 1$."; and

"We use a firm's market equity at the end of December of year $t - 1$ to compute its book-to-market, leverage, and earnings-price ratios for $t - 1$, and we use its market equity for June of year t to measure its size. Thus, to be included in the return tests for July of year t, a firm must have a CRSP stock price for December of year $t - 1$ and June of year t. It must also have monthly returns for at least 24 of the 60 months preceding July of year t (for "pre-ranking" β estimates). And the firm must have COMPUSTAT data on total book assets (A), book equity (BE) and earnings (E), for its fiscal year ending in (any month of) calendar year $t - 1$."

From this, we obtain conditions 3 to 5.

"Only firms with ordinary common equity (as classified by CRSP) are included in the tests. This means that American Depository Receipts (ADRs), Real Estate Investment Trusts (REITs) and other units of beneficial interest are excluded."

From this, we obtain condition 6.

These conditions however, remove a substantial proportion of the data – approximately 45 to 50 percent in any given year. Figure 3-1 indicates the number of observations with and without the Fama-French filtering conditions. A large proportion of the data are removed throughout the 1960s and early 1970s however 'survivorship' improves in latter periods due mainly to increased data availability and more comprehensive data coverage.

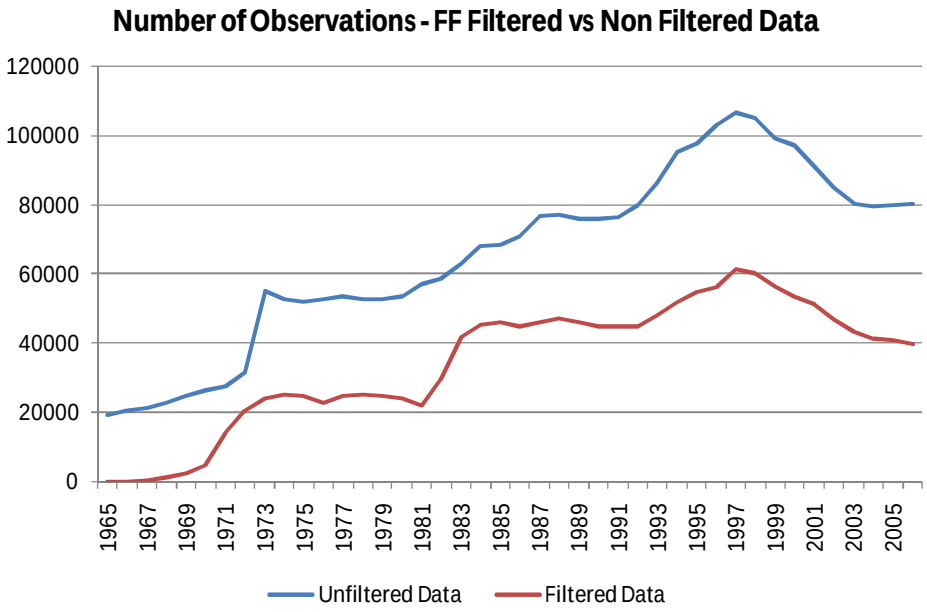


Figure 3-1 The line chart indicates the number of annual observations available on the merged CRSP/COMPUSTAT database between 1965 to 2006. The number of observations grew steadily between the 1960s and 1990s reaching a peak in 1997 after which it declined. Fama-French filtering conditions remove approximately 50 percent of data in any given year. More data are removed in earlier periods but 'survivorship' improves in latter periods due to improvements in data coverage and greater availability

Despite the substantial removal of data via the filtering conditions, approximately 40,000 to 60,000 observations per year are retained over the data period (see Chapter 5).

3.4 Conclusion

This chapter outlines the data used in this study. Financial and accounting data at the firm level from companies listed on three major U.S. stock exchanges: AMEX, NASDAQ and NYSE between 1963 and 2006 are extracted from the merged CRSP/COMPUSTAT and COMPUSTAT database.

Additionally, the Fama French (1992, 1993) filtering conditions are implemented. In total, there are 7 distinct filtering conditions. While most of these conditions are used to construct the Fama French *SMB* and *HML* factors, a number of conditions are also used to ensure data integrity. The FF filters are also implemented because a number of theoretical comparisons are made between the methodology used in this study and the FF results. Therefore, it is necessary to ensure that analysis is performed on the same dataset.

The following chapter discusses another important issue relating to the use of data – namely determining the *data period*, which refers to the starting to ending period and the *modelling interval* which refers to subsets of data used for modelling.

Chapter 4

4 Methodologies

4.1 Introduction

This study employs several innovative 'technologies'. This chapter describes these technologies in detail. How these technologies are used to address the various research objectives are described in this chapter but explored in greater detail in Section 6.2 and Section 7.3.

Cluster analysis refers to a set of statistical techniques that group elements into homogeneous groupings known as 'clusters' based on a set of measurable characteristics. Elements within a cluster are similar to each other in terms of these characteristics while elements between clusters are different.

The central technology of this study is a unique form of cluster analysis known as the Generalised Style Classification algorithm, GSC. Originally developed by Brown and Goetzman (1996) to study mutual funds styles, the GSC is applied here to individual stock data at the firm level. The GSC is a clustering algorithm that possesses several features that make it potentially superior to other forms of cluster analysis. These include a generalised least squares (GLS) correction, which adjusts for heteroskedasticity, i.e. non-constant variance and the ability to incorporate both cross sectional as well as dynamic time varying features of a dataset. Heteroskedasticity is common in financial time series such as stock returns as these variables commonly experience periods of extreme fluctuations caused by abnormal market events like stock market crashes. Furthermore, common forms of cluster analysis can typically be applied only to a cross section of data losing much of the dynamic time varying information.

It is these desirable features that make the GSC ideal for studying financial data. For these reasons the GSC is selected as the clustering technology throughout this study.

The effectiveness of the GSC is further enhanced by the incorporation of the Gap statistic test. Developed by Tibshirani, Walther and Hastie (2001), the Gap statistic test is a revolutionary method to objectively determine the optimal number of clusters, *K*. A key problem in any form of cluster analysis is determining the optimal number of clusters, *K*. The Gap statistic represents the solution to that problem. Common approaches to determining *K* often begin by establishing an objective measure of cluster performance (such as the total sum of squared deviations of individual

observations from their cluster means). Once this objective measure is established, *ad hoc* experimentation is used to find a significant reduction in the measure. The Gap statistic essentially formalises this procedure. It does so by first establishing a null distribution for the objective measure (through Monte Carlo simulation) and comparing this distribution against the actual estimated objective measure.

When combined the GSC and Gap statistic provide an ideal approach to cluster analysis, especially as they apply to financial time series.

Throughout this study the GSC is utilised in two ways. Firstly, it is used to 'profile' the industrial sectors that individual stocks belong to, creating a better understanding of these sectors – in particular as they relate to risk and return. Secondly, it is used to derive better industry cost of capital estimates, which in turn have a number of highly useful applications – one of which is to improve NPV calculations.

The remainder of this chapter is structured as follows. Section 4.2 describes the GSC algorithm. Section 4.3 describes the Gap statistic test. Section 4.4 explains how the GSC and Gap statistic can be used to profile industrial sectors. Section 4.5 explains how the GSC can be used to improve cost of capital estimates. Section 4.6 concludes.

4.2 The Generalised Style Classification algorithm

Like many forms of cluster analysis, the GSC operates on the central principle of minimising the total sum of squared deviations of individual elements from their respective cluster means. However, the innovative features inherent to the GSC allow it to form better clusters, which are robust to the adverse effects of heteroskedasticity.

The GSC was developed by Brown and Goetzmann (1996) to classify mutual fund styles. The authors were concerned with the apparent disparity between a fund's self reported 'style' classification and that which actually occurs in practice. Such misclassification (whether intentional or otherwise) has significant implications for investors who may mistakenly make an investment based on the reported style but are misled as to the actual strategy of the fund. Brown and Goetzmann also hypothesize that self misclassification occurs as management styles are used to measure performance and compensation. By pursuing a different strategy, fund managers may be able to give

the appearance of superior performance when compared to their stated benchmark objectives thus achieving better performance warranting greater compensation. They propose the use of the GSC as an objective means of classifying mutual fund styles thus addressing issues of misclassification.

Through the use of the GSC Brown and Goetzmann were able to find 8 clusters, which correspond to 8 distinct styles of fund management as opposed to the 15 styles as reported by Morningstar. To determine the appropriate number of clusters, they used the likelihood ratio, LR test developed by Quandt (1960). However, there were several issues with the use of the LR test. Firstly, it was not clear how many degrees of freedom were required to determine the critical values and secondly whether the null distribution did in fact have a χ^2 distribution. Lastly, it was hypothesized that the LR test tended to estimate too many styles.

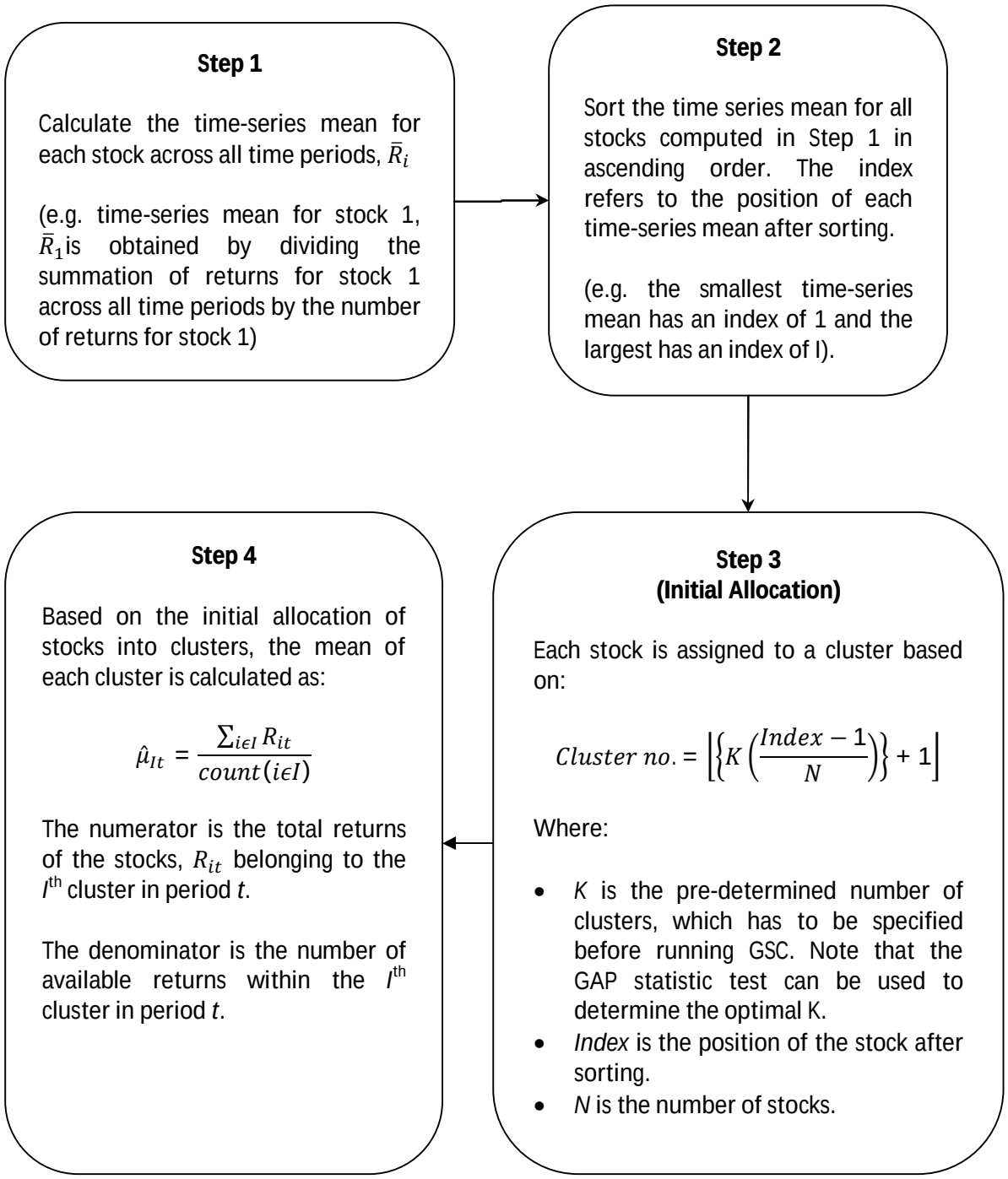
Tibshirani, Walther and Hastie (2001) circumvented these issues with the use of the Gap statistic. The gap statistic test does not assume a χ^2 distribution for testing purposes. In fact the gap statistic test generates its own null distribution through repeated applications of the GSC. Because of this, the gap statistic test is regarded as a superior approach to determining the appropriate number of clusters as opposed to the LR test initially used by Brown and Goetzmann (1997). The following section describes how the actual GSC algorithm works.

Consider an *i* x *T* array of individual stock returns:

Stocks\Time (monthly)	**1**	**2**	**3**	**…**	**T**	**Time-series Mean**
1	R_{11}	R_{12}	R_{13}	…	R_{1T}	$\overline{R_1}$
2	R_{21}	R_{22}	R_{23}	…	R_{2T}	$\overline{R_2}$
3	R_{31}	R_{32}	R_{33}	…	R_{3T}	$\overline{R_3}$
4	R_{41}	R_{42}	R_{43}	…	R_{4T}	$\overline{R_4}$
…	…	…	…	…	…	…
i	R_{i1}	R_{i2}	R_{i3}		R_{iT}	$\overline{R_i}$

Table 4-1 Monthly stock returns for *I* stocks over *T* periods where R_{iT} represents the return of the i^{th} stock during the T^{th} period.

The following flowchart is a diagrammatic representation of the GSC algorithm[20]:



[20] The notation is adapted from Brown and Goetzman (1996)

Step 5

Calculate the time-series variance of each stock as:

$$\text{var}(\hat{e}_i) = \frac{\sum_{t=1}^{T}(R_{it} - \hat{\mu}_{It})^2}{T^* - 1}$$

Calculate the cross-sectional variance at each time period as:

$$\text{var}(\hat{e}_t) = \frac{\sum_{i=1}^{I}(R_{it} - \hat{\mu}_{It})^2}{N^* - 1}$$

Where:

- R_{it} is the return of the i^{th} stock in period *t*
- $\hat{\mu}_{It}$ is the corresponding cluster mean that stock *i* belongs to in period *t*
- T^* is the number of available returns for stock *i* across all time periods.
- N^* is the number of available returns at each period *t*.

Step 6

Correct each cluster mean calculated in Step 4 by adjusting for the time-series variance:

$$\hat{\mu}_{It}^* = \sum_{i \epsilon I} \frac{R_{it}}{\text{var}(\hat{e}_i)} / \sum_{i \epsilon I} \frac{1}{\text{var}(\hat{e}_i)}$$

Where:

- R_{it} is the return of the i^{th} stock in period *t*
- $\text{var}(\hat{e}_i)$ is the time-series variance of stock *i* obtained from Step 5
- $\hat{\mu}_{It}^*$ is the Generalised Least Squares (GLS) estimate of the cluster mean. This is where the name Generalised Style Classification (GSC) is derived.

Step 7

Update both the time-series and cross-sectional variance in Step 5 replacing $\hat{\mu}_{It}$ with the corrected cluster mean, $\hat{\mu}_{It}^*$ obtained in Step 6. The formulas become:

$$\text{var}(\hat{e}_i^*) = \frac{\sum_{t=1}^{T}(R_{it} - \hat{\mu}_{It}^*)^2}{T^* - 1}$$

$$\text{var}(\hat{e}_t^*) = \frac{\sum_{i=1}^{I}(R_{it} - \hat{\mu}_{It}^*)^2}{N^* - 1}$$

Step 8

Calculate the total Sum of Squares (SSQ) for all returns of all stocks across all time periods for each switch (hence the triple summation). When a stock switches from one cluster to another, the within cluster sum of squares for the j^{th} switch is:

$$SSQ_j = \sum_{t=1}^{T} \sum_{I \epsilon I_j} \sum_{i \epsilon I} \frac{(R_{it} - \hat{\mu}_{It}^*)^2}{\text{var}(\hat{e}_i^*)\text{var}(\hat{e}_t^*)}$$

Where:

- I_j represents the clusters formed at the j^{th} switch
- $\hat{\mu}_{It}^*$ is the GLS estimate of the cluster mean calculated in Step 6
- $\text{var}(\hat{e}_i^*)$ and $\text{var}(\hat{e}_t^*)$ are the updated time-series and cross-sectional variances respectively

As evident from the flowchart, the GSC is a non-hierarchical (*k*-means) clustering technique based on minimum variance criterion. Unlike more common cluster analysis techniques which are typically applied to cross sectional data, the GSC approach can be applied to **cross sectional**, **time series** data.

Statistically, the returns data also exhibit heteroskedasticity. By taking into consideration time series and cross sectional variances, the GSC approach adjusts for this problem. This is done via generalized least squares (GLS) correction. If a stock (or group of stocks) were to exhibit heteroskedasticity, the GSC approach would minimize this effect by allocating a smaller weight to these observations. In this way, it reduces the effect of outliers on the classification thus improving the cluster profiles.

In principle, the GSC approach works by minimizing both the cross sectional and longitudinal variance via a two stage loop.

Every iteration of the GSC approach computes time series and cross sectional variances in two stages. The first stage computes times series and cross sectional variances based on cluster means obtained from initial allocation. This information is then used to update the cluster means via a generalized least squares correction (hence the name, Generalised Style Classification). The second stage then uses the updated cluster means to recalculate updated time series and cross sectional variances which in turn compute the sum of squares which forms the minimizing variable or objective function.

4.2.1 A step by step illustration

The best way to explain the logic of the GSC is to apply it to a small test dataset synthesized from the original data for which the number of clusters, *K* is known *a priori*. Table 4-2 contains the synthetic returns data.

Stocks / Time	Monthly Returns											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.0100	0.0300	0.0450	0.0050	0.0200	0.0400	0.0500	0.0100	-0.0400	-0.0200	0.0100	0.0050
2	0.0090	0.0230	0.0530	0.0060	0.0260	0.0330	0.0410	0.0200	-0.0490	-0.0230	0.0070	0.0040
3	0.0120	0.0250	0.0480	0.0050	0.0240	0.0340	0.0440	0.0080	-0.0310	-0.0290	0.0010	0.0040
4	0.0120	0.0170	0.0500	0.0060	0.0260	0.0340	0.0470	0.0220	-0.0460	-0.0310	0.0070	-0.0030
5	0.0040	0.0330	0.0540	0.0010	0.0290	0.0430	0.0410	0.0010	-0.0230	-0.0390	-0.0020	0.0020
6	0.0150	0.0270	0.0400	0.0100	0.0350	0.0440	0.0490	0.0290	-0.0550	-0.0340	0.0060	-0.0130
7	0.0090	0.0260	0.0620	0.0010	0.0320	0.0370	0.0440	0.0080	-0.0300	-0.0310	-0.0060	0.0000
8	-0.0200	-0.0500	-0.0600	-0.0200	-0.0300	-0.0400	-0.0800	-0.0400	0.0200	0.0300	-0.0300	-0.0100
9	-0.0290	-0.0420	-0.0650	-0.0210	-0.0230	-0.0350	-0.0860	-0.0500	0.0180	0.0330	-0.0310	-0.0080
10	-0.0210	-0.0460	-0.0670	-0.0120	-0.0270	-0.0450	-0.0860	-0.0520	0.0150	0.0330	-0.0390	-0.0180
11	-0.0180	-0.0560	-0.0740	-0.0050	-0.0260	-0.0430	-0.0900	-0.0600	0.0170	0.0260	-0.0370	-0.0090
12	-0.0120	-0.0470	-0.0840	-0.0150	-0.0180	-0.0510	-0.0990	-0.0580	0.0150	0.0360	-0.0430	-0.0070
13	-0.0130	-0.0550	-0.0800	-0.0240	-0.0210	-0.0530	-0.0910	-0.0580	0.0120	0.0360	-0.0430	-0.0090
14	-0.0010	0.0010	0.0020	0.0010	-0.0100	0.0060	0.0020	0.0040	0.0010	0.0020	-0.0020	-0.0010
15	-0.0001	0.0012	0.0029	0.0003	-0.0099	0.0053	0.0029	0.0033	0.0006	0.0015	-0.0027	-0.0002
16	-0.0001	0.0004	0.0036	-0.0001	-0.0093	0.0049	0.0027	0.0041	0.0006	0.0020	-0.0020	-0.0010
17	0.0002	-0.0001	0.0045	-0.0010	-0.0090	0.0059	0.0037	0.0049	0.0010	0.0029	-0.0011	-0.0007
18	0.0000	-0.0009	0.0037	-0.0003	-0.0097	0.0060	0.0043	0.0049	0.0019	0.0027	-0.0020	-0.0016
19	-0.0010	-0.0012	0.0044	-0.0002	-0.0095	0.0053	0.0053	0.0045	0.0021	0.0032	-0.0030	-0.0021
20	-0.0020	-0.0008	0.0041	-0.0007	-0.0098	0.0048	0.0054	0.0046	0.0013	0.0023	-0.0022	-0.0018

Table 4-2 Hypothetical returns for 20 stocks over a 12 month period

Figure 4-1 shows a line chart of the synthetic returns:

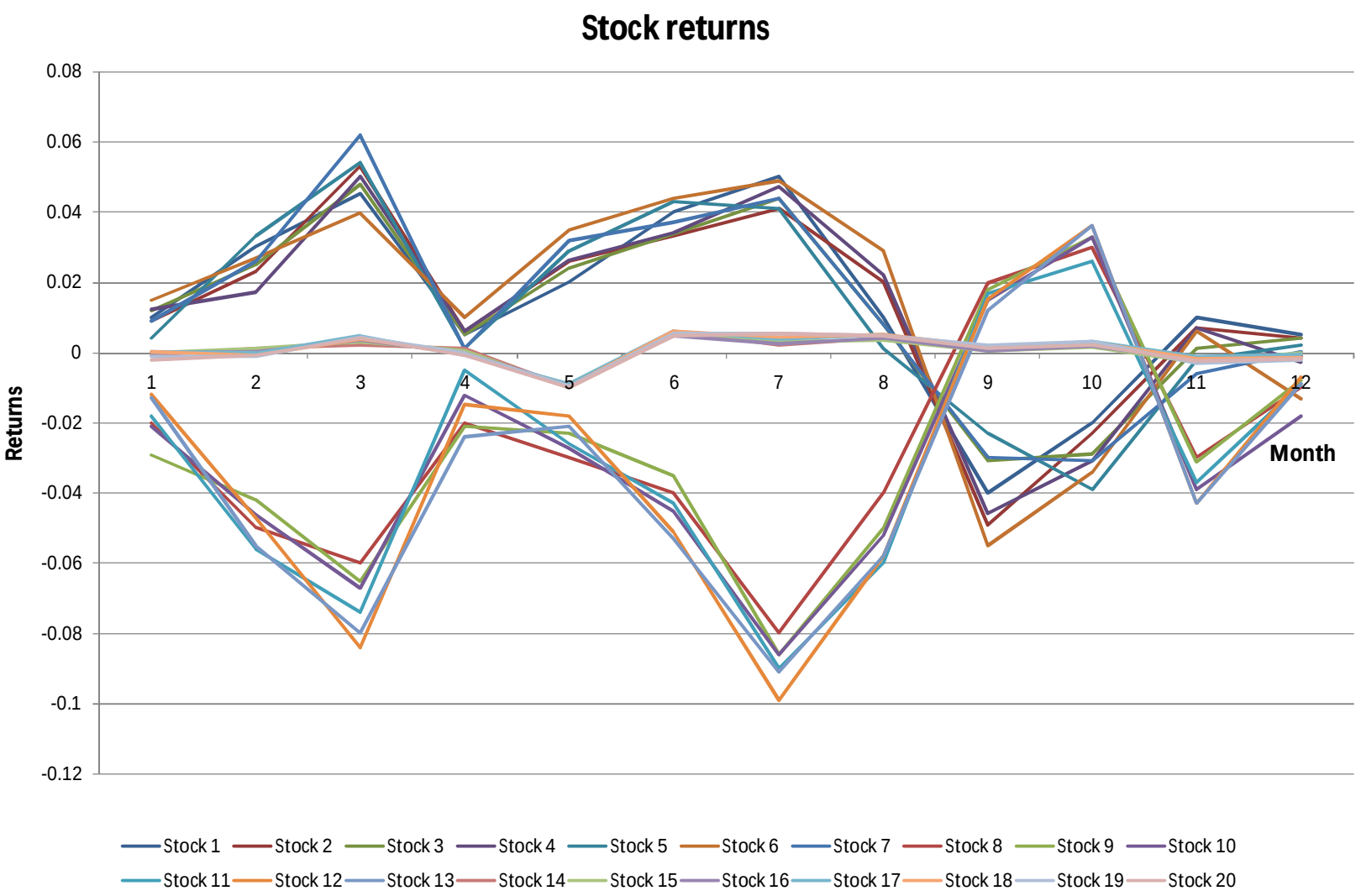


Figure 4-1 Line chart of 20 synthetic stock returns

From Figure 4-1, it is apparent that there are three distinct groups of stocks. The first group consists of stocks 1 to 7. These stocks appear to have steady growth in returns from month 1 to 3 reaching a peak of 4 to 6%, followed by a sharp decline in month 4. There is steady growth until another peak in month 7 followed by a period of decline until month 9 to 10, after which returns stabilize at approximately 0 to 1%. These will be referred to as the *positive return* stocks.

The second group of stocks consists of stocks 8 to 13. This group appears to move in an opposite direction to group 1 indicating negative covariance between these two groups. Since this group consists of stocks predominantly negative returns, they will be referred to as the *negative return* stocks.

The third group of stocks consists of stocks 14 to 20 and they exhibit lower volatility than the other two groups with returns fluctuating about 0%. These will be referred to as the *zero return* stocks.

A priori, there appear to be three distinct groups of stocks or more specifically three distinct patterns of returns. Performing k-means cluster analysis on the returns data should yield an optimal solution with $K = 3$. Furthermore, stocks 1 to 7 would be expected to form one cluster, stocks 8 to 13 to form another cluster and stocks 14 to 20 to form another cluster as this would be consistent with their pattern of variation.

Step 1: Calculating the time series mean

The first step in the GSC approach is to calculate time series means for each of the *i* stocks across all time periods. Time series means are calculated as:

$$\overline{R}_i = \frac{1}{n}\sum_{t=1}^{T} R_{it}$$

Calculating time series means for the hypothetical data in Table 4-2 yields:

Stocks / Time	Monthly Returns							
	1	2	3	⋯	10	11	12	$\overline{R}_i$
1	0.0100	0.0300	0.0450	…	-0.0200	0.0100	0.0050	0.0138
2	0.0090	0.0230	0.0530	…	-0.0230	0.0070	0.0040	0.0125
3	0.0120	0.0250	0.0480	…	-0.0290	0.0010	0.0040	0.0121
4	0.0120	0.0170	0.0500	…	-0.0310	0.0070	-0.0030	0.0118
5	0.0040	0.0330	0.0540	…	-0.0390	-0.0020	0.0020	0.0120
6	0.0150	0.0270	0.0400	…	-0.0340	0.0060	-0.0130	0.0128
7	0.0090	0.0260	0.0620	…	-0.0310	-0.0060	0.0000	0.0127
8	-0.0200	-0.0500	-0.0600	…	0.0300	-0.0300	-0.0100	-0.0275
9	-0.0290	-0.0420	-0.0650	…	0.0330	-0.0310	-0.0080	-0.0283
10	-0.0210	-0.0460	-0.0670	…	0.0330	-0.0390	-0.0180	-0.0304
11	-0.0180	-0.0560	-0.0740	…	0.0260	-0.0370	-0.0090	-0.0313
12	-0.0120	-0.0470	-0.0840	…	0.0360	-0.0430	-0.0070	-0.0319
13	-0.0130	-0.0550	-0.0800	…	0.0360	-0.0430	-0.0090	-0.0333
14	-0.0010	0.0010	0.0020	…	0.0020	-0.0020	-0.0010	0.0004
15	-0.0001	0.0012	0.0029	…	0.0015	-0.0027	-0.0002	0.0004
16	-0.0001	0.0004	0.0036	…	0.0020	-0.0020	-0.0010	0.0005
17	0.0002	-0.0001	0.0045	…	0.0029	-0.0011	-0.0007	0.0009
18	0.0000	-0.0009	0.0037	…	0.0027	-0.0020	-0.0016	0.0008
19	-0.0010	-0.0012	0.0044	…	0.0032	-0.0030	-0.0021	0.0007
20	-0.0020	-0.0008	0.0041	…	0.0023	-0.0022	-0.0018	0.0004

Step 2: Sorting the time series mean

The second step in the GSC sorts the time series means calculated in step 1 into ascending order and assigning an index to the sorted array of stocks. The stock with the lowest time series mean has an index of 1 while the stock with the highest time series mean has an index of *I*.

Step 3: Initial allocation

The third step in the GSC approach is to initially allocate stocks to clusters based on the ranked means. Each stock is assigned to a cluster using the formula:

$$\left\{K\left(\frac{index-1}{N}\right)\right\}+1$$

Where K = pre-determined number of clusters

N = number of stocks

index = rank Index from step 2

Performing steps 2 and 3 yields:

Stocks / Time	Monthly Returns 1	2	...	11	12	$\overline{R}_i$	Index	Initial allocation
13	-0.0130	-0.0550	...	-0.0430	-0.0090	-0.0333	1	1
12	-0.0120	-0.0470	...	-0.0430	-0.0070	-0.0319	2	1
11	-0.0180	-0.0560	...	-0.0370	-0.0090	-0.0313	3	1
10	-0.0210	-0.0460	...	-0.0390	-0.0180	-0.0304	4	1
9	-0.0290	-0.0420	...	-0.0310	-0.0080	-0.0283	5	1
8	-0.0200	-0.0500	...	-0.0300	-0.0100	-0.0275	6	2
14	-0.0010	0.0010	...	-0.0020	-0.0010	0.0004	7	2
15	-0.0001	0.0012	...	-0.0027	-0.0002	0.0004	8	2
20	-0.0020	-0.0008	...	-0.0022	-0.0018	0.0004	9	2
16	-0.0001	0.0004	...	-0.0020	-0.0010	0.0005	10	2
19	-0.0010	-0.0012	...	-0.0030	-0.0021	0.0007	11	2
18	0.0000	-0.0009	...	-0.0020	-0.0016	0.0008	12	2
17	0.0002	-0.0001	...	-0.0011	-0.0007	0.0009	13	2
4	0.0120	0.0170	...	0.0070	-0.0030	0.0118	14	2
5	0.0040	0.0330	...	-0.0020	0.0020	0.0120	15	2
3	0.0120	0.0250	...	0.0010	0.0040	0.0121	16	3
2	0.0090	0.0230	...	0.0070	0.0040	0.0125	17	3
7	0.0090	0.0260	...	-0.0060	0.0000	0.0127	18	3
6	0.0150	0.0270	...	0.0060	-0.0130	0.0128	19	3
1	0.0100	0.0300	...	0.0100	0.0050	0.0138	20	3

Step 4: Initial cluster means

After step 3, the stocks have been allocated into *K* clusters. The next step in the GSC approach is to calculate the cluster means. This will produce a *K* × *T* array as there are *K* cluster means for each of the *T* time periods. Cluster means are calculated as:

$$\hat{\mu}_{It} = \frac{\sum_{i \in I} R_{it}}{count(i \in I)}$$

Where $\hat{\mu}_{It}$ is the mean of the I^{th} cluster in month *t*

R_{it} is the return of the i^{th} stock in month *t*

That is the mean of each cluster is the sum of the Returns belonging to that cluster divided by the number of stocks belonging to that cluster.

For example, there are 5 stocks in cluster 1 after initial allocation. Therefore, the mean of cluster 1 in period 1, $\hat{\mu}_{11}$ is that given by the sum of the returns belonging to cluster 1 in period 1 divided by the number of stocks in cluster 1, which is 5.

$$\hat{\mu}_{11} = \frac{\sum_{i \in I} R_{it}}{count(i \in I)}$$
$$= \frac{R_{13,1} + R_{12,1} + R_{11,1} + R_{10,1} + R_{9,1}}{5}$$
$$= \frac{-0.0130 - 0.0120 - 0.0180 - 0.0210 - 0.0290}{5} = -0.0186$$

Applying this to the data yields:

Cluster /					**Cluster means**			
Time	1	2	3	...	9	10	11	12
1	-0.0186	-0.0492	-0.0740	...	0.0154	0.0328	-0.0386	-0.0102
2	-0.0008	0.0000	0.0069	...	-0.0041	-0.0023	-0.0040	-0.0019
3	0.0110	0.0262	0.0496	...	-0.0410	-0.0274	0.0036	0.0000

Step 5: Time series and cross sectional variances

The next step in the GSC approach is to calculate the time series and cross sectional variances. Variance is calculated as the sum of the squared deviations from the mean divided by the number of observations minus 1. In the current context, the time series variance is the sum of the squared deviations of time series returns from their respective cluster means divided by the number of available time periods minus 1. It is calculated thus:

$$\text{var}(\hat{e}_i) = \left(\frac{\sum_{t=1}^{T} (R_{it} - \hat{\mu}_{It})^2}{T^* - 1} \right)$$

Conversely, cross sectional variance is the sum of the squared deviations of cross sectional returns from their respective cluster means divided by the number of available cross sectional observations minus 1. It is calculated thus:

$$\text{var}(\hat{e}_t) = \left(\frac{\sum_{i=1}^{I} (R_{it} - \hat{\mu}_{It})^2}{N^* - 1} \right)$$

Where T^* is the number of available time series observations for the i^{th} stock

N^* is the number of available cross sectional observations for month t

For example, stock 13 is the first stock in cluster 1 after initial allocation. The time series variance for stock 13 is given by:

$$\text{var}(\hat{e}_{13}) = \left[\frac{(R_{13,1} - \hat{\mu}_{1,1})^2 + (R_{13,2} - \hat{\mu}_{1,2})^2 + (R_{13,1} - \hat{\mu}_{1,3})^2 + \cdots + (R_{13,12} - \hat{\mu}_{1,12})^2}{12 - 1} \right]$$

$$= \left[\frac{(-0.031 + 0.0186)^2 + (-0.0550 + 0.0492)^2 + (-0.0800 + 0.0740)^2 + \cdots + (-0.0090 + 0.0102)^2}{12 - 1} \right]$$

$$= 2.595 \times 10^{-5}$$

The cross sectional variance for month 1 is given by:

$$\text{var}(\hat{e}_1) = \left[\frac{(R_{13,1} - \hat{\mu}_{1,1})^2 + (R_{12,1} - \hat{\mu}_{1,1})^2 + (R_{11,1} - \hat{\mu}_{1,1})^2 + \cdots + (R_{1,1} - \hat{\mu}_{3,1})^2}{12 - 1} \right]$$

$$= \left[\frac{(-0.031 + 0.0186)^2 + (-0.0120 + 0.0186)^2 + (-0.0180 + 0.0186)^2 + \cdots + (0.0100 + 0.0110)^2}{12 - 1} \right]$$

$$= 4.080 \times 10^{-5}$$

The following table helps illustrate the calculation of time series and cross sectional variances:

Stocks	$(R_{it} - \hat{\mu}_{it})^2$					Time series variance
	1	2	...	11	12	
13	$(R_{13,1} - \hat{\mu}_{1,1})^2$	$(R_{13,2} - \hat{\mu}_{1,2})^2$	⋯	$(R_{13,11} - \hat{\mu}_{1,11})^2$	$(R_{13,12} - \hat{\mu}_{1,12})^2$	$\frac{\sum_{t=1}^{12}(R_{13,t} - \hat{\mu}_{1t})^2}{12 - 1}$
12	$(R_{12,1} - \hat{\mu}_{1,1})^2$	$(R_{12,2} - \hat{\mu}_{1,2})^2$	⋯	$(R_{it} - \hat{\mu}_{1,11})^2$	$(R_{12,12} - \hat{\mu}_{1,12})^2$	$\frac{\sum_{t=1}^{12}(R_{12,t} - \hat{\mu}_{1t})^2}{12 - 1}$
...			...			
6	$(R_{6,1} - \hat{\mu}_{3,1})^2$	$(R_{6,2} - \hat{\mu}_{3,2})^2$	⋯	$(R_{it} - \hat{\mu}_{3,11})^2$	$(R_{6,12} - \hat{\mu}_{3,12})^2$	$\frac{\sum_{t=1}^{12}(R_{6,t} - \hat{\mu}_{3t})^2}{12 - 1}$
1	$(R_{1,1} - \hat{\mu}_{3,1})^2$	$(R_{1,2} - \hat{\mu}_{3,2})^2$	⋯	$(R_{it} - \hat{\mu}_{3,11})^2$	$(R_{1,12} - \hat{\mu}_{3,12})^2$	$\frac{\sum_{t=1}^{12}(R_{1,t} - \hat{\mu}_{3t})^2}{12 - 1}$
Cross sectional variance	$\frac{\sum_{i=1}^{20}(R_{i,1} - \hat{\mu}_{I1})^2}{20 - 1}$	$\frac{\sum_{i=1}^{20}(R_{i,2} - \hat{\mu}_{I2})^2}{20 - 1}$		$\frac{\sum_{i=1}^{20}(R_{i,11} - \hat{\mu}_{I11})^2}{20 - 1}$	$\frac{\sum_{i=1}^{20}(R_{i,12} - \hat{\mu}_{I12})}{20 - 1}$	

The time series and cross sectional variances are below:

Stocks	$(R_{it} - \hat{\mu}_{It})^2$					Time series variance
	1	2	...	11	12	
13	3.14×10^{-5}	3.36×10^{-5}	...	1.94×10^{-5}	1.44×10^{-6}	2.60×10^{-5}
12	4.36×10^{-5}	4.84×10^{-6}	...	1.94×10^{-5}	1.02×10^{-5}	2.95×10^{-5}
...	...	...	...	...	...	...
6	1.60×10^{-5}	6.40×10^{-7}	...	5.76×10^{-6}	1.69×10^{-4}	7.73×10^{-5}
1	1.00×10^{-6}	1.44×10^{-5}	...	4.10×10^{-5}	2.50×10^{-5}	2.39×10^{-5}
Cross sectional variance	4.08×10^{-5}	2.14×10^{-4}	...	5.70×10^{-5}	2.07×10^{-5}	

This will produce a vector of time series variances with 20 elements (one for each stock) and a vector of cross sectional variances with 12 elements (one for each month).

Step 6: Corrected cluster means

The next step is to adjust the cluster means obtained in step 4 with the time series variances via the following correction:

$$\hat{\mu}_{It}^{*} = \sum_{i \epsilon I} \frac{R_{it}}{\text{var}(\hat{e}_i)} / \sum_{i \epsilon I} \frac{1}{\text{var}(\hat{e}_i)}$$

$\hat{\mu}_{It}^{*}$ is the generalised least squares (GLS) estimate of the cluster mean. This is where the name generalised style classification (GSC) is derived. In principle, correcting the cluster means in this way integrates longitudinal variance such that it is now captured in the cluster means.

More importantly, the inverse of the time series variance acts as a weighting mechanism to reduce the impact or contribution that outliers make on the corrected cluster means. As Brown and Goetzmann (1996, p.380) explain:

"A modification of the basic algorithm is a generalized least squares procedure, which allows time-varying and fund-specific residual return variance. By scaling observations by the inverse of the estimated standard deviation, we decrease the influence of extreme observations in the classification process."

To obtain the corrected cluster means, the ratios of the cross sectional returns to their respective time series variances are summed over each cluster and divided by the sum of the inverse time series variances for each cluster.

For example, there are 5 stocks in cluster 1 after initial allocation. The corrected cluster mean is calculated thus:

$$\hat{\mu}_{11}^{*} = \frac{\left[\frac{R_{13,1}}{\text{var}(\hat{e}_{13})} + \frac{R_{12,1}}{\text{var}(\hat{e}_{12})} + \frac{R_{11,1}}{\text{var}(\hat{e}_{11})} + \frac{R_{10,1}}{\text{var}(\hat{e}_{10})} + \frac{R_{9,1}}{\text{var}(\hat{e}_{9})}\right]}{\left[\frac{1}{\text{var}(\hat{e}_{13})} + \frac{1}{\text{var}(\hat{e}_{12})} + \frac{1}{\text{var}(\hat{e}_{11})} + \frac{1}{\text{var}(\hat{e}_{10})} + \frac{1}{\text{var}(\hat{e}_{9})}\right]}$$

$$= \frac{\left[\frac{-0.0130}{2.60 \times 10^{-5}} + \frac{-0.0120}{2.95 \times 10^{-5}} + \frac{-0.0180}{2.20 \times 10^{-5}} + \frac{-0.0210}{1.69 \times 10^{-5}} + \frac{-0.0290}{4.55 \times 10^{-5}}\right]}{\left[\frac{1}{2.60 \times 10^{-5}} + \frac{1}{2.95 \times 10^{-5}} + \frac{1}{2.20 \times 10^{-5}} + \frac{1}{1.69 \times 10^{-5}} + \frac{1}{4.55 \times 10^{-5}}\right]}$$

$$= -0.01812$$

Applying this to the data yields:

Cluster	Updated Cluster means							
	1	2	3	...	9	10	11	12
1	-0.0181	-0.0498	-0.0738	...	0.0152	0.0325	-0.0391	-0.0112
2	-0.0005	0.0000	0.0037	...	0.0011	0.0022	-0.0021	-0.0012
3	0.0105	0.0258	0.0500	...	-0.0398	-0.0258	0.0042	0.0025

Step 7: Updated time series and cross sectional variance

As with step 5, time series and cross sectional variances are calculated. However, the initial cluster means are replaced with the corrected cluster means obtained from step 6. Time series and cross sectional variances are calculated thus:

$$\text{var}(\hat{e}_i^{\,*}) = \left(\frac{\sum_{t=1}^{T} (R_{it} - \hat{\mu}_{It}^{*})^2}{T^{*} - 1} \right)$$

$$\text{var}(\hat{e}_t^{\,*}) = \left(\frac{\sum_{i=1}^{I} (R_{it} - \hat{\mu}_{It}^{*})^2}{N^{*} - 1} \right)$$

Where $\hat{\mu}_{It}^{*}$ = corrected cluster mean of cluster *I* in month *t*

$\text{var}(\hat{e}_i^{\,*})$ = updated time series variance for stock *i*

$\text{var}(\hat{e}_t^{\,*})$ = updated cross sectional variance in month *t*

The updated time series and cross sectional variances are below:

Stocks	$(R_{it} - \hat{\mu}_{It}^{*})^2$					Updated Time series variance
	1	2	...	11	12	
13	2.62×10^{-5}	2.75×10^{-5}	...	1.51×10^{-5}	4.93×10^{-6}	2.66×10^{-5}
12	3.74×10^{-5}	7.61×10^{-6}	...	1.51×10^{-5}	1.78×10^{-5}	3.05×10^{-5}
...	...	...	...	...	...	...
6	2.01×10^{-5}	1.39×10^{-6}	...	3.07×10^{-6}	2.41×10^{-4}	1.00×10^{-4}
1	2.64×10^{-7}	1.75×10^{-5}	...	3.31×10^{-5}	6.09×10^{-6}	1.80×10^{-5}
Updated Cross sect var	4.09×10^{-5}	2.14×10^{-4}	...	5.90×10^{-5}	2.30×10^{-5}	

This will again produce a vector of time series variances with 20 elements (one for each stock) and a vector of cross sectional variances with 12 elements (one for each month).

Step 8: Calculate the sum of squares

The last step is to calculate the sum of squares (SSQ) for returns of all stocks across all time periods for each switch. Therefore, the sum of squares is a triple summation. When a stock switches from one cluster to another, the within-cluster sum of squares for the j^{th} switch is:

$$SSQ_j = \sum_{t=1}^{T} \sum_{I \in I_j} \sum_{i \in I} \frac{(R_{it} - \hat{\mu}_{It}^*)^2}{\text{var}(\hat{e}_i^*)\,\text{var}(\hat{e}_t^*)}$$

SSQ at the j^{th} switch is calculated as the squared deviation of a return from its cluster mean divided by the product of the time series and cross sectional variances for all *i* stocks across all *t* periods at each j^{th} switch[21].

The following table illustrates how SSQ is calculated:

Stocks	$\frac{(R_{it} - \hat{\mu}_{It}^*)^2}{\text{var}(\hat{e}_i^*)\,\text{var}(\hat{e}_t^*)}$				
	1	2	...	11	12
13	$\frac{(R_{13,1} - \hat{\mu}_{11}^*)^2}{\text{var}(\hat{e}_{13}^*)\,\text{var}(\hat{e}_1^*)}$	$\frac{(R_{13,2} - \hat{\mu}_{12}^*)^2}{\text{var}(\hat{e}_{13}^*)\,\text{var}(\hat{e}_2^*)}$	...	$\frac{(R_{13,11} - \hat{\mu}_{1,11}^*)^2}{\text{var}(\hat{e}_{13}^*)\,\text{var}(\hat{e}_{11}^*)}$	$\frac{(R_{13,12} - \hat{\mu}_{1,12}^*)^2}{\text{var}(\hat{e}_{13}^*)\,\text{var}(\hat{e}_{12}^*)}$
12	$\frac{(R_{12,1} - \hat{\mu}_{11}^*)^2}{\text{var}(\hat{e}_{12}^*)\,\text{var}(\hat{e}_1^*)}$	$\frac{(R_{12,2} - \hat{\mu}_{12}^*)^2}{\text{var}(\hat{e}_{12}^*)\,\text{var}(\hat{e}_2^*)}$	...	$\frac{(R_{12,11} - \hat{\mu}_{1,11}^*)^2}{\text{var}(\hat{e}_{12}^*)\,\text{var}(\hat{e}_{11}^*)}$	$\frac{(R_{12,12} - \hat{\mu}_{1,12}^*)^2}{\text{var}(\hat{e}_{12}^*)\,\text{var}(\hat{e}_{12}^*)}$
...	...	...	...	...	...
6	$\frac{(R_{6,1} - \hat{\mu}_{31}^*)^2}{\text{var}(\hat{e}_6^*)\,\text{var}(\hat{e}_1^*)}$	$\frac{(R_{6,2} - \hat{\mu}_{32}^*)^2}{\text{var}(\hat{e}_6^*)\,\text{var}(\hat{e}_2^*)}$	...	$\frac{(R_{6,11} - \hat{\mu}_{3,11}^*)^2}{\text{var}(\hat{e}_6^*)\,\text{var}(\hat{e}_{11}^*)}$	$\frac{(R_{6,12} - \hat{\mu}_{3,12}^*)^2}{\text{var}(\hat{e}_6^*)\,\text{var}(\hat{e}_{12}^*)}$
1	$\frac{(R_{1,1} - \hat{\mu}_{31}^*)^2}{\text{var}(\hat{e}_1^*)\,\text{var}(\hat{e}_1^*)}$	$\frac{(R_{1,2} - \hat{\mu}_{32}^*)^2}{\text{var}(\hat{e}_1^*)\,\text{var}(\hat{e}_2^*)}$	...	$\frac{(R_{1,11} - \hat{\mu}_{3,11}^*)^2}{\text{var}(\hat{e}_1^*)\,\text{var}(\hat{e}_{11}^*)}$	$\frac{(R_{1,12} - \hat{\mu}_{3,12}^*)^2}{\text{var}(\hat{e}_1^*)\,\text{var}(\hat{e}_{12}^*)}$

SSQ is the sum of the all elements within the table for a given j^{th} switch.

[21] N.B. A switch is defined as when a stock is re-allocated from one cluster to another.

Applying this to the data for the initial allocation yields:

Stocks	$\frac{(R_{it} - \hat{\mu}^*_{It})^2}{\text{var}(\hat{e}^*_i)\,\text{var}(\hat{e}^*_t)}$				
	1	2	...	11	12
13	24092.27	4848.71	...	9636.23	8101.13
12	29914.53	1166.44	...	8371.67	25421.28
...	...	...	...	...	...
6	4913.96	65.15	...	519.81	105085.35
1	358.33	4539.85	...	31107.91	14718.75

SSQ = 2,141,067

The SSQ forms the objective function, which is to be minimized through iterative re-allocation of stocks from one cluster to another. In other words:

$$\min SSQ_j = \sum_{t=1}^{T} \sum_{I \in I_j} \sum_{i \in I} \frac{(R_{it} - \hat{\mu}^*_{It})^2}{\text{var}(\hat{e}^*_i)\,\text{var}(\hat{e}^*_t)}$$

By changing: cluster membership

Subject to: cluster membership $\in$ 1, 2, 3 ... *K*
cluster membership $\geq$ 1 (non-negativity, non-zero)

where: *K* = number of pre-specified clusters

Solution sets

Recall that in step 3, clusters were formed via initial allocation, which is based on ranked simple means. Recall also from figure 1, there appear to be three distinct groups of stocks. Stocks 1 to 7 appear to follow a distinct pattern, while stocks 8 to 13 and stocks 14 to 20 also appear to follow a separate pattern.

A priori, a three cluster solution would be ideal. Furthermore, stocks 1 to 7 is expected to be allocated to one cluster while stocks 8 to 13 and stocks 14 to 20 would also be allocated to one cluster each as this would minimize the within cluster variation (cross sectional and time series) given the pattern observed in Figure 4-1.

Under initial allocation, stocks 9, 10, 11, 12, 13 are allocated to cluster 1, while stocks 4, 5, 8, 14, 15, 16, 17, 18, 19, 20 are allocated to cluster 2 and stocks 1, 2, 3, 6, 7 are allocated to cluster 3. This produces an SSQ of 2,141,067.

Figure 4-2 shows the clustering pattern after initial allocation:

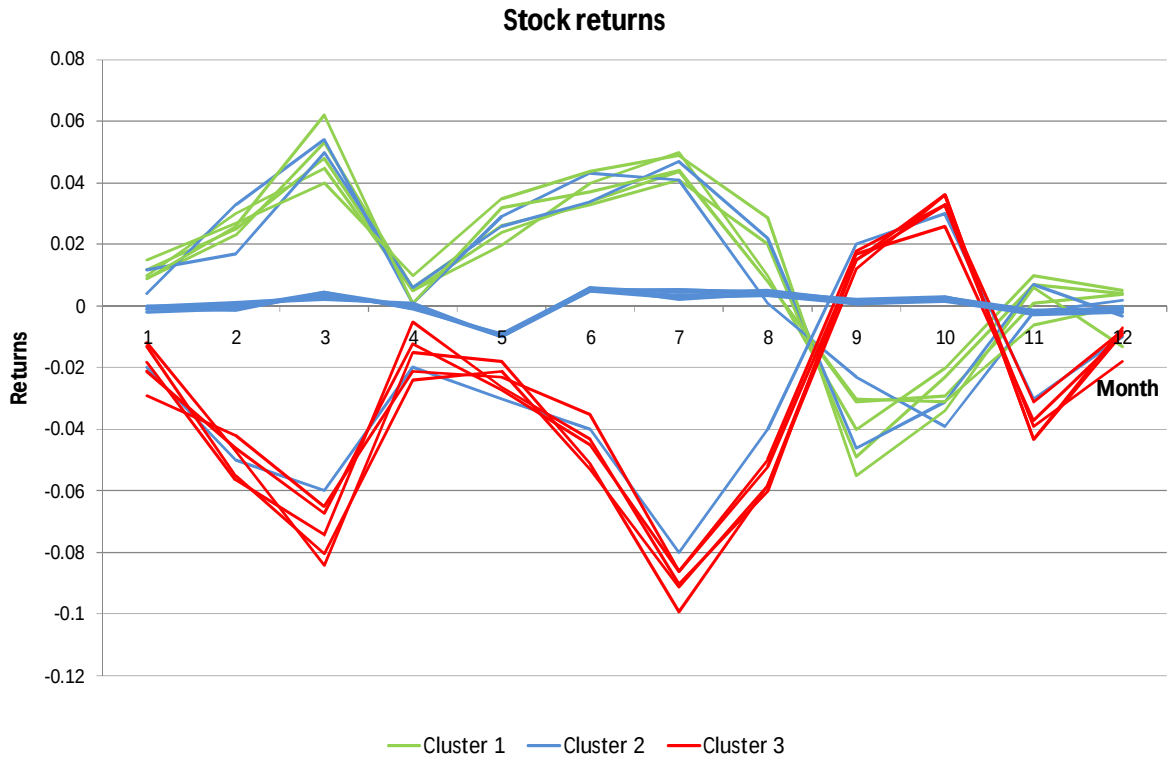


Figure 4-2 Cluster solution after initial allocation

Initial allocation appears to have already clustered stocks roughly in accordance with the *a priori* solution. Cluster 1 primarily consists of the *positive return* stocks, Cluster 2 consists of the *negative return* stocks, with some 'spillover' into cluster 1 and 3; and Cluster 3 primarily consists of the *zero return* stocks. Cluster 2 is the largest cluster with some spillover into clusters 1 and 3. With the exception of the spillover, initial allocation appears to have already performed reasonably well.

SSQ can be further reduced with the final solution set:

Cluster 1: 8, 9, 10, 11, 12, 13

Cluster 2: 14, 15, 16, 17, 18, 19, 20

Cluster 3: 1, 2, 3, 4, 5, 6, 7

SSQ: 1,588,958

Figure 4-3 shows the clustering pattern for the final solution:

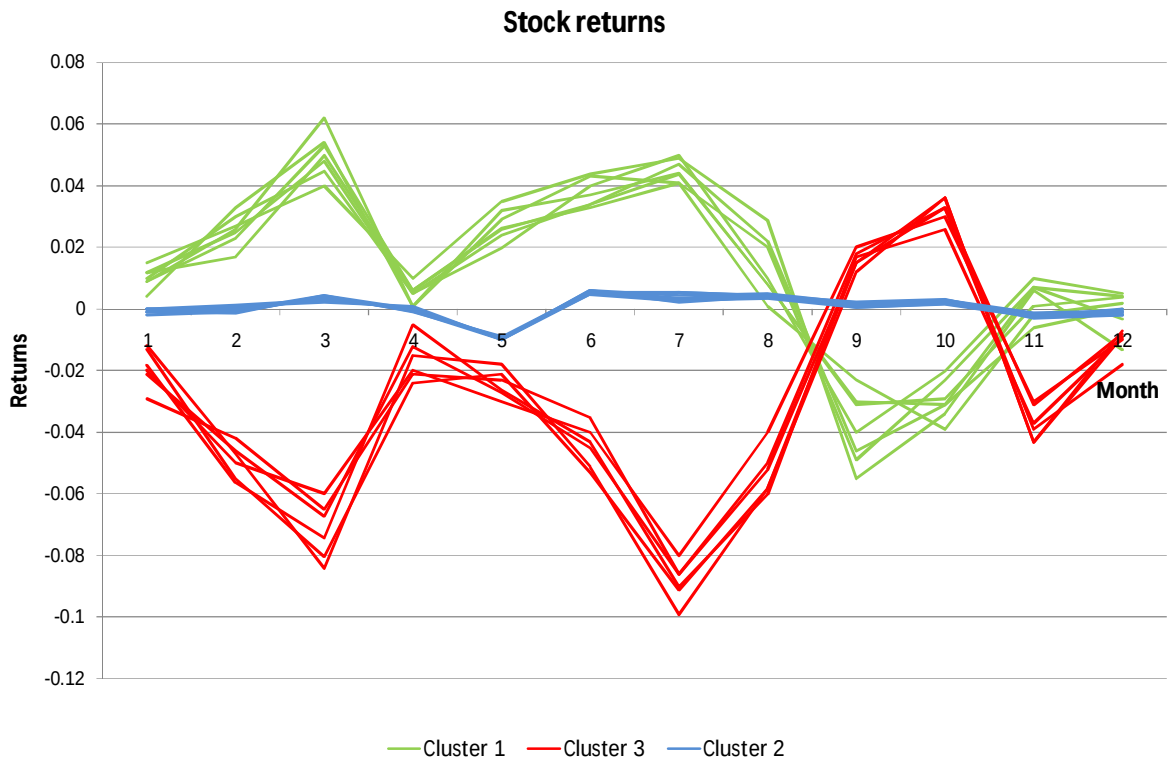


Figure 4-3 Cluster solution after final allocation

The final solution set is consistent with the *a priori* solution. That is, the *positive return* stocks are placed into one cluster, the *negative return* stocks are placed into another cluster and the *zero return* stocks are placed into their own cluster.

This simple example illustrates how the GSC can successfully allocate stocks into separate returns groups. While it is easily apparent from visual analysis of Figure 4-1 that three distinct groups of stocks exist, the final GSC 'result' is not trivial. The purpose of this example is to illustrate how the GSC *would* allocate a group of stocks into separate returns groups. The actual returns data encountered in practice are far greater in volume and complexity however the underlying principles of the GSC remain unchanged. This example explicates how the GSC can take an initial group of seemingly random stocks and place them into distinct and interpretable clusters. Analysis of the real data can take place with confidence that the GSC is able to identify true underlying patterns in the returns data.

This innovative technology has never been directly applied to individual returns data at the firm level. In previous work, the GSC has only ever been applied to mutual funds data in the context of

identifying mutual fund 'styles'. While this work undoubtedly has merits in the area of mutual funds research, no similar attempt has been made to apply it directly to stock returns at the firm level. Individual returns are 'noisy', which makes it hard to reliably estimate the cross section of returns. The GSC overcomes this problem in several ways. Firstly, it forms clusters based not only the cross section of returns but also incorporates the dynamic time series variation. This places stocks into homogeneous risk classes in a way that maximises between cluster heterogeneity while minimising within cluster homogeneity. This means that high risk stocks are placed into one cluster while low risk stocks are placed into another. Secondly, the incorporation of the GLS correction minimises the impact of noisy periods and observations allowing the underlying returns structure to be revealed. Reducing such noise is of itself is a significant achievement and it minimises the distorting effects of extreme observations or periods. In a regression context, this improves model fit, reduces bias and improves the power of statistical tests. In a clustering context, it allows the true underlying clustering patterns to be observed allowing for better cluster interpretation and profiling.

While the many innovative features of the GSC make it a superior clustering methodology[22], it is the application of the GSC that makes it an exciting technology as it has to potential to improve the current state of understanding in the literature in areas such as profiling stock returns and the pricing of risky assets. The GSC algorithm is implemented in Matlab.

4.3 The Gap statistic test

The Gap statistic test developed by Tibshirani, Walther and Hastie (2000) is a procedure to determine the optimal number of clusters in cluster analysis. It essentially formalises the 'elbow' test common in cluster analysis. It does so by generating a null distribution for the total sum of squares, SSQ under the assumption that no clustering patterns exist and then compares the actual SSQ obtained from clustering to find the largest statistically significant difference.

Without the aid of the null distribution, testing for the optimal *K* becomes a subjective task. Consider the scree plot depicted in Figure 4-4.

[22] A superior clustering methodology is defined as one which is able to minimise within cluster dispersion and maximise between cluster dispersion. These results are shown in Chapter 7, specifically Table 7-5 to Table 7-8.

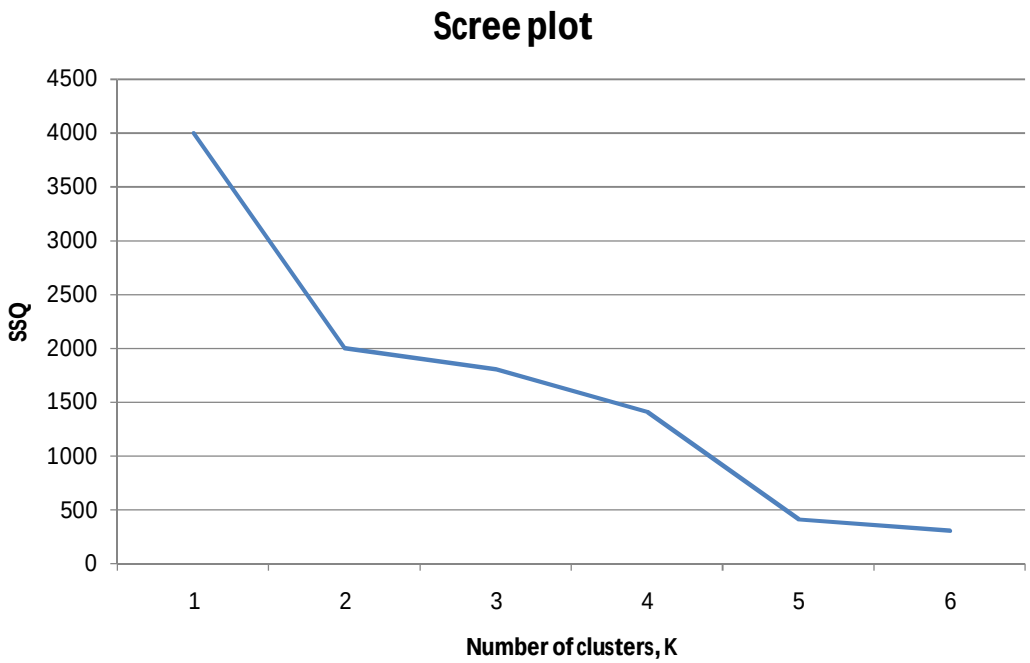


Figure 4-4 Typical scree plot. According to the 'elbow' test, there are two possible values for the optimal number of clusters, *K*. The first occurs at *K* = 2 and the second at *K* = 5

The 'elbow' test searches for the value of *K* which results in a large reduction in SSQ. While the SSQ will always be a monotonically decreasing function of *K*, the intuition is that a large reduction in SSQ indicates a substantial improvement in the cluster solution compared to neighbouring values of *K*.

In this example, this elbow occurs at two possible values: *K* = 2 and *K* = 5. Without any formal testing procedure to determine the optimal value of *K*, determining this value becomes a matter of subjective judgement.

One way to formalise the test would be to compare the SSQ against a null series. To generate the null series, the dataset is randomly re-sampled and the clustering algorithm is applied.

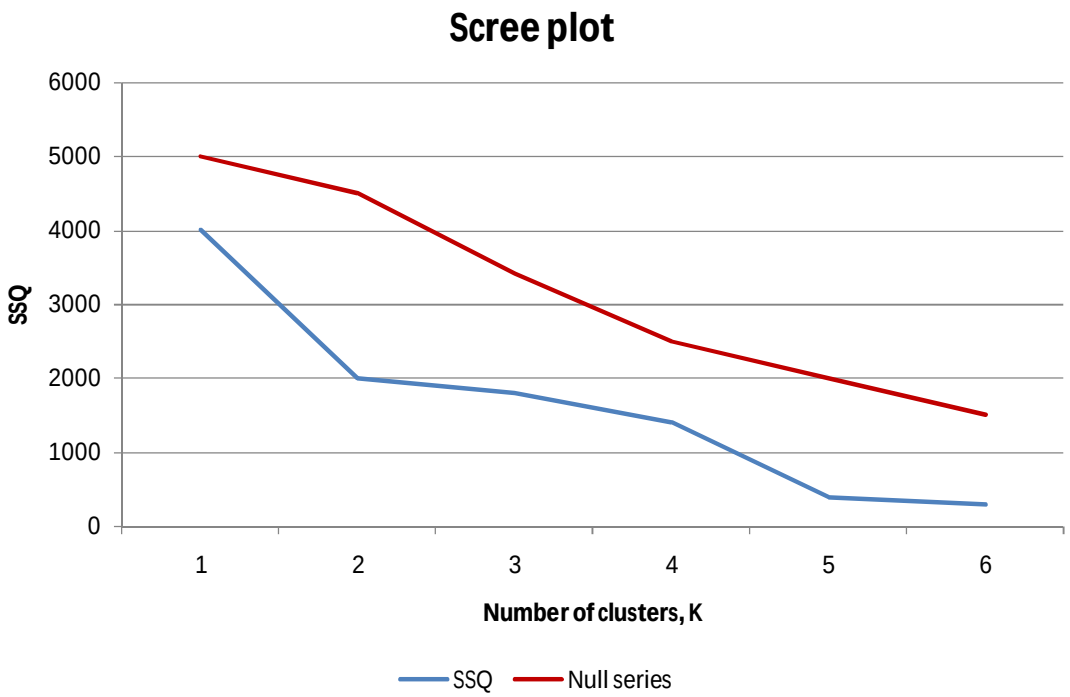


Figure 4-5 Scree plot with a null series. The null series represents the values of SSQ that would occur under random assignment of observations to clusters.

By comparing the null series to the actual SSQ, it is apparent that the largest difference between the null series and the SSQ occurs at $K = 2$ indicating that a 2 cluster solution would be superior to a 5 cluster solution.

While $K = 2$ results in the largest difference between the null series and the SSQ, there is no guarantee that this is a statistically significant difference. In order for a statistically significant difference to exist, the value under investigation (or test statistic) must be compared against a null distribution.

One way to generate this distribution would be to repeatedly resample the data under investigation to derive a set of reference datasets and then apply the clustering algorithm to these reference sets to form clusters. Figure 4-6 depicts how such a distribution would appear.

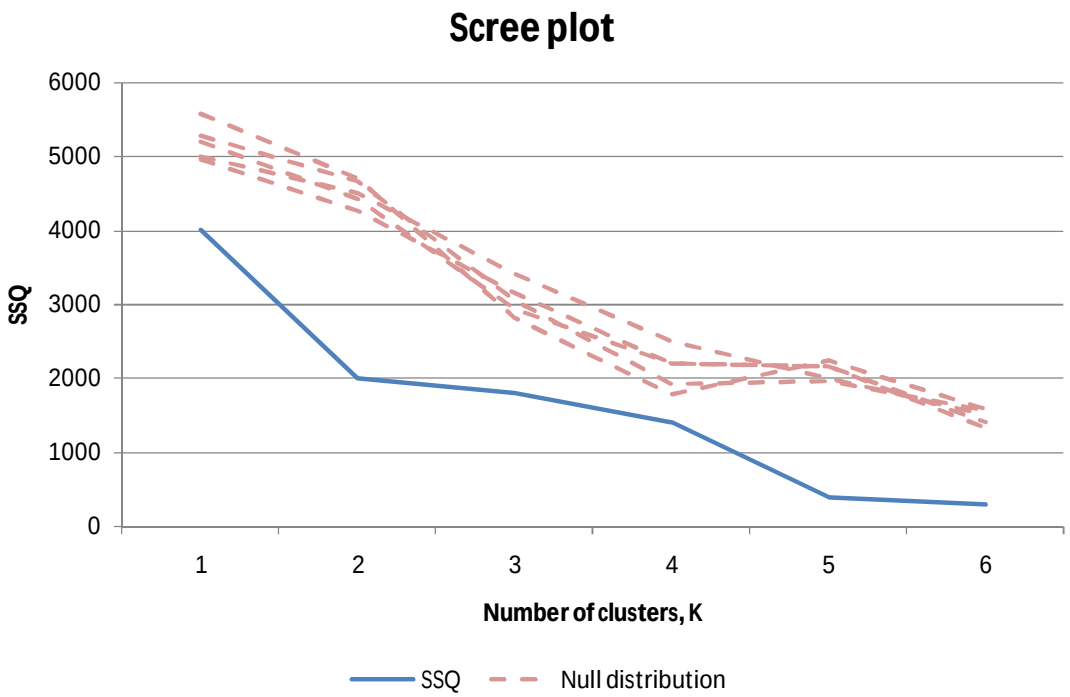


Figure 4-6 Scree plot with a null distribution. The null distribution represents the values of SSQ that would occur under random repeated re-sampling of the null series.

A null distribution is generated from random repeated re-sampling of the data. The SSQ falls far below the null distribution indicating a statistically significant difference. Furthermore, the largest statistically significant difference occurs at $K = 2$ indicating that a 2 cluster solution is optimal. This simple example illustrates the underlying principles of the Gap statistic test.

Formally, the Gap statistic test operates by first calculating the squared Euclidean distance between observations. Following the notation of Tibshirani et al, consider a dataset given by $\{x_{ij}\}$ with n observations and p characteristics, i.e. $i = 1, 2 \ldots n$ and $j = 1, 2 \ldots p$. The squared Euclidean distance, $d_{ii'}$ between observation i and i' is given by:

$$d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$$

Suppose that k clusters are estimated. Let $C_1, C_2, \ldots, C_k$ with C_r denote the indices of observation in cluster r. Therefore, the pairwise distances for the r^{th} cluster is given by:

$$D_r = \sum_{i,i' \epsilon C_r} d_{ii'}$$

The pooled within cluster sum of squares is given by:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

The central premise behind the Gap statistic test is to compare $log(W_k)$ against the expectation of a null distribution for all k clusters between *k* = 1, 2, … *K*. Thus the actual Gap statistic itself is calculated by:

$$\begin{aligned} Gap_n(k) &= E_n^*\{log(W_k)\} - log(W_k) \\ &= \frac{1}{B}\sum_b\{log(W_{kb}^*)\} - log(W_k) \end{aligned}$$

E_n^* denotes the expectation under a sample of size *n* from the reference distribution which is randomly drawn via Monte Carlo simulation. $E_n^*\{log(W_k)\}$ is the average of B copies of $log(W_k^*)$, each of which is computed from a Monte Carlo sample $X_1^*, \dots, X_n^*$ drawn from the reference distribution.

The Gap statistic procedure is to find the value of *k* for which $log(W_k)$ falls farthest below the null distribution curve as this maximizes the Gap statistic. Therefore, the actual Gap statistic 'test' itself is to determine the value of k which maximises $Gap_n(k)$ after accounting for the sampling and simulation error in drawing the expectations. That is, the optimal occurs when:

$$Gap(k) > Gap(k+1) - s_{k+1}$$

Where s_{k+1} represents the sampling and simulation error from an additional cluster

Since the expectations are randomly generated, the sampling error must be taken into account when determining the optimal *K*. s_k is calculated by taking into account both the sampling and simulation error in drawing $E_n^*\{log(W_k)\}$. It is calculated thus:

$$s_k = sd_k\sqrt{1 + \frac{1}{B}}$$

sd_k represents the standard deviation of each cluster, computed to account for the sampling error, which is calculated by:

$$sd_k = \sqrt{\frac{1}{B}\sum_b \left(log(W^*_{kb}) - \bar{l}\right)^2}$$

Where $\bar{l} = \frac{1}{B}\sum_b log(W^*_{kb})$

Figure 4-7 explains the steps in the Gap statistic test:

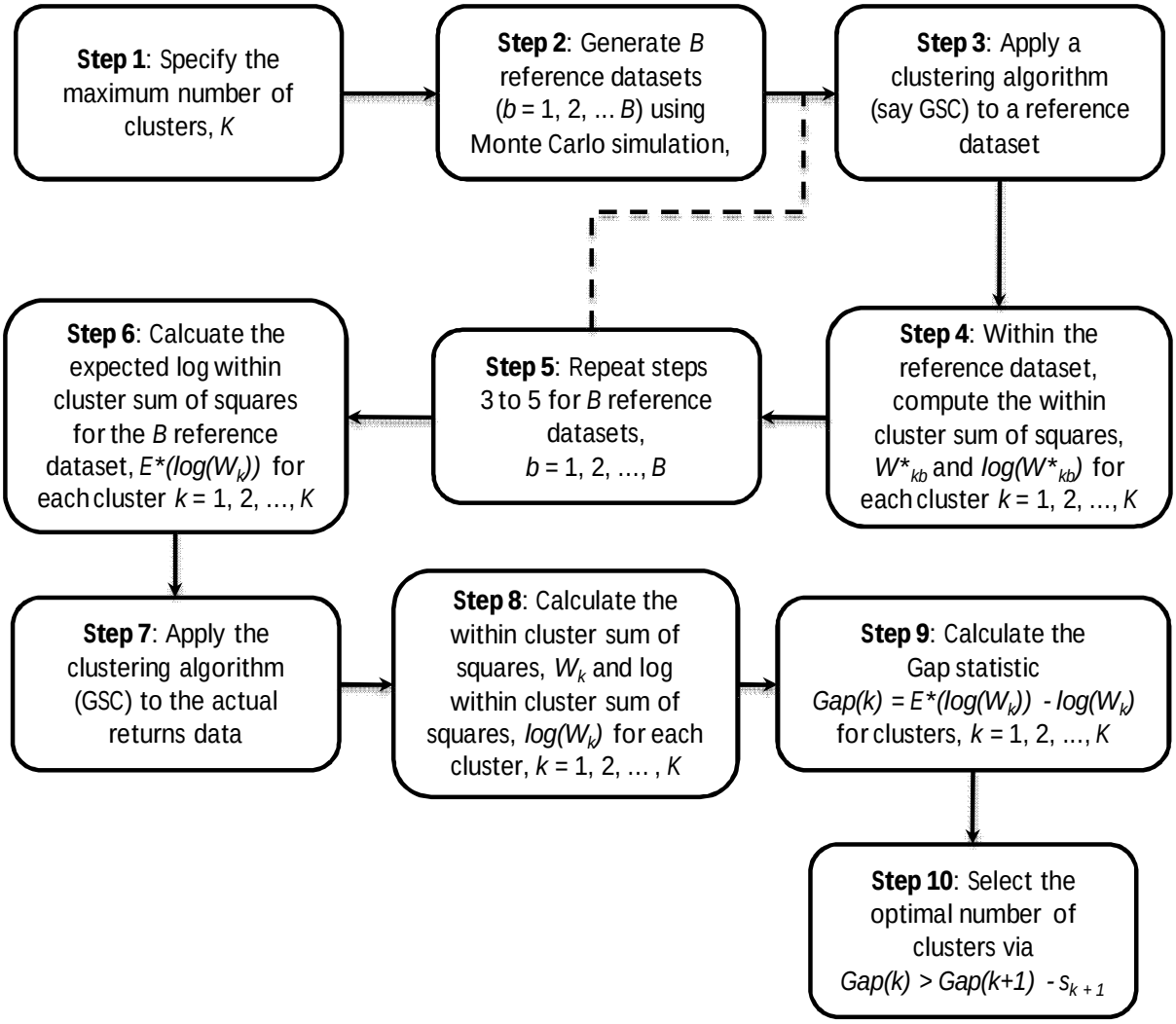


Figure 4-7 Flowchart indicating the steps of the Gap statistic test.

The Gap statistic test provides a powerful way for researchers to objectively determine the optimal number of clusters, *K*. As previously mentioned, selecting the number of clusters is a key challenge in any k-means cluster analysis. An improperly specified number of clusters will worsen the cluster solution as elements are not properly separated into distinct clusters preventing meaningful interpretation of those clusters. Unfortunately, researchers are often left with few truly objective means to determine the optimal *K* and instead rely on subjective assessment. The Gap statistic test solves this problem by developing a null distribution which is based on repeated random re-sampling of the data. Researchers also have control over how many reference sets to form thus improving the power of statistical tests. Furthermore it can be applied to any clustering algorithm making it extremely flexible and adaptable.

The only real drawback of the Gap statistic is the computational requirements of repeated applications of the clustering algorithm to develop the null distribution. However with improvements in computational resources and careful management of the data, these issues are mitigated.

These issues aside, the Gap statistic test represents an elegantly simple yet highly effective means to resolve a central issue which historically has been a major problem in k-means cluster analysis. When combined with the GSC, which already possesses many desirable statistical properties, the Gap statistic test provides researchers with a potent tool for studying the clustering patterns of stock returns or any financial variable of interest for that matter. Like the GSC algorithm, the Gap statistic test is also implemented in Matlab.

4.4 Innovative clustering of stocks by industrial sector

The GSC-Gap combination provides researchers with an innovative way to study stock returns in a unique way, which has never been attempted in the literature. One way to apply this innovative technology is to cluster and profile individual stocks at the firm level. To cluster a group of stocks by returns implies that the stocks within the cluster share a similar profile in terms of returns. There are many reasons why researchers may wish to form such homogenous groups (see Chapter 2).

Such applications can be improved by a better understanding of the universe of stocks and their behaviour. Toward this end, the GSC is applied to the universe of stocks (post Fama-French filters) and ‘profiled’ on the basis of industrial sector.

In Chapter 6, the GSC is applied to the returns data to identify clusters. Individual stocks are then grouped by their respective clusters and their risk/returns pattern is profiled against industry membership as indicated by an industry classification scheme.

A key step therefore is the choice of industry classification scheme. The Standardised Industry Classification (SIC) scheme managed by the United States Office of Management and the Budget is a commonly used industry classification scheme. Despite the many shortcomings of the SIC scheme, it is still the most commonly used industry classification scheme in the literature (Clarke, 1989; Bhojraj, Lee and Oler, 2003; Kahle and Walking, 1996). Given its established use, the SIC scheme is selected as the industry classification scheme for profiling clusters.

Note that the SIC codes are used only to profile the GSC clusters not in the actual formation of the clusters themselves. The GSC clusters are formed independently of the SIC codes. Other classification schemes such as NAICS, GICS of FF may also be used for profiling however additional profiling based on these other schemes are beyond the scope of this research.

The GSC clusters are then 'matched' to the SIC categories by via maximum squared rank correlation. That is, if an industry group mean return correlates highly with a GSC cluster group mean return, then the industry group will be matched to the GSC cluster.

To visualise the GSC clusters and their relation to certain industry groups, a perceptual mapping technique known as MultiDimensional Scaling (MDS) is used. MDS is a visualisation technique that takes a matrix of similarity of dissimilarity measures between elements and calculates a set of Cartesian coordinates where the distances between elements are scaled based on the degree of similarity or dissimilarity. Elements which share a high degree of similarity will have be separated by a short distance while elements which are dissimilar will be separated by a long distance. The elements are then charted onto *N* dimensional space (usually 2) where the distances between elements reflect their degree of similarity or dissimilarity. Thus, elements that are 'close' to each other in the visual space do in fact share a high degree of similarity and the opposite is true for elements which are 'far' from each other (Hair et al).

In the current study, the matrix of squared rank correlation coefficients between the cluster means and industry group means is used as a measure of similarity. Higher numbers indicate greater correlation between the clusters and a given industry while lower numbers indicate the opposite.

Thus when the clusters and industry groups are charted onto 2-dimensional space, industries that are close to particular clusters indicate a high degree of correlation. Note because of the way the squared rank correlation matrix is set up, this can only happen if a cluster correlates highly to an industry and the industry correlates highly to the cluster (See Section 6.2.2). The resultant MDS charts thus provide a useful visual aid for identifying and profiling the GSC clusters on the basis of industry. Note however that the MDS chart is just a visual aid which in essence provides a visual counterpart to the squared rank correlation matrix. It is not solely relied upon for the purposes of cluster interpretation.

To determine the optimal *K*, the Gap statistic test is used, which as previously discussed, represents an objective way to determine the optimal level of *K*.

The result is a set of representative clusters that can be readily interpreted on the basis of industrial and economic sector of operation. These industry clusters can be further profiled on the basis of risk/return to not only identify the underlying sectors of the economy but also the kind of investment prospect they represent to investors. To measure risk, standard deviation of the time series of industry average returns is used. To measure return, the average of the industry average returns is used. Approximately 10 clusters are identified in each modelling interval[23] and these clusters represent various sectors of the economy.

This includes sectors such as primary resources, which deal mainly with the extraction and refinement of primary resources such as oil, gas and petroleum. This sector typically experiences high-risk but generates medium-high returns. This sector is exposed to a number of risks related to exploration and extraction as well as economic factors such as fluctuations in commodity prices, exchange rate risk, changes in political climate and external risk. The cluster means capture these sources of risk which are inherent to the operational environment of the sector. Other sectors identified include Utilities which have low risks and generate low-moderate returns. The proximity of

[23] Chapter 5 explores issues relating to the data period and modelling interval. The data period refers to the entire length of the data used in this study from start to finish. However, within the data period, the data are further subdivided into discrete intervals referred to as modelling intervals. This is done to allow a degree of flexibility in the clustering solutions so that stocks are not locked in to a particular cluster across all time. By contrast, stocks are allowed to move between clusters as their returns profile changes over time.

this sector to the Government and the regulated nature of the market provide a stabilising effect to this sector. Again, these risks (or lack thereof) are captured in the cluster means which are inherent to the operational environment of the industry. Sectors such as Elaborately Transformed Manufactures (ETMs) experience high risk but also generate high returns. ETMs represent a major export sector in the U.S. and as such are exposed to external risk, exchange rate risk and changes in foreign economic conditions. Furthermore, many of the firms in this sector are involved in the production of high-tech goods and as such require heavy investment in research and development. By contrast, sectors such as basic manufacturing have less risk but generate lower returns. Basic manufacturing tends to be isolated to the production of 'basic' goods such as primary metals, rubber and plastics and paper and allied products.

These findings are all consistent with the risk-return paradigm in finance. The difference between this approach and those advocated by Fama French is that no attempt is made to impose an arbitrary interpretation of risk that is applied broadly across all stocks in all industries.

The various sources of risk are complex, vary on a case by case basis and are unique to the operational environment of the industry. The GSC clusters capture these unique sources risk and interpretation needs to be made on a case by case basis.

4.5 Towards better cost of capital estimates

The cost of capital is an integral component in many financial applications. For example, it is commonly used as a discount rate in Net Present Value (NPV) calculations. The cost of capital represents the cost to a firm of raising capital. Firms raise capital in one of two ways: through issuing stock (equity) or by assuming debt. The cost of raising this capital is the return to stockholders (cost of equity) in the case of equity or interest payments in the case of debt. The Weighted Average Cost of Capital (WACC) takes into account the cost of both forms of raising capital. It is weighted to take into account the relative proportions of equity or debt that a firm raises. For example, if a firm raises 90 percent of its capital through stock, then the WACC will be heavily weighted toward the cost of equity.

While the cost of debt is relatively easy to observe and measure, the same is not true for the cost of equity. Investors buy equity from firms with no guarantee of return. The cost of this equity must therefore be estimated. Estimating the cost of equity however is far more complex. If successful, it

has the potential to improve WACC calculations, which in turn can improve NPV calculations; and this leads to better allocation of funds and ultimately profit maximisation.

Although the cost of capital is firm and project specific, industry costs of capital may be used as benchmark rates. Firms operating within a given industry may turn to these benchmark rates in evaluating/comparing their individual costs of capital relative to other firms operating in the same industry.

The asset pricing tests in Chapter 7 are related to the cost of capital in the following way: The cost of capital or more specifically, the weighted average cost of capital is based on the cost of debt and the cost of equity. Estimating the cost of debt is relatively straightforward however estimating the cost of equity is a far more contentious issue in the literature. A common approach is to quantify the relationship between returns (cost of equity) and firm characteristics, which are then interpreted as representing various dimensions of risk. The results in Chapter 7 indicate that the GSC cluster means are able to explain a high proportion of the cross section of returns. Costs of equity estimated with lower variability lead to more precise cost of capital estimates and potentially better budget allocations.As different industries are exposed to different sources of risk, their cost of capital must vary. For example, a mining firm is exposed to greater risk than say a utilities firm. Therefore, two cost of capital estimates are required – one for the mining industry and one for the utilities sector.

Again, common industry classification schemes may be used to form industry groups. Industry average returns may then be used to explain the cross section of stock returns via:

$$R_{it} = \alpha + \beta R_{It} + \varepsilon_{it} \qquad (4.1)$$

Where R_{it} = Return of I^{th} stock in period t belonging to I^{th} category

R_{It} = Average return of I^{th} category in period t

The problem with industry averages formed via industry classification schemes is that it is unclear how these industry groups are formed. As such, industry averages can only explain up to 30 percent of the cross section of stock returns (Bhojraj et al). This is because common industry classification schemes do not group firms based on returns but rather economic activity. The GSC however does group stocks based on returns. For many financial applications however, it matters less that two firms share similarity in economic/business activity but rather whether they share similarity in returns (Ritter, 1991; Spiess and Affleck-Graves, 1994; Hendricks and Singhal, 2001).

The GSC however does group firms on the basis on returns providing a solution to the ongoing problem in the literature of not being able to form return homogeneous groups. Rather than calculating industry averages, the GSC cluster averages may be used to estimate Equation 4.1. As evidence of the overall superiority of the GSC, the GSC cluster averages explain up to 60 percent of the cross section of returns. In some cases, there is outperformance by a factor of 2 to 3. Since the clusters themselves can be interpreted as industries (Section 4.4), the GSC provides a vastly superior alternative to common industry classification schemes for the purpose of calculating industry cost of capital.

Out of sample testing is also performed to test the robustness of the GSC. To perform out of sample testing, a proportion of the data is randomly removed, henceforth referred to as the validation set. The remaining data, henceforth referred to as the estimation set is used to re-calculate the industry/cluster averages and the following equation is estimated:

$$R_{it}^{*} = \alpha + \beta R_{It} + \varepsilon_{it} \qquad (4.2)$$

Where R_{it}^{*} = Return of ith stock in the validation set in period *t* belonging to Ith category

R_{It} = Average return of Ith category in the estimation set in period *t*.

Again the out of sample performance for the GSC is high compared to common industry classification schemes indicating robustness in the GSC. This outstanding result is evidence that the GSC is better able to form the returns homogeneous groups so desperately needed in the literature (Miller and Modigliani, 1966; Litzenberger and Rao, 1972; Fama French, 1997; Boness and Frankfurter, 1977; Chan et al, 2007; Rapach, Strauss, Tu and Zhou, 2010; Asness, Porter and Stevens, 2000). Such returns homogeneous groups have the potential to revolutionise current approaches to estimating industry costs of capital leading to improved and more precise estimates.

At its core, this is an exercise in calculating the cost of equity. The cost of equity is an integral component in cost of capital estimates. The high adjusted R^2 obtained from estimating Equation 4.1 indicates that the GSC based approach will lead to better cost of capital estimates.

4.6 Conclusion

The central technology in this research is the Generalised Style Classification (GSC) algorithm used to cluster individual stock returns at the firm level. The GSC has many innovations which reduce the impact of noise such as the ability to integrate both cross sectional and time series variation as well as GLS correction to minimise the impact of volatile periods and extreme observations. Although historically applied to the mutual funds data, the GSC has never been used on individual stock returns at the firm level making this research a novel contribution to the literature.

The addition of the Gap statistic further improves the effectiveness of the GSC. The Gap statistic test is an objective method to determine the optimal number of clusters which is a common issue in any form of k-means cluster analysis. It does so by drawing multiple reference sets and applies the clustering algorithm to derive the within cluster sum of squares to form a null distribution. The actual within cluster sum of squares is compared against the expected within cluster sum of squares from the null distribution and the optimal *K* is reached at the point where there is the largest statistically significant difference after accounting for sampling and simulation error. Again, such a GSC-Gap combination has never been implemented in the literature.

The GSC-Gap combination is applied to the universe of stocks to group individual stocks into clusters. These clusters are then matched against SIC-delineated industry groups and profiled on the basis of industry. This matching is performed on the basis of squared rank correlations. The Gap statistic test is used to determine the optimal number of clusters. MDS perceptual mapping is also used to help visualise the clusters and their industry groups. The results show that various industry groups exist and these groups are highly distinct and separate to each other in terms of risk and return.

The second application is in the area of corporate finance. The cost of capital is an important variable used in many financial applications. One such application is in calculating NPV. The cost of capital is based on the cost of equity and the cost of debt. Industry costs of capital represent the cost of raising capital for different industries. The cost of equity however is difficult to estimate. Industry costs of capital in particular are difficult to estimate because the industry classification schemes used to group the stocks are not based on returns but rather economic activity. The GSC however is used as an alternative and explains up to 60 percent of the cross section of returns which represents outperformance by a factor of 2 – 3 when compared to common industry classification schemes.

Both these applications represent a novel contribution to the literature by virtue of the fact that they are driven by the GSC-Gap combination which has never been used in such a way. The success of this combined approach is a testament to the effectiveness of these innovative procedures.

Chapter 5

5 Determining the appropriate Data period and Modelling interval

5.1 Introduction

Two crucial steps before any modelling can be performed are firstly to determine the appropriate data period and secondly the modelling interval. Here, the data period refers to the entire period under investigation. Data on the merged CRSP/COMPUSTAT database is available from the early 1960s onwards. However, there is very little data available in the early periods. The modelling interval refers to the individual sections of the data period. For example a 5 year modelling interval means the data will be divided into 5 yearly sections.

The rationale for determining the appropriate data period is to ensure sufficient data is available, which is especially important given the substantial amount of data removed from the Fama-French filtering conditions and the lack of available data in the early periods.

The rationale for determining the appropriate modelling interval is to allow some degree of flexibility in the clustering procedure. Under the GSC approach, once a stock is allocated to a cluster, it remains there for the entire modelling interval. This would not be a problem if every firm's industry membership and the very definition of an industry itself remained constant throughout time. This assumption however is invalid in practice. Consider the case of conglomerate firms or firms which move across industries over time. The Dunlop company for example, at various periods in time has diversified operations in tyre manufacturing, footwear and clothing, aerospace, sporting equipment and even contraceptive goods[24]. Recall that according to our theory, the clusters in fact represent industry or industry groups. Allocating a firm such as Dunlop to a particular cluster may only be valid in one particular period but less so in other periods when the firm migrates to another industry. If the modelling interval were too long, say the entire data period, firms such as Dunlop may be forced to remain in one industry even though they may potentially migrate across industries thus leading to misallocation and potential bias. Since the data is available in monthly frequency, one extreme position would be set the modelling interval to 1-month for maximum flexibility. However doing so would result in an insufficient number of observations and more importantly increase noise which is especially troublesome given the volatility of stock prices and the returns

[24] They were involved in the production of condoms.

being modelled[25]. This noise would inflate the time series variances of returns which are used as weights during various stages in the GSC algorithm ultimately reducing the ability of the GSC to form clusters and allocate stocks accordingly.

Furthermore, various industries change substantially across time. The information and communications technology sector for example has evolved from simple communications to include various information and technology services in recent periods. Having a 'long' modelling period may not allow the GSC enough flexibility to accommodate such emerging and constantly evolving industries.

So a trade-off exists between having a long modelling period which allows for data stability and noise reduction and a short modelling period which allows greater flexibility in clustering allocations. Due to the severe consequences of an improperly specified modelling interval, further investigation to determine the appropriate interval is necessary.

[25] In any case, the GSC would not operate under such conditions as it calculates time-series as well as cross sectional standard deviations and uses these as weights during various stages of the clustering. Having a 1-month interval would result in a zero time series standard deviation and subsequent division / multiplication by zero.

5.1.1 Determining the appropriate data period

The CRSP/COMPUSTAT database offers data from the 1960s onwards.

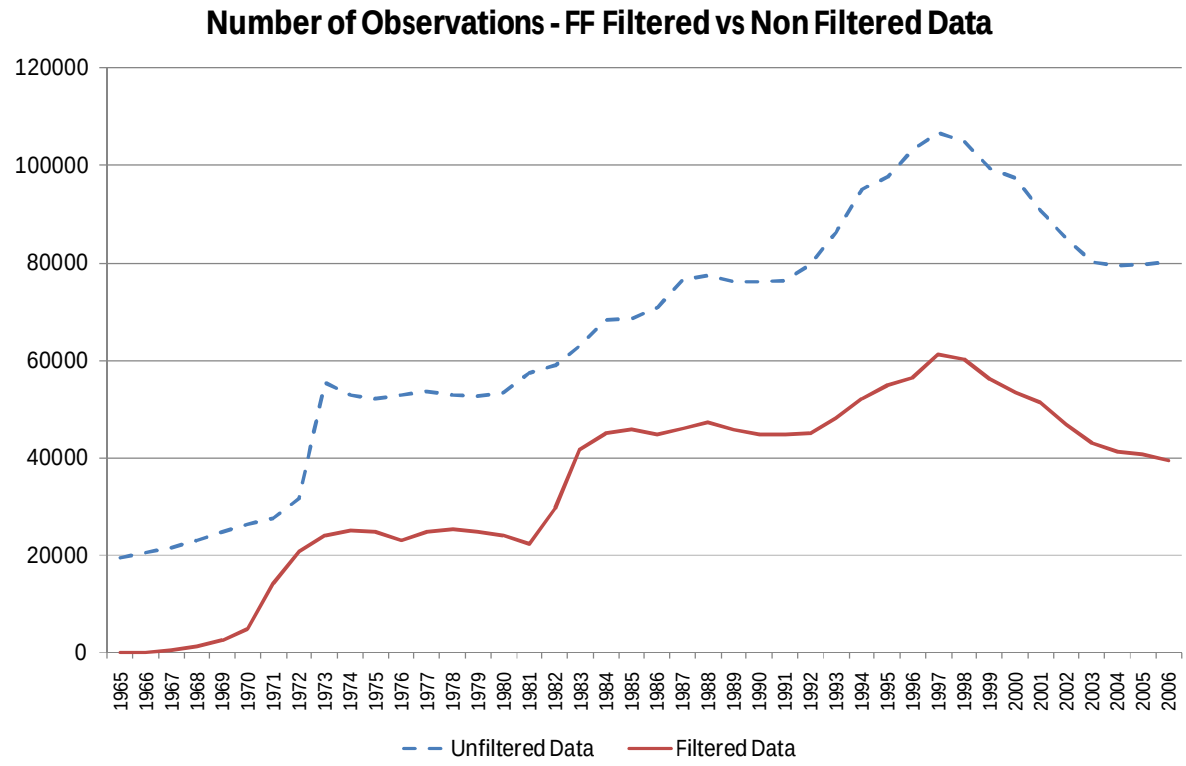


Figure 5-1 This figure displays the number of observations available from the CRSP/COMPUSTAT database between 1965 and 2006 before and after the application of the Fama-French (1992, 1993) filtering conditions. The dotted line represents the number of observations from the unfiltered data while the solid line represents the number of observations from the filtered data.

As Figure 5-1 indicates, the Fama French (1992, 1993) filtering conditions remove a substantial proportion of the data (approximately 40 to 50 percent on average). There is a sharp increase in the number of observations in the early 1970s (probably due to the incorporation of the NASDAQ) to approximately 20,000 observations and another increase in the early 1980s to about 45,000 observations in the filtered data.

Due to the large increase in the number of observations and relative stability afterwards, 1984 was selected as the start of the data period. At the time of writing, only data up to 2006 was available. Therefore, the selected data period is 1984 to 2006.

5.1.2 Determining the appropriate modelling interval

A common approach in many heuristic methods is to firstly define some kind of objective function or measure of performance. This performance measure is then calculated under various permutations of the decision variable(s) until either an optimal performance measure is reached or no further changes/improvements to the performance measure can be achieved with additional incremental changes of the decision variable.

Here the decision variable of interest is the length of the modelling interval. Given the nature of the clustering algorithm, an appropriate performance measure would be the sum of squared errors, SSQ. In the GSC context, an error is defined as the deviation of a return from its cluster mean in a given time period. The sum of such squared errors is therefore a measure of performance. In a 'good' cluster solution, within cluster homogeneity is maximised while between cluster homogeneity is minimised (or between cluster heterogeneity is maximised). In such cases, the SSQ would be relatively low. For further details on how the SSQ is derived, refer to the Section 4.2.

So a heuristic approach to determining the appropriate modelling interval would be to compare the SSQ obtained from different interval lengths. The SSQ will always increase with greater interval lengths. As the interval length is increased, more observations/data enter the solution space increasing the number of pairwise deviations inevitably inflating the SSQ. To overcome this problem, the SSQ may be 'standardised' by dividing by the number of observations. Therefore, the optimal interval length will be reached when the SSQ per Observation (analogous to MSQ) ceases to change with additional increments of interval length – effectively reaching a 'steady state'.

Beginning at 1984 and an initial modelling interval of 12 months, the modelling interval was increased at 1 month increments. At each increment, GSC cluster analysis was performed and the SSQ per observation was recorded. To ensure that the optimal modelling interval was not biased by the choice of K, the analysis was repeated for K = 5, 10 and 15. The results are reproduced in Figure 5-2.

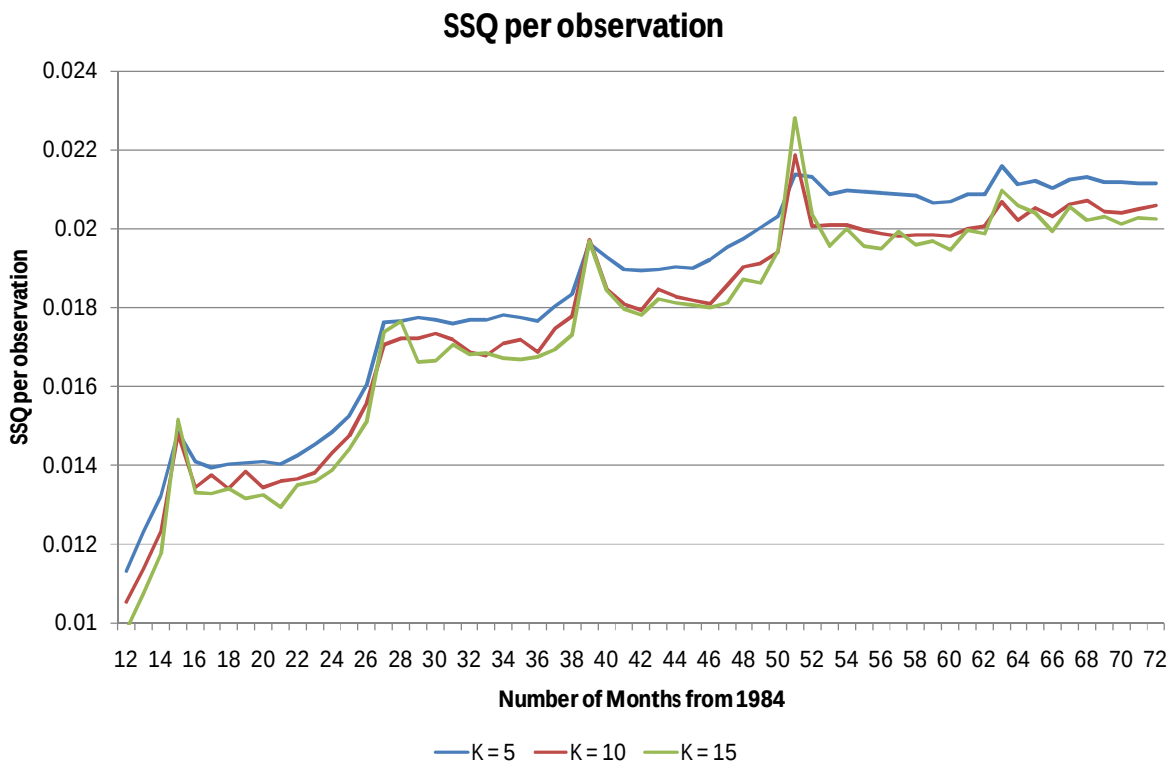


Figure 5-2 This figure shows the SSQ per observation as the data is increased at 1 month increments beginning at 1984 for K = 5, 10 and 15.

There are several noteworthy points. Firstly, a clear pattern of diminishing returns exists indicating less improvement to the SSQ per observation with longer modelling intervals[26]. From this one may conclude that additional months in the modelling interval beyond some point are redundant. Secondly, there are several "plateaus" in the SSQ per observation. This first occurs after approximately 28 months and the second after approximately 54 months. Furthermore, very little improvement is observed after 60 months. From this, one may conclude that the appropriate modelling interval should not exceed 5 years. Given the first two plateaus, there are two potential options for the appropriate modelling interval: 3 or 5 years. The change in levels however is only minor between 36 and 60 months when compared to the change between 12 and 36 months indicating only minor improvements to the cluster solution with the longer modelling interval of 60 months. Given the minor improvement with the longer modelling interval and the need to maintain flexibility with shorter modelling intervals as discussed in Section 5.1, the shorter option of 36 months will be chosen as the appropriate modelling interval.

Although the issue of determining the optimal K will be explored in more detail in later chapters, at this early stage, it may be possible to determine an approximate range for *K*. As *K* increases, the SSQ

[26] Note also that the Y-axis has been scaled to start at 0.01 not zero.

per observation will decrease as there are more cluster centres and provided these cluster centres are well dispersed, pairwise deviations from cluster means will be smaller thus reducing the overall SSQ. However, using the same heuristic intuition, improvements to the SSQ will be lower with additional increments of *K* indicating an optimal solution. The overall reduction in SSQ per observation between *K* = 5 and 10 is dramatic while the overall reduction between *K* = 10 and 15 is marginal by comparison indicating that additional clusters beyond 10 are potentially redundant. The optimal *K* therefore lies between *K* = 5 and 10. Although preliminary, this is clear evidence that the optimal *K* lies between 5 and 10. This is a significant finding as one of the main issues in k-means cluster analysis is determining the optimal number of clusters. A more objective test to determine the optimal number of clusters is presented in Chapter 6.

To ensure that the optimal modelling interval of 36 months is not biased by the choice of the start period, 1984, the starting period was varied between 1984, 1985, 1986, 1987 and 1988. The procedure was repeated and the same statistic, the SSQ per observation was averaged across the 5 sets of results. The average SSQ per observation across these 5 sets are reproduced in Figure 5-3.

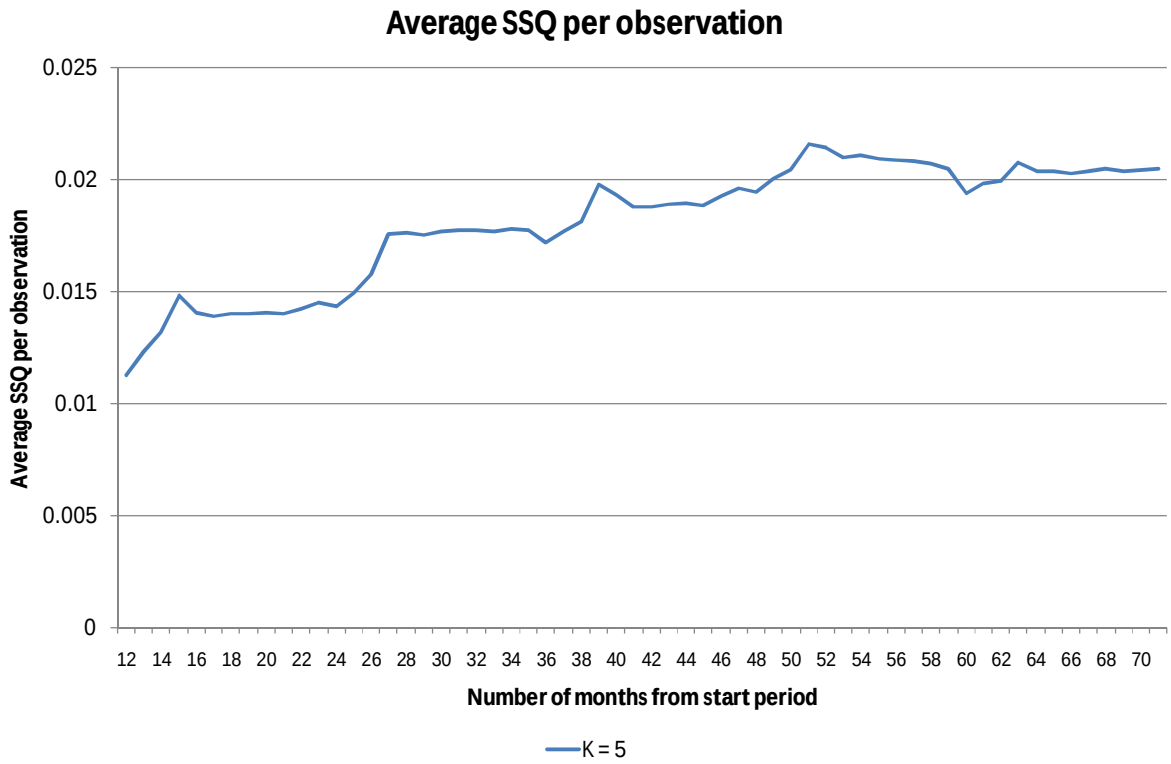


Figure 5-3 The starting period was varied between 1984, 1985, 1986, 1987 and 1988. The GSC was applied to the 5 datasets and the SSQ per observation was calculated. This figure indicates the average SSQ per observation. Although the increase in SSQ per observation is less dramatic between 12 and 28 months, there still appears to be a plateau at 28 months and another after 54 months which is the same pattern observed in Figure 5-2.

As Figure 5-3 indicates, the average SSQ per observation still plateaus once after 28 months and again after approximately 54 months indicating once again two potential choices for the optimal modelling interval: 3 or 5 years. However, to maintain some degree of flexibility, 36 months is chosen as the optimal modelling interval. Furthermore, the variation in the starting period indicates that this result is not affected by the choice of start period.

5.2 Conclusion

Before data modelling can proceed, two key issues must first be resolved: the appropriate data period and the modelling interval. Given the substantial increase in the number of observations after 1984, data post 1984 provides a rich source of data for modelling. The selected data period is therefore 1984 to 2006.

By adopting a heuristic approach, a modelling interval of 36 months was selected. This interval is 'optimal' in the sense that very little change in the SSQ per observation is observed beyond 36 months effectively reaching a 'steady state'. Furthermore the 3 year modelling interval was chosen over the longer 5 year option to allow greater flexibility in stock allocation to clusters.

These important findings will help ensure the integrity of modelling results presented in later chapters. Selecting a data period that is rich in data has many desirable statistical properties such as minimising small sample bias, ensuring all members of the population are represented and providing power to statistical tests among others.

Selecting an 'optimal' modelling interval is also important as it helps to balance two key factors. Having a shorter modelling interval allows more flexibility in the sense that firms are not locked into industry sectors throughout the entire data period and also allows for structural shifts in the industry sectors themselves. On the other hand, a longer modelling interval ensures data stability (sufficient observations) and noise reduction, which is important given the volatility of financial data. The modelling interval of 36 months will balance these two considerations by allowing enough flexibility without sacrificing data stability and noise reduction.

Chapter 6

6 Innovative Clustering of Stocks by Industrial Sector

6.1 Introduction

In the previous chapter, issues relating to the appropriate data period and modelling interval were discussed. The consequences of an improperly specified data period and modelling interval range from; estimation bias, to low power on statistical tests, to complete failure of the clustering algorithm to form clusters and allocate elements correctly. Now that these issues have been addressed, focus may shift to the actual cluster analysis and interpretation of those clusters with some degree of confidence that the results are free from the aforementioned errors.

This chapter explores the application of the GSC clustering algorithm on the returns data to derive clusters and obtain interpretations of those clusters in the context of industrial sectors. Recall that the pioneering work of King (1966) found that up to 20 percent of the variation in an asset's cross sectional returns could be explained by industry factors. The ability of the GSC clusters to explain the cross section of returns is presented in the next chapter. The objective of this chapter is to explore the meaning of these industry factors and whether the universe of stocks can be summarised / described by a discrete set of manageable, representative clusters.

In essence the findings in this chapter address a key shortcoming in the literature: namely that no method currently exists for forming homogeneous groups that is useful for capital market research. As Chapter 2 indicates, there are a number of applications in the literature that require homogeneous groupings of stocks such as: identifying control firms, describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate. However, these all rely on the use of existing industry classification schemes such as SIC, NAICS and GICS to create homogeneous groupings but it is not clear how such classification schemes make their groupings. In some cases, classifications are made based on similarity in production method but in others they are based on similarity in product use. The GSC on the other hand is transparent in the way it groups stocks. Stocks are grouped on returns. In fact, the GSC algorithm is flexible enough that it can group stocks on virtually any metric of interest, e.g. financial ratios or operational characteristics. This provides researchers with a powerful way to make homogeneous groupings in any way they require for the study at hand. In the current study, stocks

are grouped on returns but there is no reason to believe they cannot be adapted to any other variable of interest with equal success.

In this study, approximately 10 clusters are identified in each modelling interval. These clusters roughly correspond to the primary, secondary and tertiary sectors in Clarke's (1940) famous three-sector model of the economy[27]. Clusters in the primary sector consist of sub-sectors that are largely involved in the extraction of primary resources. Clusters in the secondary sector consist of sub-sectors such as metal working, construction, chemicals, food and textile production. Clusters in the tertiary sector consist of sub-sectors such as business services, information and communications technology, elaborately transformed manufactures, research and development, retailing, transportation and distribution, entertainment and food and beverage.

Although each cluster has its own unique risk/return profile, in general stocks in the primary sector exhibit moderate to high levels of risk, but generate low to moderate returns. However, during periods of economic prosperity, this sector also generates high returns. Stocks in the secondary sector generally experience moderate levels of risk and generate moderate returns. Lastly, stocks in the tertiary sector experience high levels of risk but also generate high returns. Note this finding is consistent with the principle of risk-return trade-off in finance.

The remainder of this chapter is structured as follows: Section 6.2 discusses the research design, Section 6.3 presents the results, Section 6.4 provides a synthesis of the results, Section 6.5 describes the limitations of this research and Section 6.6 concludes.

6.2 Research Design

Cluster analysis here is a two stage process. The first stage deals with the allocation of stocks into clusters, i.e. clustering, based on returns. The second stage deals with the interpretation of those clusters in terms of industry sectors. The clustering algorithm used in the first stage is the Generalised Style Classification (GSC) algorithm described in Section 4.2. As previously mentioned, the advantages of the GSC are that it can be used on cross sectional, time series data; and possesses

[27] Clarke is credited with developing the three-sector model which divides economies into three sectors of activity: the primary, secondary and tertiary sector. The primary sector involves the extraction and production of raw materials. The secondary sector involves manufacturing or the transformation of raw materials into finished or semi-finished products. The tertiary sector consists of services. Additional sectors not originally identified by Clarke include the quaternary and quinary sectors. The quaternary sector consists of intellectual activities while the quinary sector consists of non-profit activities such as public services.

a GLS correction for heteroskedasticty making it ideal for financial data which is prone to periods of volatility.

The cluster interpretation in the second stage compares the time series cluster mean returns with industry mean returns. This stage deals with the interpretation of clusters in terms of the industrial sectors they represent, discussion of those industrial sectors in terms of their risk and return profile and comparison against other industrial sectors in the context of the wider macroeconomy. Clusters are 'profiled' against industries based on shared correlation. For example, if the cluster 1 mean return correlates highly with the oil and gas extraction industry mean return, then cluster 1 will be interpreted as a cluster representing the energy sector. Clearly, this stage of cluster interpretation requires some degree of subjective assessment but such judgements are common in exploratory data analysis and certainly required in k-means cluster analysis. Another key issue in any form of k-means cluster analysis is the choice of clusters, *K*

While other studies commonly rely on subjective judgement or a priori beliefs to inform the choice of *K*, this study utilizes new technology known as the Gap statistic test developed by Tibshirani, Walter and Hastie (2000) which objectively searches for the optimal *K* thereby minimising the effect of subjective input, which is important as it minimises the possibility of data snooping (see Section 6.2.4). The only subjective analysis that remains is assigning meaningful names and interpretation of the clusters.

6.2.1 The Industry Classification scheme

To calculate industry mean returns, there must be a way to first classify stocks into industries. Industry classification schemes classify firms into categories that best describe their industry of operation. The Standard Industry Classification, SIC is an industry classification scheme developed and maintained by the United States Office of Management and the Budget. The SIC scheme is organised into 4 levels of detail. These are: Division, Major Group, Industry Group and Industry with Division being the broadest level of classification, Major Group the second broadest and so forth. There are 10 Divisions (1st digit), 83 Major Groups (2nd digit), 445 Industry Groups (3rd digit) and several thousand Industries (4th digit)[28]. The actual SIC code is a 4-digit structured code with each digit representing a different level of detail. For example, SIC codes beginning with 0 belong to the

[28] The SIC scheme undergoes multiple revisions, usually at the 4th digit making it difficult to obtain an exact estimate of the number of industry categories at this level of detail.

Agriculture, Forestry and Fishing Division. SIC codes beginning with 1 belong to the *Mining* or *Construction* Divisions. SIC codes beginning with 2 or 3 belong to the *Manufacturing* Division and so on[29].

At the second digit, SIC codes beginning with 01 belong to *Agricultural Production Crops* Major Group under the *Agriculture, Forestry and Fishing* Division. SIC codes beginning with 02 belong to the *Agriculture production livestock and animal specialties* Major Group under the *Agriculture, Forestry and Fishing* Division and so forth[30].

While the SIC classification scheme was established in 1937, it has undergone multiple revisions to accommodate new and changing industrial composition and organisation. Other industry classification schemes include the:

- North American Industry Classification System, NAICS, which is a 6-digit code established in 1997 and designed to supplant the SIC
- Global Industry Classification Standard, GICS, which is a 8-digit code developed by Morgan Stanley Capital International and Standard and Poor's
- Fama-French, FF, Industry classification scheme developed by the academic community. See Fama French (1997).

SIC, NAICS and GICS are available at the individual stock level on the CRSP/Compustat database. The FF industry classification scheme[31] is simply a recoded SIC that Fama and French developed for studying the cost of capital (Fama French, 1997).

[29] Note there are several divisions which overlap at the first digit, e.g. Mining and Construction which both begin with 1 making it impossible to tell in these cases which division a firm belongs to by examining the first digit only. This is a common source of confusion and a noted criticism of SIC codes. See Kahle and Walking (1996) and Clarke (1989).
[30] United States Department of Labour, Occupational Health and Safety Administration, *SIC manual* [online] Available at: <http://www.osha.gov/pls/imis/sic_manual.html> [accessed January 2011]
[31] In addition to Fama-French (1997), the recoded FF industry definitions are available at Kenneth French's homepage. French, K., *Data Library* [online] Available at: <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html> [accessed January 2011]

The following table, adapted from Bhojraj, Oler and Lee (2003) indicate the number of categories by industry classification scheme:

		Title	Categories	Digits
SIC	Level 1 (broadest)	Division	10	First digit
	Level 2	Major Group	83	First 2 digits
	Level 3	Industry Group	445	First 3 digits
	Level 4 (narrowest)	Industry	1004	All 4 digits
NAICS	Level 1 (broadest)	Sector	20	First 2 digits
	Level 2	Subsector	100	First 3 digits
	Level 3	Industry Group	311	First 4 digits
	Level 4 (narrowest)	Industry	1170	First 5 digits
FF	Level 1		48	
GICS	Level 1 (broadest)	Sector	10	First 2 digits
	Level 2	Industry Group	23	First 4 digits
	Level 3	Industry	68	First 6 digits
	Level 4 (narrowest)	Sub-industry	123	All 8 digits

Table 6-1 Number of categories by industry classification scheme.

Despite its age and many shortcomings, the SIC scheme is still the most commonly used industry classification scheme in academic studies and in general (Clarke, 1989; Bhojraj, Lee and Oler, 2003; Kahle and Walking, 1996). Given its established use in the literature, the SIC scheme will be used as the industry classification scheme for 'profiling' the GSC clusters (the second stage of cluster analysis described in Section 6.2).

The next issue is to determine the appropriate level of detail for the SIC code. Using too few digits over-generalises the cluster interpretations while using too many digits allows excessive detail. For example, there are only 10 industry classifications (Divisions) at the first-digit. If $K = 10$, then the clusters will likely form an approximate one-to-one mapping onto each of the 10 Divisions, which may not be a bad solution per se but the Divisions are too broad for any form of specific analysis. On the other hand, there are over one thousand industry classifications at the fourth-digit, which makes interpretation difficult for the opposite reason.

There are 83 industry classifications (Major Groups) at the second-digit. This level of detail provides sufficient detail without over generalising the cluster interpretation. For this reason, 2-digit SIC codes will be used. This level of detail is also commonly used in other studies of industrial organisation for the same reason (Bhojraj, Lee and Oler, 2003; Guenther and Rosman, 1994; Kahle and Walking, 1996).

6.2.2 Correlation analysis

As previously mentioned, cluster mean returns are compared against industry mean returns and these clusters are profiled based on the degree of shared correlation. In essence the following correlation matrix is estimated:

	Cluster				
SIC (2-digit) Industry Group	1	2	3	…	K
Industry 1	$\rho_{1,1}$	$\rho_{1,2}$	$\rho_{1,3}$	…	$\rho_{1,K}$
Industry 2	$\rho_{2,1}$	$\rho_{2,2}$	$\rho_{2,3}$	…	$\rho_{2,K}$
Industry 3	$\rho_{3,1}$	$\rho_{3,2}$	$\rho_{3,3}$	…	$\rho_{3,K}$
…	…	…	…	…	
Industry J	$\rho_{J,1}$	$\rho_{J,2}$	$\rho_{J,3}$	…	$\rho_{J,K}$

Where $\rho_{J,K}$ = correlation coefficient between the J^{th} industry mean return and the K^{th} cluster mean return

To enhance the correlation patterns, squared rank correlations will be used. The original correlation matrix is transformed into the following:

	Cluster				
SIC (2-digit) Industry Group	1	2	3	…	K
Industry 1	$R^R_{1,1} \cdot R^C_{1,1}$	$R^R_{1,2} \cdot R^C_{1,2}$	$R^R_{1,3} \cdot R^C_{1,3}$	…	$R^R_{1,K} \cdot R^C_{1,K}$
Industry 2	$R^R_{2,1} \cdot R^C_{2,1}$	$R^R_{2,2} \cdot R^C_{2,2}$	$R^R_{2,3} \cdot R^C_{2,3}$	…	$R^R_{2,K} \cdot R^C_{2,K}$
Industry 3	$R^R_{3,1} \cdot R^C_{3,1}$	$R^R_{3,2} \cdot R^C_{3,2}$	$R^R_{3,3} \cdot R^C_{3,3}$	…	$R^R_{3,K} \cdot R^C_{3,K}$
…	…	…	…	…	
Industry J	$R^R_{J,1} \cdot R^C_{J,1}$	$R^R_{J,2} \cdot R^C_{J,2}$	$R^R_{J,3} \cdot R^C_{J,3}$	…	$R^R_{J,K} \cdot R^C_{J,K}$

Where $R^R_{J,K}$ = Ranked row correlation
$R^C_{J,K}$ = Ranked column correlation

A good ‘match’ will occur when an industry is highly correlated to a cluster and when a cluster is highly correlated to an industry. The ranked row correlation will be high when a cluster correlates highly with an industry for a given industry; and the ranked column correlation will be high when a cluster correlates highly with an industry for a given cluster. In order for a cluster and industry to be highly ’correlated’ under these conditions, it must have both a high row rank and a high column rank, i.e. the cluster must correlate highly to the industry for that industry only and the industry must correlate highly to the cluster for that cluster only.

For example, in order for cluster 1 to be 'matched' to industry 1, $R^R_{1,1}$ must be high compared to $R^R_{1,2}$, $R^R_{1,3}$ and so forth, i.e. out of all the industry 1 correlations, cluster 1 ranks highly. At the same time, $R^C_{1,1}$ must be high compared to $R^C_{2,1}$, $R^C_{3,1}$ and so forth, i.e. out of all the cluster 1 correlations, industry 1 ranks highly. If only the row rank is high but the column rank is low or vice versa, the squared rank correlation will not be high and cluster 1 and industry 1 will not be matched. Transforming the correlation matrix in this fashion ensures that clusters and industries will only be matched when they are highly correlated to each other and no other cluster/industry.

6.2.3 Cluster visualisation

While not essential, visual representations of the cluster centres relative to industry groups aid interpretation and profiling of the clusters. Quite often many beneficial insights can be made from visual analysis of the solution space. To visualise the solution, a mapping technique known as Multi-Dimensional Scaling, MDS is used. MDS is a perceptual mapping technique for exploring similarities and/or dissimilarities within a data set making it ideal for cluster interpretation.

The MDS algorithm uses similarity/dissimilarity measures (in this case, squared rank correlation coefficients), calculates Cartesian coordinates and maps these coordinates onto *N*-dimensional space, usually 2-dimensional space. Elements which are similar (or specifically have a high similarity measure) will have coordinates close to each other while elements that are dissimilar will have coordinates far from each other.

As previously mentioned, a combination of approaches is used for cluster interpretation. In addition to the squared rank correlations, the MDS perceptual maps will assist in making those interpretations.

6.2.4 Determining *K*

All forms of k-means cluster analysis struggle with the issue of determining the number of clusters, *K*. *K* needs to be specified with care as it can have a drastic impact on the final cluster solution. Common approaches to determining *K* typically involve the use of a scree plot (or scree test). The rationale behind the scree plot test is simple. First, an objective goodness-of-fit measure is determined. The number of *K* is then incremented over a specified range. At each increment, the

goodness-of-fit measure is recorded. The scree plot simply charts the goodness-of-fit measure against *K*. An optimal *K* is reached when there is a large change in the goodness-of-fit measure for that given *K*[32].

In cluster analysis, a common goodness-of-fit measure is the sum of squares, SSQ. This measures the sum of squared deviations of each element from their respective cluster mean. A good cluster solution is one where elements within each cluster are close to their cluster mean (or centroid) hence achieving within cluster homogeneity. SSQ will typically decrease in a monotonic fashion with additional *K* since there are more cluster centres dispersed across the solution space making each pairwise deviation smaller (or at the very least no bigger) from the previous *K*. The fact that the scree plot decreases monotonically therefore is not unexpected. However, when there is a relatively large decrease in the SSQ going from one *K* to the next, this indicates that a substantial improvement in goodness-of-fit has been achieved with the most recent addition of *K* thereby indicating a potentially optimal solution has been found. In a scree plot, this phenomenon is usually referred to as an 'elbow' or 'L-shape' where the charted SSQ decreases substantially at the optimal *K* but then resumes a normal rate of decay for subsequent increments of *K*.

Unfortunately, such tests tend to be *ad hoc* in nature and as such have been criticised for lacking objective statistical criterion. While the scree test is not a formal testing procedure per se, it has been shown to produce reliable and accurate decisions in empirical studies (Milligan, Cooper, 1985). Note however, that even though the scree plot test provides researchers with a quasi-objective criterion for determining *K*, it does not supersede common sense or subjective assessments. That is, once a researcher has determined the optimal *K* through a scree plot, the cluster solution must still be interpreted. There is no guarantee that the cluster interpretation will be sensible or consistent with prior beliefs or theories. In all cases, careful judgement must be exercised to determine whether or not to retain *K* or continue searching.

As the GSC is a form of k-means cluster analysis, the scree plot test may be used to determine the optimal *K*. The Gap statistic test developed by Tibshirani et al (2000) however formalises the scree plot test thus addressing the lack of objective statistical criterion. It does so by generating a null distribution (via random resampling of the data) for the objective goodness-of-fit measure (in this case, SSQ) under varying *K* and then compares the actual scree plot against the null distribution after

[32] Such scree plot tests are also common in factor analysis where instead of determining the optimal number of clusters; researchers are interested in determining the optimal number of factors to extract.

accounting for sampling error. At each level of *K*, the actual SSQ may be compared against the null distribution of SSQ and tested for statistical significance.

The Gap statistic test then proceeds to search for the level of *K* which results in the largest statistically significant difference between the null distribution and the actual *log(SSQ)*. As Tibshirani et al explain: "Our estimate for the optimal number of clusters is then the value of *K* for which $log(W_k)$[33] falls farthest below this reference curve (null distribution)" (p.412). Consider the following:

Figure 6-1 This figure shows the null distribution of SSQ and estimated SSQ for the GAP statistic test estimated via GSC.

As Figure 6-1 indicates, the Gap statistic test will estimate the optimal *K* to be 10 as this results in the largest statistically significant difference between the null distribution and the actual *log(SSQ)*. It is the point where the *log(SSQ)* "falls farthest below (the) reference curve".

[33] Utilising the notation of Tibshirani et al (2000). They are basically referring the objective function in the clustering algorithm. In this case, that is the sum of squares, SSQ.

6.3 Results

Recall from the previous chapter that the data period was restricted to 1983 to 2006. Recall also that the modelling interval was 3 years. The results will therefore be organised accordingly.

6.3.1 1983 to 1985

The first step is to determine the optimal *K*. As Figure 5-2 indicates, significant improvements to the SSQ are achieved between *K* = 5 to 10 however negligible improvements are achieved between *K* = 10 to 15. Regardless, the 'search range' for *K* is set to *K* = 5 to 15 to allow for the possibility of higher optimal *K*. The choice to restrict the search range to *K* = 5 to 15 is made for several reasons.

The first is for ease of interpretability. As Section 6.2.1 explains, 2nd level SIC codes were chosen for the industry classification scheme which produces roughly 80 industry major groups. Setting *K* at too high a level would be counterproductive to the objective of cluster analysis, which is to summarise or reduce the data.

Secondly setting *K* between 5 and 15 is consistent with the findings of Brown and Goetzmann (1997) who originally developed the GSC algorithm for mutual funds data and found approximately 8 mutual fund 'styles'.

Thirdly, restrictions in computational resources limit the choice of *K*. Estimating higher levels of *K* require exponentially greater amounts of computational resources, especially when performing the Gap statistic test as this requires repeated resampling of the data to simulate a null distribution. At the time of writing, performing the Gap statistic test with higher levels of *K* is computationally infeasible.

The Gap statistic test is implemented for *K* = 5 to 15. Recall from Figure 5-2 that incremental changes in *K* between 10 and 15 generate far less improvement than changes between 5 and 10. Therefore, it is likely that the optimal *K* lies between 5 and 10. This notwithstanding, 'searching' between *K* = 5 to 15 provides an adequate range of *K* to accommodate higher levels of *K*.

Figure 6-2 shows the scree plot of SSQ for K = 5 to 11[34] as well as the null distribution obtained from the Gap statistic procedure. According to the scree plot, the optimal K may be 10 as this produces the largest absolute difference between the null value of the SSQ and the SSQ produced by the GSC.

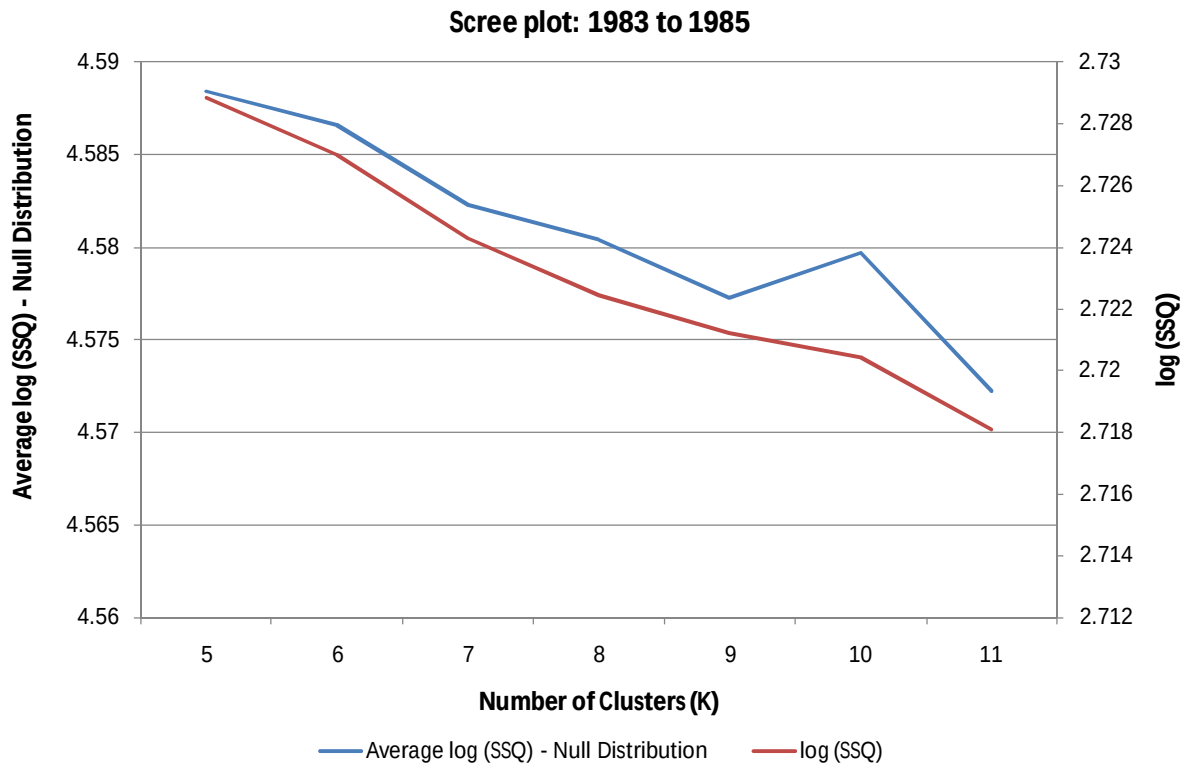


Figure 6-2 The scree plot for the interval 1983 to 1985. The average log(SSQ) for the null distribution is given by the blue line and measured on the left vertical axis while the estimated log(SSQ) is given by the red line and measured on the right vertical axis. The largest statistically significant difference occurs at K = 10 hence the optimal K suggested by this test is 10.

The next step is to evaluate the squared ranked correlation matrix to determine which industries correlate to particular clusters. The squared ranked correlation matrix is reproduced in Table 6-2. Darker shading indicates higher squared ranked correlation while lighter shading indicates lower correlation. A quick visual analysis of the matrix indicates some clear patterns of correlation. For example, cluster 8 correlates highly with the Metal mining; Oil and gas extraction; Petroleum refining and related industries and Water transportation industries indicating that this cluster may represent the primary resources sector. Cluster 6 correlates highly with the Communications sector. Cluster 3 correlates highly with Electric, Gas and Sanitary services which would be representative of the public utilities sector.

[34] additional levels of K were calculated and are available but not reproduced as the higher levels do not provide any substantial differences between the null distribution and SSQ

Additionally, Figure 6-3 shows the MDS perceptual map generated from the ranked correlation matrix for the *K* = 10 solution. The 'diamonds' represent the cluster 'centroids' while the circles represent the industry groups. Industry groups are colour coded to match the cluster which they have been assigned to via GSC. The cluster 'centroids' are computed based on the time series of cluster means. That is, for each cluster, an average is computed from each stock allocated to that cluster for each month in a given 36 month modelling interval generating *k* vectors of cluster means. These cluster means are then mapped onto perceptual space along with industry group means based on squared rank correlation. As previously discussed, greater proximity to the cluster 'centroids' indicates higher correlation, while lower proximity indicates lower correlation. The size of the circles have also been scaled to reflect the number of firms within respective industry groups. Larger circles are industry groups which contain more firms and smaller circles contain fewer firms.

Note the following colour key applies to Table 6-2:

0 – 20th percentile
20 – 40th percentile
40 – 60th percentile
60 – 80th percentile
80 – 100th percentile

SIC Major Group (2-digit)	Cluster 1	2	3	4	5	6	7	8	9	10
Agricultural Production Crops	108	49	16	80	28	330	30	26	60	40
Agricultural Services	22	171	6	112	270	90	120	116	168	33
Metal Mining	64	72	3	18	10	6	4	520	12	42
Coal Mining	40	45	1	14	40	12	6	30	6	5
Oil And Gas Extraction	207	80	4	24	40	48	10	550	12	63
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	70	10	56	72	48	42	4	60	18	12
Building Construction General Contractors And Operative Builders	287	256	39	390	369	60	128	64	198	160
Heavy Construction Other Than Building Construction Contractors	400	189	21	90	56	110	18	352	39	133
Construction Special Trade Contractors	260	144	49	91	152	75	45	42	84	40
Food And Kindred Products	51	115	110	261	174	144	530	19	224	320
Tobacco Products	4	12	192	35	80	140	98	1	108	42
Textile Mill Products	216	342	19	430	352	126	156	52	259	170
Apparel And Other Finished Products Made From Fabrics And Similar Materials	124	387	23	550	204	123	329	68	336	195
Lumber And Wood Products, Except Furniture	204	264	108	360	224	114	164	8	490	155
Furniture And Fixtures	315	480	17	336	504	176	90	74	264	180
Paper And Allied Products	45	378	36	168	96	378	185	40	560	352
Printing, Publishing, And Allied Industries	108	280	102	520	222	102	495	14	273	175
Chemicals And Allied Products	450	470	46	352	180	141	176	86	240	357
Petroleum Refining And Related Industries	112	112	7	72	55	48	22	560	27	225
Rubber And Miscellaneous Plastics Products	344	390	31	238	120	240	99	76	210	432
Leather And Leather Products	250	120	12	153	96	57	161	56	60	90
Stone, Clay, Glass, And Concrete Products	175	288	37	224	414	87	116	60	480	246
Primary Metal Industries	322	550	10	180	200	184	48	147	486	432
Fabricated Metal Products, Except Machinery And Transportation Equipment	343	405	29	368	540	255	126	72	282	148
Industrial And Commercial Machinery And Computer Equipment	504	530	9	376	357	180	81	94	300	245
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	486	560	14	336	424	196	93	84	312	250
Transportation Equipment	255	540	38	477	385	165	204	80	440	330

Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	477	520	32	408	364	200	114	82	230	312
Miscellaneous Manufacturing Industries	308	460	43	328	450	260	70	153	120	258
Railroad Transportation	72	490	45	144	140	140	315	92	360	504
Motor Freight Transportation And Warehousing	110	243	35	161	126	128	66	18	430	336
Water Transportation	147	180	18	84	39	115	32	540	64	216
Transportation By Air	390	132	33	175	279	33	104	24	232	110
Transportation Services	180	370	26	182	342	96	57	54	288	115
Communications	378	217	50	264	198	560	144	44	125	84
Electric, Gas, And Sanitary Services	15	45	560	54	18	243	147	11	44	136
Wholesale Trade-durable Goods	416	510	22	315	282	129	184	78	477	230
Wholesale Trade-non-durable Goods	423	400	30	296	92	200	84	70	287	270
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	65	88	11	190	153	52	36	8	119	78
General Merchandise Stores	36	84	106	279	90	30	560	6	184	75
Food Stores	80	196	41	324	336	63	520	36	168	150
Automotive Dealers And Gasoline Service Stations	40	91	13	162	144	10	480	72	108	80
Apparel And Accessory Stores	145	261	44	540	150	90	392	30	182	116
Home Furniture, Furnishings, And Equipment Stores	72	240	27	342	480	51	280	50	144	100
Eating And Drinking Places	228	306	40	245	344	93	500	62	135	104
Miscellaneous Retail	112	396	47	560	234	111	432	32	357	190
Hotels, Rooming Houses, Camps, And Other Lodging Places	9	48	520	24	12	81	12	2	35	35
Personal Services	320	136	8	180	140	48	68	34	114	70
Business Services	495	500	34	400	294	212	129	100	190	371
Automotive Repair, Services, And Parking	42	234	20	154	350	130	72	14	248	126
Miscellaneous Repair Services	8	7	25	3	4	6	2	230	1	36
Motion Pictures	370	216	42	168	154	125	26	99	80	144
Amusement And Recreation Services	171	180	2	77	72	21	14	360	32	60
Health Services	231	250	24	216	130	56	136	318	66	297
Educational Services	60	21	15	45	24	12	64	6	25	8
Engineering, Accounting, Research, Management, And Related Services	384	369	25	490	315	156	50	144	170	282

Table 6-2 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit). Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

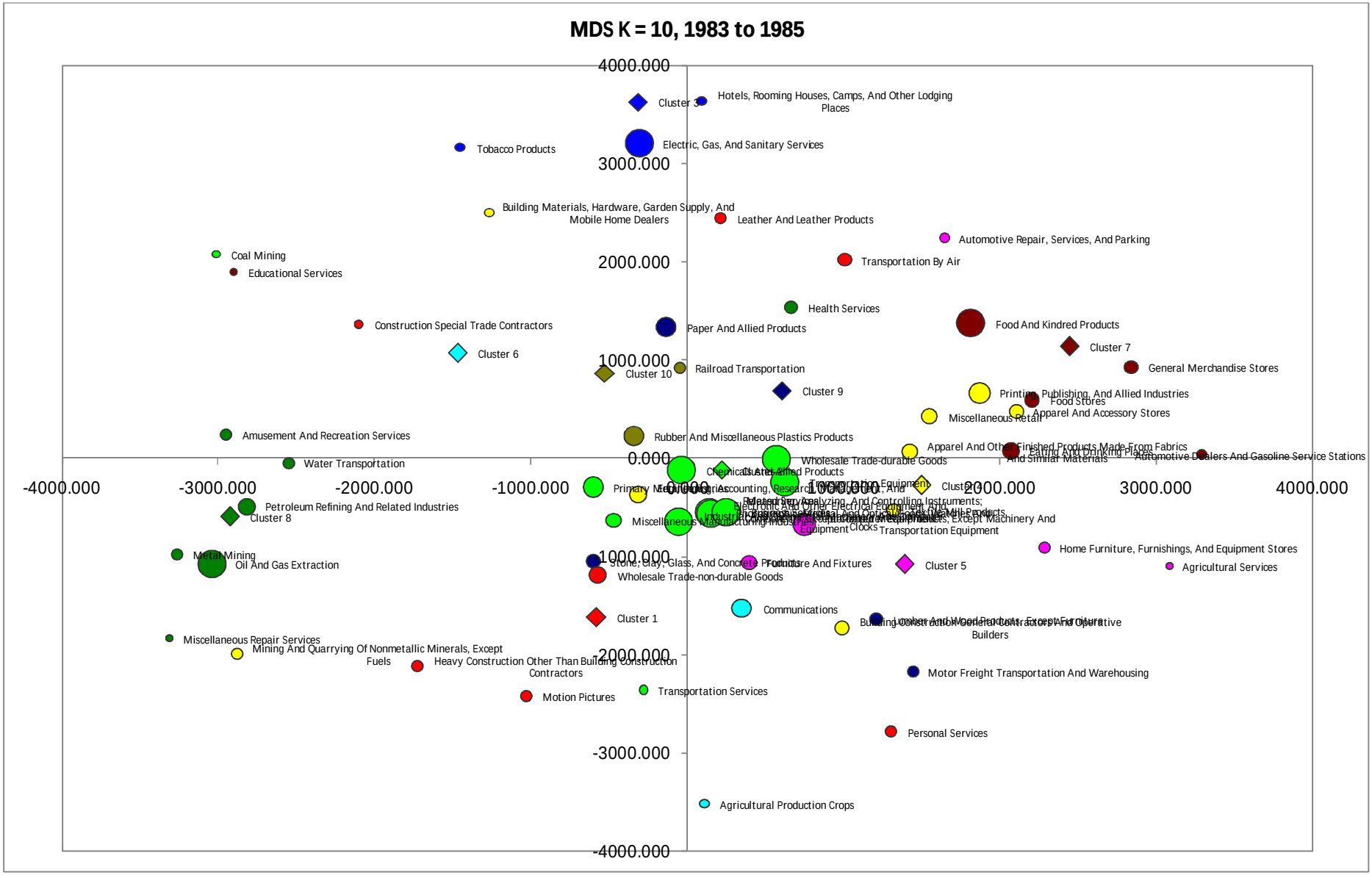


Figure 6-3 MDS chart for the K = 10 solution. Industry mean returns are mapped against cluster mean returns. Diamonds represent the cluster 'centroids' while circles represent the industries. Industries are colour coded to match the cluster which they have been assigned to via GSC. The size of the circle reflects the number of firms within the industry. Greater proximity to the cluster 'centroids' indicate higher correlation.

6.3.2 Cluster Interpretation

In interpreting the clusters, a combination of approaches are used including analysis of the squared ranked correlation matrix as well as visual analysis of the MDS perceptual map. Thresholds for the squared ranked correlations are also established to aid interpretation. For example, Table 6-2 contains 56 SIC major groups at the 2nd digit level[35]. Additionally, there are 10 clusters. Therefore, the theoretical maximum squared rank correlation an industry-cluster pair may achieve is 560. This is because the maximum row rank an industry-cluster pair may achieve is 10 (as there are 10 clusters). The maximum column rank an industry-cluster pair may achieve is 56. Therefore, when an industry-cluster pair are highly correlated, its row rank will be very high (10 is the maximum) and the same is true for its column rank (56 is the maximum). In such a case, its squared rank correlation will have a maximum of 560[36]. 95 percent of the theoretical maximum may be used to establish an arbitrary threshold to filter out 'low' industry-cluster correlations. In this case, the 95 percent 'cut-off' limit (given a maximum of 560) is 532. Table 6-3 shows the industries that exceed the cut off limit as well as potential cluster interpretations. The average industry monthly return ('Mean' column) and standard deviation ('SD' column) across the 36-month modelling interval is also shown[37].

[35] Although there are 83 Major Groups at the 2nd SIC digit level, some of these are removed during data filtering. For example, one of the filtering conditions removes financial firms. Therefore, any industry major groups that correspond to this sector would also be removed from the analysis.

[36] This occurs for example between the Petroleum refining and related industries industry group and Cluster 8; as well as the Electric, Gas and Sanitary services industries and Cluster 3 indicating very strong correlation between these industries and their respective clusters.

[37] The average industry return is calculated as the time series, cross sectional average or the 'average of the averages'. For each month, the average return for all stock within a given industry is calculated giving the cross sectional monthly average. The average of the cross sectional averages across 36 months are then calculated giving the time series cross sectional monthly average. The standard deviation is calculated from the 36 cross sectional average returns for a given industry and is commonly used in finance as an indicator of risk.

Cluster	Industry	Mean	SD	Interpretation
1	• N/A			Generic
2	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0124	0.0712	Manufacturing
	• Primary Metal Industries	0.0090	0.0518	
	• Transportation Equipment	0.0205	0.0553	
3	• Electric, Gas, And Sanitary Services	0.0226	0.0259	Utilities
4	• Miscellaneous Retail	0.0154	0.0622	Retail
	• Apparel And Other Finished Products Made From Fabrics And Similar Materials	0.0277	0.0608	
	• Apparel And Accessory Stores	0.0350	0.0806	
5	• Fabricated Metal Products, Except Machinery And Transportation Equipment	0.0157	0.0423	Basic metal products
6	• Communications	0.0219	0.0440	Communications
7	• General Merchandise Stores	0.0264	0.0651	General merchandise
8	• Petroleum Refining And Related Industries	0.0114	0.0482	Primary resources
	• Oil And Gas Extraction	-0.0041	0.0714	
	• Water Transportation	0.0098	0.0664	
9	• Paper And Allied Products	0.0216	0.0458	Paper
10	• N/A			Generic

Table 6-3 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations

Analysis of Table 6-3 is consistent with the squared rank correlation matrix in Table 6-2 as well as visual analysis of the MDS perceptual map in Figure 6-3. For example, Cluster 8 is highly correlated to the Petroleum Refining and Related Industries, Oil and Gas Extraction and Water Transportation industries. These industries are also very close to the cluster 8 centroid shown in the MDS perceptual map.

Cluster 2 represents the **Manufacturing sector**. The *Electronic And Other Electrical Equipment And Components, Except Computer Equipment* Major Group consists of industries such as: *Electronic & Other Electrical Equipment, Motors & Generators, Household Appliances, Printed Circuit Boards, Semiconductors & Related Devices* and so forth. The *Primary Metal Industries* Major Group consists of industries such as: *Iron & Steel Foundries, Primary Production of Aluminium, Steel Works, Blast Furnaces & Rolling & Finishing Mills* and so forth. Lastly, the *Transportation Equipment* Major Group consists of industries such as: *Motor Vehicles & Passenger Car Bodies, Motor Vehicle Parts & Accessories, Aircraft & Parts, Truck & Bus Bodies, Railroad Equipment* and so forth. The common theme among these industries is that they all involve some form of manufacturing process. These industries typically take some form of semi finished material (or raw materials in the case of Primary Metal Industries) and transform them into other forms of semi finished goods (e.g. motor vehicle

parts and accessories) or finished goods (e.g. household appliances). This sector has relatively moderate to high levels of risk (standard deviation is between 0.05 and 0.07) and has relatively moderate levels of return.

Cluster 3 represents the **Utilities** sector. The *Electric, Gas and Sanitary Services* Major Group consists of industries such as: *Electric Services, Natural Gas Transmission & Distribution, Water Supply, Sanitary Services, Refuse Systems, Hazardous Waste Management* and so forth, which are all public utilities services. The proximity of this sector to the Government (it is likely the Government would comprise the majority of clientele for firms in this sector) as well as the fact that these are necessity services[38] means these firms operate in a lower risk environment (standard deviation is 0.0259). Its returns patterns however is moderately high compared to that of other stocks (average monthly return is 0.0226).

Cluster 4 represents the **Retail** sector. The *Miscellaneous Retail* Major Group consists of industries such as: *Miscellaneous Shopping Goods Stores, Jewelry Stores, Hobby, Toy & Game Shops* and so forth. The *Apparel And Other Finished Products Made From Fabrics And Similar Materials* Major Group consists of industries such as: *Men's & Boys' Furnishings, Work Clothing, & Allied Garments, Women's, Misses', and Juniors Outerwear, Women's, Misses', Children's & Infant's Undergarments* and so forth. The *Apparel and Accessory Stores* Major Group consists of industries such as: *Women's Clothing Stores, Family Clothing Stores, Shoe Stores* and so forth. These industries are not involved in the manufacture of finished consumer goods but rather the on-selling of such goods to the general public, i.e. retail sales. Many of these retail goods involve clothing/fashion and entertainment. Given the 'leading' nature of these industries and the 'luxury' nature of these goods[39], industries in this sector are more sensitive to periods of economic downturns such as recessions. Consumer Confidence Indices for instance are commonly used as leading indicators of economic conditions. Once again, these industries are exposed to different sources of risk causing their returns patterns to vary from other sectors. This sector has relatively higher levels or risk (standard deviation is between 0.06 and 0.08) but in accordance with finance theory[40] also has higher levels of return (average monthly return is between 0.015 and 0.035).

[38] As opposed to normal or luxury goods to borrow a term from the Economics literature.
[39] As opposed to necessity goods
[40] As previously discussed, a fundamental principle in finance theory is that there exists a relationship between risk and return. The riskiness of a firm's operating environment (or perhaps industry) would therefore translate to riskiness in the firm's equity (specifically risk to the equity holders), which will in turn affect equity prices and returns. Much of the highly influential work of Fama and French (1992; 1993) for example is based on the fundamental tenet that risk is correlated to returns and it is these risk factors (assuming they are properly identified and priced) that drive patterns of return.

Cluster 5 represents the production of **basic metal** products. The *Fabricated Metal Products, Except Machinery And Transportation Equipment* Major Group consists of industries such as: *Cutlery, Handtools & General Hardware, Fabricated Structural Metal Products, Fabricated Plate Work (Boiler Shops), Sheet Metal Work, Prefabricated Metal Buildings & Components, Bolts, Nuts, Screws, Rivets & Washers, Metal Forgings & Stampings* and so forth. Note that these industries do not involve elaborately transformed manufactures such as computer equipment, consumer electronics or scientific equipment but rather basic (non-elaborately transformed) metal products - some of which are finished goods (e.g. handtools and general hardware) and some of which are intermediate goods (e.g. prefabricated metal buildings and components). This is different to Cluster 2 which involve more elaborately transformed manufactured goods. Once again, the argument can be made that industries in this cluster operate in a different risk environment to that of other industries hence their returns pattern are unique. Many of the industries in this cluster are linked to the construction sector (e.g. *prefabricated metal buildings and components, Fabricated Structural Metal Products*, etc.) which operates in a highly challenging, complex and financially risky environment. This sector has moderate levels of risk (standard deviation is 0.0423) and moderate levels of return (average monthly return is 0.0157).

Cluster 6 represents the **Communications** sector. The *Communications* Major Group consists of industries such as: *Radiotelephone Communications, Telephone Communications, Radio Broadcasting Stations, Television Broadcasting Stations, Cable & Other Pay Television Service*s. The communications sector faces risk from developing and maintaining large infrastructure networks, foreign competition, changing political landscapes and rapidly changing technologies. This sector has moderate levels of risk (standard deviation is 0.0440) and moderate levels of return (average monthly return is 0.0219).

Cluster 7 represents the **General merchandise** sector. The *General Merchandise Stores* Major Group consists of industries such as: *Department Stores, Variety Stores* and so forth. While similar to the Retail cluster, industries within this cluster appear to be more general (e.g. Department stores vs. Children's apparel). It is likely that the larger scale of these firms places them in a slightly different risk category to those appearing in the retail cluster. This sector has moderate to high levels of risk (standard deviation is 0.0651) and moderate levels of return (average monthly return is 0.0264).

Cluster 8 represents the **Primary resources** sector. The *Petroleum Refining and Related Industries; Oil and Gas extraction; and Water Transportation* Major Groups consist of industries such as: *Petroleum Refining, Crude Petroleum & Natural Gas, Drilling Oil & Gas Wells, Oil & Gas Field Exploration Services, Deep Sea Foreign Transportation of Freight* and so forth. These industries deal with the extraction and refining of energy resources (fossil fuels) – namely petroleum, oil and gas. A very significant and unavoidable source of risk is that of exploration. Furthermore, many oil and gas wells are located offshore (hence the need for water transportation of freight) resulting in a very complex and risky business environment. Large amounts of capital are required with a high degree of uncertainty regarding return on investment. This sector has moderate to high levels of risk (standard deviation is between 0.04 and 0.07) but low levels of return (average monthly return is between 0.00 and 0.01). In particular, the *Oil and Gas Extraction* Major group suffers from high risk but generates low returns (average monthly return is -0.0041) indicating poor performance during this period.

Cluster 9 represents the **Paper** sector. The *Paper and Allied Products* Major Group consists of industries such as: *Pulp Mills, Paper Mills, Paperboard Mills, Paperboard Containers & Boxes, Plastics, Foil & Coated Paper Bags* and so forth. This sector has moderate levels of risk (standard deviation is 0.0458) and moderate levels of return (average monthly return is 0.0216).

Of the 10 clusters, 8 (Clusters 2 to 9) contain industries that exceed the 95 percent cut off discussed before. This does not necessarily mean that Clusters 1 and 10 are redundant and/or have no interpretation. Clusters 1 and 10 have been interpreted as 'generic' clusters and their presence is necessary to ensure that 'orthogonality' is preserved among the remaining 8 clusters. If Clusters 1 and 10 were removed, the industries assigned to those clusters (which do not have a unique returns pattern) will in some sense 'pollute' the remaining clusters making interpretation more difficult.

Another purpose for these 'generic' clusters is to absorb firms that have no clearly defined industry of operation. This may be the case for conglomerates, which operate across a number of industries hence allocation to one particular cluster would be inappropriate. Likewise, firms transitioning between one industry and another after divesture may also fall into this category.

In summary, the best performing sector in this modelling interval is the Utilities sector as it has relatively low levels of risk but generates relatively moderate to high levels of return. By contrast,

the poorest performing sector is the Primary resources sector which has relatively high levels of risk but relatively low levels of return.

6.3.3 1986 to 1988

As with the previous modelling interval, the Gap statistic test is used to determine the optimal *K*, henceforth referred to as *K**. From there, cluster interpretation will be based on analysis of the squared rank correlation matrix and the MDS perceptual map.

As before, the search range for *K** is between 5 and 15 (although only the results *K* = 5 to 11 are displayed as higher levels of K do not indicate and substantial differences between the null distribution and *log(SSQ)*). Figure 6-4 shows the scree plot for the modelling interval: 1986 to 1988. The largest absolute difference between the null value of the SSQ and the SSQ produced by the GSC occurs at *K* = 9 so *K** suggested by this test is 9.

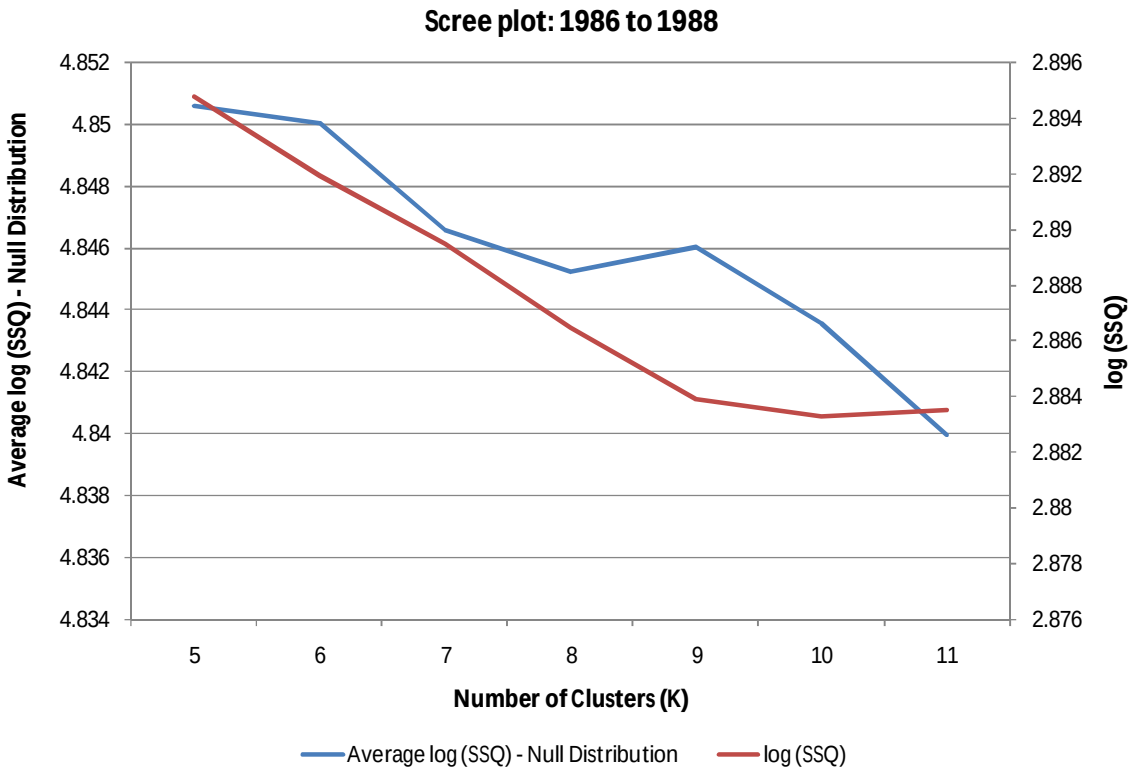


Figure 6-4 The scree plot for the interval 1986 to 1988. The average log(SSQ) for the null distribution is given by the blue line and measured on the left vertical axis while the estimated log(SSQ) is given by the red line and measured on the right vertical axis. The largest statistically significant difference occurs at *K* = 9 hence the optimal *K* suggested by this test is 9.

The squared rank correlation matrix is located in the appendices in Table 10-2. Figure 6-5 shows the MDS perceptual map.

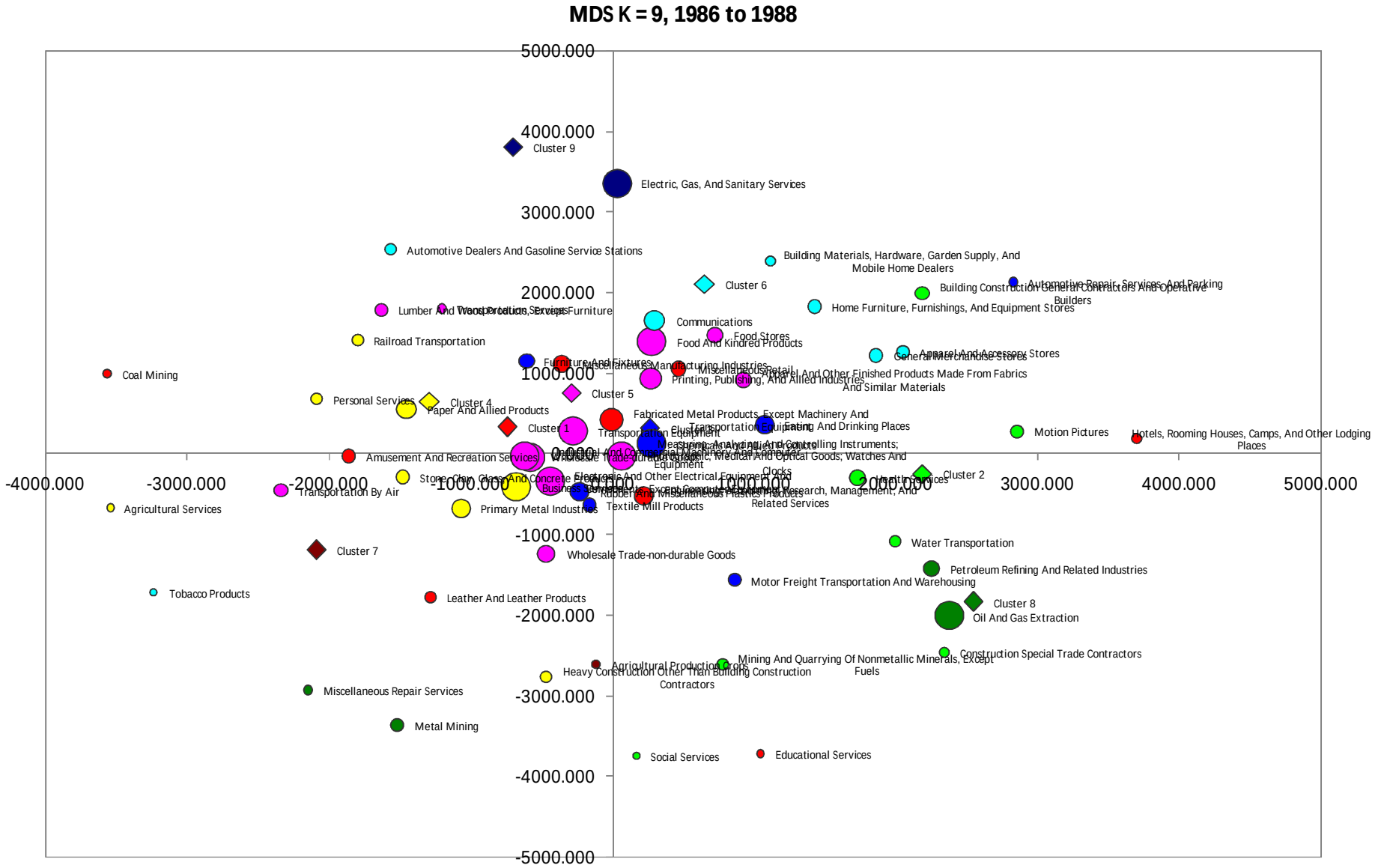


Figure 6-5 MDS chart for the K = 9 solution, 1986 to 1988. Industry mean returns are mapped against cluster mean returns. Diamonds represent the cluster 'centroids' while circles represent the industries. Industries are colour coded to match the cluster which they have been assigned to via GSC. The size of the circle reflects the number of firms within the industry. Greater proximity to the cluster 'centroids' indicate higher correlation.

Visual analysis of Figure 6-5 indicates some clear patterns of clustering. Cluster 8 correlates highly with the Petroleum, Refining and Related Industries; and Oil and Gas Extraction Major Group, which is similar to Primary resources cluster from the 1983 to 1985 modelling interval. Cluster 9 correlates highly with the Electric, Gas and Sanitary services Major Group which is similar to the Utilities cluster from the previous modelling interval.

6.3.4 Cluster Interpretation

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-4 shows these industries as well as cluster interpretations, the average monthly industry return and standard deviation across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Fabricated Metal Products, Except Machinery And Transportation Equipment	0.0065	0.0668	Basic metal products
2	• N/A			Generic
3	• Chemicals and Allied Products	0.0164	0.0791	Chemicals
4	• Paper And Allied Products	0.0175	0.0741	Paper
5	• Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0103	0.0750	Elaborately transformed manufactures
	• Industrial And Commercial Machinery And Computer Equipment	0.0106	0.0764	
6	• N/A			Generic
7	• N/A			Generic
8	• Petroleum Refining And Related Industries	0.0173	0.0678	Primary resources
	• Oil And Gas Extraction	0.0072	0.0775	
9	• Electric, Gas and Sanitary Services	0.0127	0.0395	Utilities

Table 6-4 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 1986 to 1988 modelling interval

There are several recurring clusters from the previous modelling interval. Clusters 1, 4, 8 and 9 have similar industries and interpretations as the previous modelling interval[41].

[41] Note that in cluster analysis, the numerical ordering of the clusters is not important. For example, in the previous modelling interval, the utilities sector was represented by cluster 3, yet in the current modelling interval, it is represented by cluster 9. They are essentially the same cluster even though the numerical order is different.

Cluster 1 represents the **Basic metal** products sector and has relatively moderate levels of risk and slightly below average levels of return.

Cluster 3 represents the **Chemicals** sector. The *Chemicals and Allied Products* Major Group consists of industries such as: Industrial *Inorganic Chemicals, Plastic Material, Synth Resin/Rubber, Cellulos (No Glass), Medicinal Chemicals & Botanical Products, Pharmaceutical Preparations, In Vitro & In Vivo Diagnostic Substances, Specialty Cleaning, Polishing and Sanitation Preparations, Agricultural Chemicals* and so forth. Many of the industries are involved with the production of medicines and pharmaceutical products. As these are necessities, such industries would be less exposed to market conditions and as such belong to a different risk category. Furthermore, these industries operate differently in that they are characterised by high research and development costs with uncertain payoffs resulting in high costs of capital, which again places them in a different risk category. Other industries within this cluster are involved in the production of chemicals and other materials used in industrial processes which exposes them to certain manufacturing sectors. Note that the United States is also a major exporter of chemicals and medicinal products exposing this sector to risk from external sources. This sector has relatively moderate to high levels of risk and relatively moderate to high levels of return.

Cluster 4 represents the **Paper** sector and has relatively moderate to high levels of risk and relatively moderate to high levels of return.

Cluster 5 represents the **Elaborately transformed manufacturing** sector. The *Measuring, Analyzing, And Controlling Instruments* etc. Major Group consists of industries such as: *Construction Machinery & Equipment, Industrial Trucks, Tractors, Trailers & Stackers, Special Industry Machinery, Computer & office Equipment, Electronic Computers, Computer Communications Equipment, Calculating & Accounting Machines* and so forth. The *Industrial And Commercial Machinery And Computer Equipment* Major Group consists of industries such as: *Search, Detection, Navigation, Guidance, Aeronautical Systems, Laboratory Apparatus & Furniture, Auto Controls For Regulating Residential & Commercial Environments, Laboratory Analytical Instruments, Measuring & Controlling Devices, Surgical & Medical Instruments & Apparatus, Orthopedic, Prosthetic & Surgical Appliances & Supplies, X-Ray Apparatus & Tubes & Related Irradiation Apparatus, Electromedical & Electrotherapeutic Apparatus* and so forth. These industries have a common element in that they all involve the production of elaborately transformed manufactures, ETMs. ETMs such as semiconductors, aircraft, computer equipment, pharmaceutical preparations and medical

equipment form the majority of U.S. exports hence this sector is heavily exposed to external risk. This sector has relatively moderate to high levels of risk but relatively moderate to low levels of return.

Cluster 8 represents the **Primary resources** sector. It has moderately high levels of risk with low to high levels of returns. As before, the *Oil and Gas Extraction* Major group is the poorest performing sub-sector within this sector with high levels of risk but offering very low levels of return.

Cluster 9 represents the **Utilities** sector. As before, it has moderate to low levels of risk but moderate to high levels of return.

Lastly, there are several **Generic** clusters which as previously discussed are required to ensure orthogonality among the remaining clusters and/or account for conglomerates and divestures.

This modelling interval is characterised by overall higher levels of risk but lower levels of return indicating poorer performance in the stock market relative to the previous 36 months. Within this modelling interval, the **Utilities** sector again performs well with generally lower levels of risk but moderate to high levels of return. The **Primary resources** sector, in particular the *Oil and Gas Extraction* sub-sector is one of the poorest performers with high levels of risk but low levels of return.

6.3.5 1989 to 1991

The next modelling interval is 1989 to 1991. As with the previous modelling interval, the Gap statistic test is used to help determine K^*. Figure 6-6 shows the scree plot. According to the scree plot, there are two possible values for K^*. These are 5 and 9. Analysis of the $K = 5$ solution however, indicates too much clustering with the clusters containing disparate industries. Therefore, $K = 9$ was chosen as the optimal level of K. Furthermore, setting $K = 5$ would be inconsistent with earlier findings from Chapter 5, which indicated that further improvement in SSQ beyond $K = 5$ is minimal; and the optimal K lies between 5 and 10.

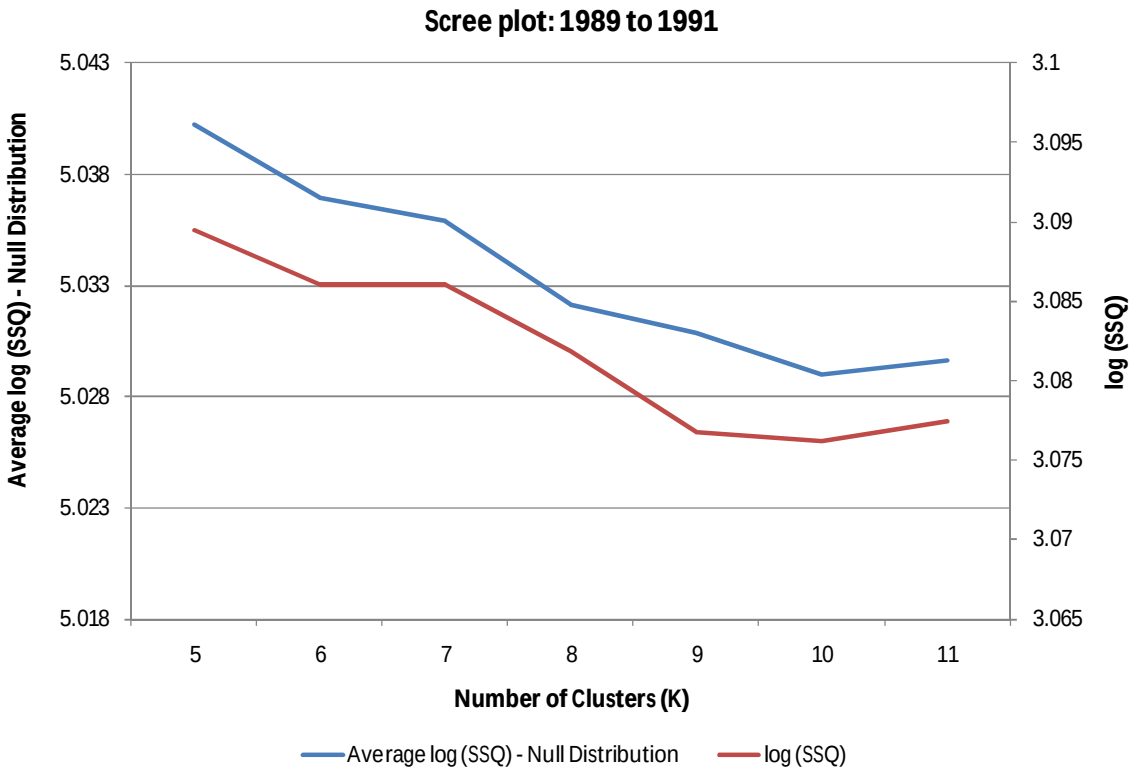


Figure 6-6 The scree plot for the interval 1989 to 1991. The largest statistically significant difference occurs at $K = 9$ hence K^* suggested by this test is 9.

The squared rank correlation matrix is located in the appendices in Table 10-3. Figure 6-7 shows the MDS perceptual map.

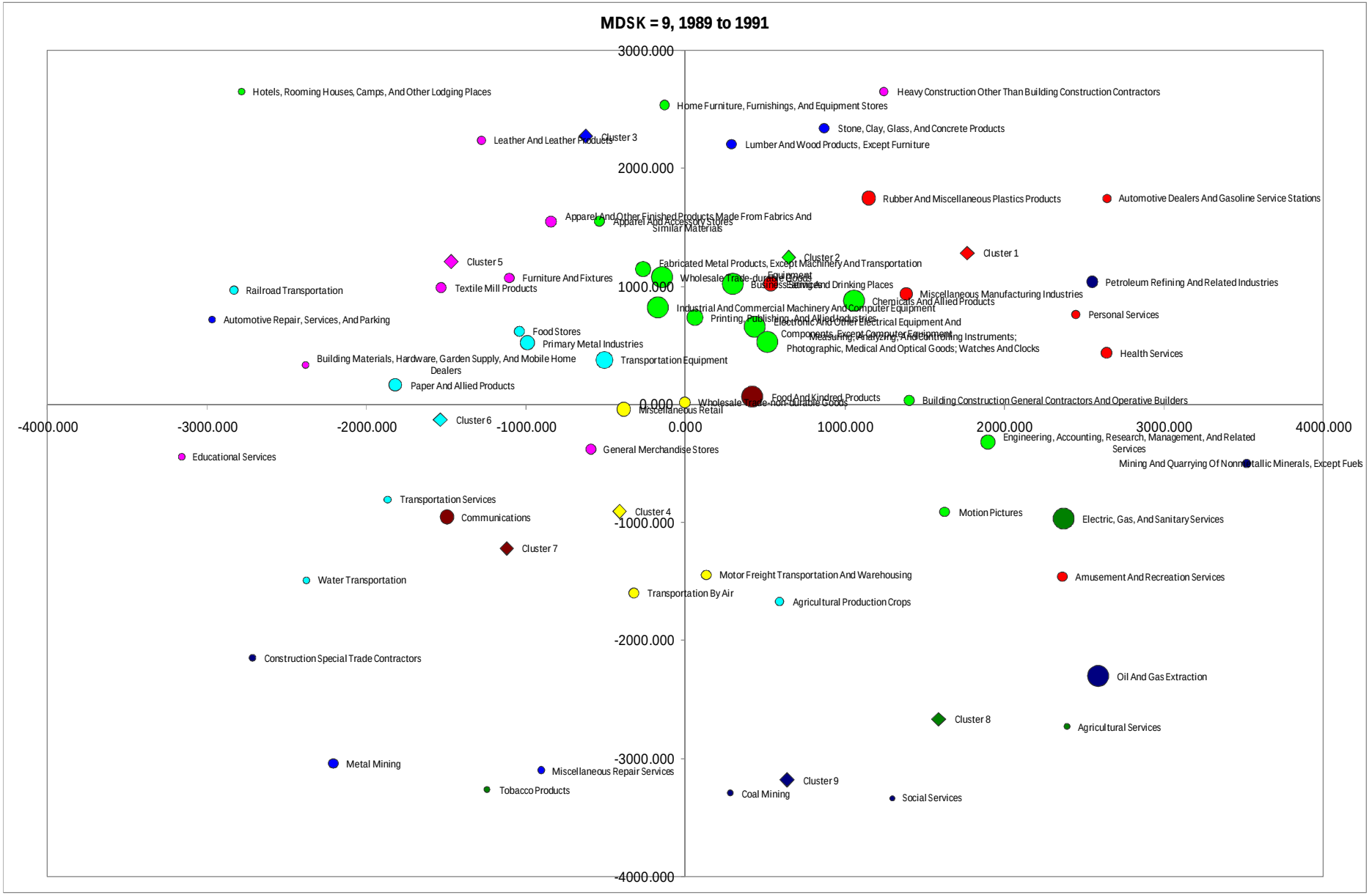


Figure 6-7 MDS chart for the K = 9 solution, 1989 to 1991.

6.3.6 Cluster Interpretation

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-5 shows these industries as well as cluster interpretations, average monthly industry returns and standard deviations across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Miscellaneous Manufacturing Industries	0.0212	0.0627	Miscellaneous manufacturing/Generic
2	• Business Services	0.0163	0.0548	Business services and Elaborated transformed manufactures
	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0107	0.0569	
	• Industrial And Commercial Machinery And Computer Equipment	0.0149	0.0610	
3	• Lumber And Wood Products, Except Furniture	-0.0004	0.0617	Construction materials
	• Stone, Clay, Glass, And Concrete Products	0.0071	0.0572	
4	• N/A			Generic
5	• Furniture And Fixtures	0.0074	0.0468	Retail – Apparel and Furnishing
	• Apparel And Other Finished Products Made From Fabrics And Similar Materials	0.0097	0.0700	
6	• Paper And Allied Products	0.0134	0.0560	Generic/Transportation
	• Transportation Equipment	0.0097	0.0561	
	• Food Stores	0.0067	0.0462	
7	• Food And Kindred Products	0.0160	0.0432	Food production
8	• Electric, Gas, And Sanitary Services	0.0129	0.0272	Utilities
9	• Oil And Gas Extraction	0.0139	0.0525	Primary resources
	• Coal Mining	0.0245	0.0770	

Table 6-5 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 1989 to 1991 modelling interval

Cluster 1 represents the **Miscellaneous manufacturing** sector however given the nature of the industries contained within may be better interpreted as a **Generic** cluster. Visual analysis of the MDS perceptual map in Figure 6-7 shows that Cluster 1 also includes other major industry groups such as *Rubber and Miscellaneous Plastics, Personal Services, Health Services, Eating and Drinking places and Automotive Dealers and Gasoline Service Stations*. These industries do not appear to share very much in common in terms of risk/return profile.

Cluster 2 represents the **Business services** and **Elaborately transformed manufacturing** sector. It is similar to the ETMs cluster in the previous set of results except for the addition of the *Business*

services Major group. The *Business services* Major group consists of industries such as: *Advertising, Consumer Credit, Equipment rental and leasing, Employment agencies, Computer programming, Prepackaged software, Computer integrated systems design, telephone interconnect systems* and so forth. The inclusion of Business services to the ETMs cluster may at first seem odd however analysis reveals that many of the industries in the Business services Major group involve technology related services and products (e.g. *computer programming, prepackaged software*, etc.) which are similar to industries in the *Industrial And Commercial Machinery And Computer Equipment* Major group (e.g. *Computer & office Equipment, Electronic Computers, Computer Communications Equipment*, etc.). Therefore, industries in the **Business services** Major group may share a similar risk/return profile to those in the ETMs sector. This sector has relatively moderate to high levels of risk and relatively high levels of return.

Cluster 3 represents the **Construction materials** sector. It may appear odd that the *Lumber And Wood Products, Except Furniture Major group is clustered with the Stone, Clay, Glass, And Concrete Products* Major group. However, analysis of the industries contained within indicate that these sectors are closely related to the construction sector, in particular the manufacture of construction materials. The *Lumber And Wood Products, Except Furniture Major group consists of industries such as: Millwood, Veneer, Plywood, & Structural Wood Members, Mobile Homes, Prefabricated Wood Buildings & Components* and so forth while the *Stone, Clay, Glass, And Concrete Products Major group consists of industries such as: Cement (Hydraulic), Structural Clay Products, Concrete, Gypsum & Plaster Products, Cut Stone & Stone Products, Abrasive, Asbestos & Misc Nonmetallic Mineral Products* and so forth. Many of the materials produced by these industries are used in the construction of homes, buildings and structures. Given the unique risk profile of the construction sector (see Section 6.3.2), the stocks in this sector would have a unique returns pattern to that of the general market. The construction sector operates in a highly challenging, complex and financially risky environment. Projects require large amounts of capital with significant risk of default and/or non-completion. Other risk factors include non availability of funds, inflation of project costs (cost blowouts), exchange rate risk, insurance risk and risk from natural disasters. Complicating factors include changing industrial relations, compliance requirements, sourcing raw materials, changing political environments, etc. This sector has relatively moderate to high levels of risk but low levels of return.

Cluster 5 represents the **Retail – Apparel and Furnishing** sector. The *Apparel etc*. Major group consists of industries that relate mainly to clothing while the Furniture and Fixtures Major group

consists of industries involved in home and office furnishing. Visual analysis of the MDS perceptual map also indicates that the *Leather and leather products* Major group is allocated to this cluster which has obvious links to the Apparel and Furnishing sectors. This cluster is similar to the Retail clusters from previous results. This sector has relatively moderate levels of risk but slightly below average levels of return.

Cluster 6 may be a potential **Generic** cluster as its industries appear to be quite disparate. The *Paper and allied products* Major group is involved in paper products. The *Food stores* Major group is involved with food retailing (and not so much production). Lastly, the *Transportation Equipment* Major group contain industries that relate to the production and maintenance of various transportation equipment such as motor vehicles, trucks, buses, commercial transportation, aircraft, ship and rail.

However, visual analysis of the MDS perceptual map indicates that this cluster contains several other Major groups that relate to the transportation sector such as Railroad transportation, Transportation services and Water transportation Major groups. Although these other Major groups do not meet the 95 percent cut-off limit described in Section 6.3.2, they do score very close. For example, the Transportation services Major group has a squared rank correlation of 468 out of a potential maximum of 513. This sector has moderate levels of risk and moderate levels of return.
Cluster 7 represents the **Food production** sector. Note that this sector does not deal with Food retailing (although these sectors are obviously related). The *Food and Kindred products* Major group consists of industries such as: *Meat Packing Plants, Poultry Slaughtering and Processing, Dairy Products, Canned, Frozen & Preserved Fruit, Veg & Food Specialties, Canned, Fruits, Veg, Preserves, Jams & Jellies, Grain Mill Products, Bakery Products, Sugar & Confectionery Products, Beverages, Prepared Fresh or Frozen Fish & Seafood* and so forth. Food production belongs to a different risk category because of its exposure to various environmental risk factors such as draught, flooding, pests, extreme weather conditions, etc. This sector also faces considerable risk from external sources as the U.S. is both an exporter and importer of food and beverage. This sector has moderate levels of risk but above average levels of return.

Clusters 8 and 9 represent the **Utilities** and **Primary resources** sectors respectively and are identical to the previous results. The **Utilities** sector has relatively low levels of risk and moderate to high levels of return. The **Primary resources** sector has moderate levels of risk and moderate levels of return.

Overall, the **Utilities** sector again exhibits strong performance with a relatively high return to risk ratio. The **Primary resources** sector does not perform too poorly compared to previous modelling intervals indicating some improvement in performance for this sector. By contrast, the **Construction materials** sector performs poorly with a low return to risk ratio.

6.3.7 1992 to 1994

The next modelling interval is 1992 to 1994. As with the previous results, the Gap statistic test is used to help determine K^*. This is shown in Figure 6-8. There are two possible values for K^*. These are 8 and 9. Although both solutions are highly similar, $K = 9$ offers slightly better, i.e. more intuitive clustering so K^* will be set to 9.

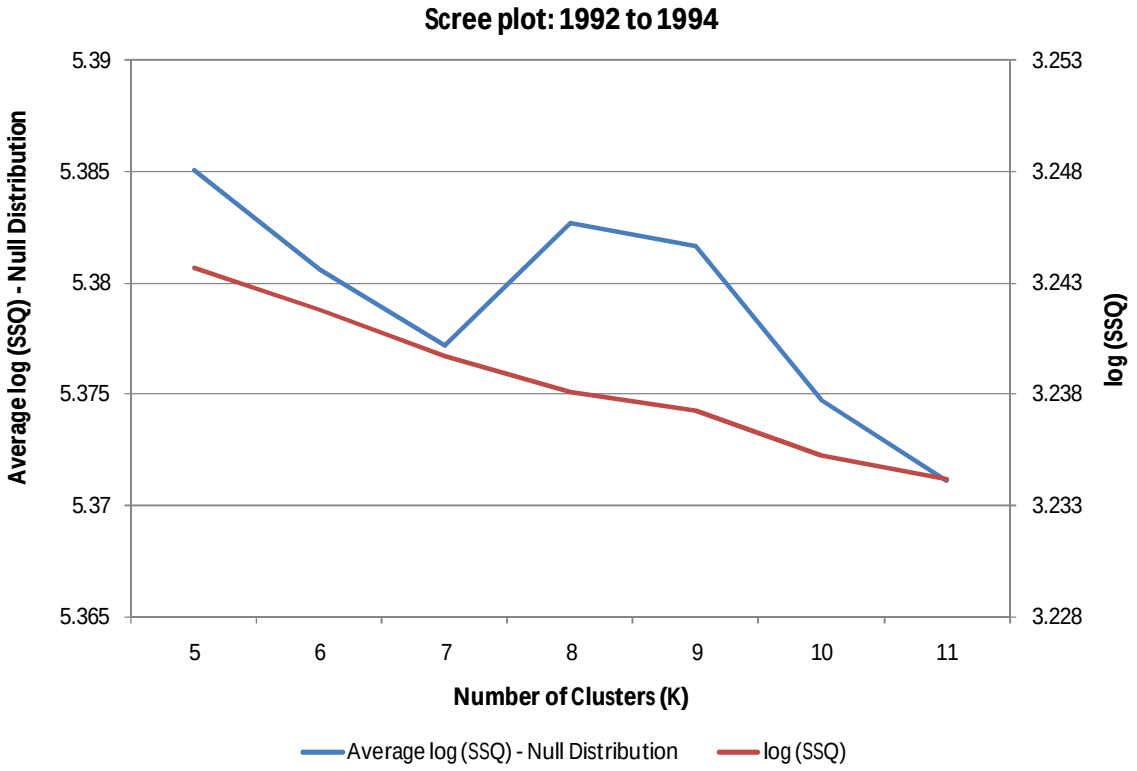


Figure 6-8 The scree plot for the interval 1992 to 1994. The largest statistically significant difference occurs at $K = 9$ hence K^* suggested by this test is 9.

The squared rank correlation matrix is located in the appendices in Table 10-4. Figure 6-9 shows the MDS perceptual map.

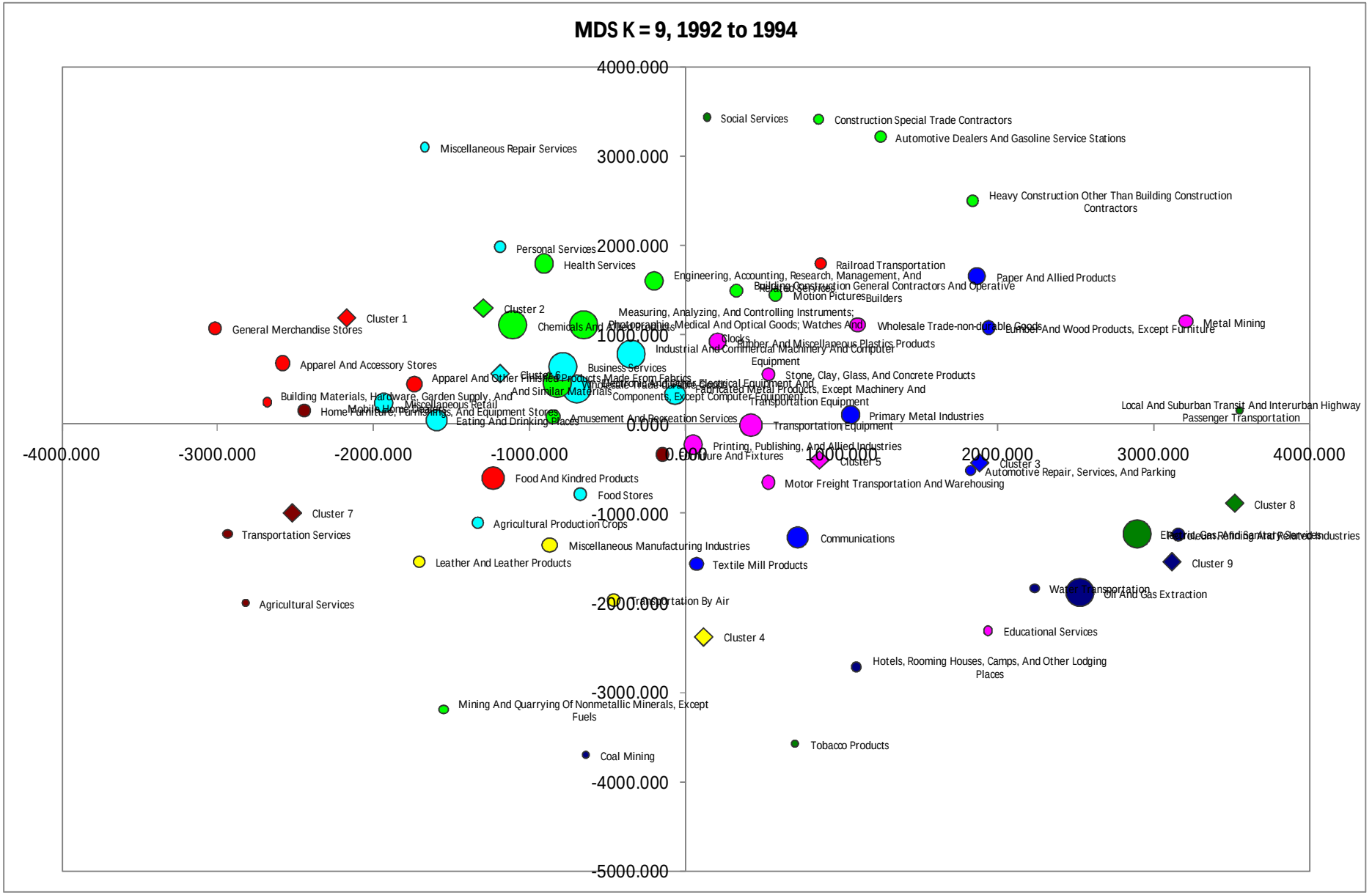


Figure 6-9 MDS chart for the K = 9 solution, 1992 to 1994.

6.3.8 Cluster Interpretations

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-6 shows these industries as well as cluster interpretations.

Cluster	Industry	Mean	SD	Interpretation
1	• Apparel And Accessory Stores	0.0010	0.0583	Retail
2	• Chemicals and Allied Products	-0.0022	0.0491	Health
	• Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0068	0.0458	
	• Health Services	0.0060	0.0586	
3	• Primary Metal Industries	0.0177	0.0490	Metal and paper
	• Paper And Allied Products	0.0096	0.0354	
4	• Leather And Leather Products	0.0166	0.0463	Apparel and Furnishing
5	• Motor Freight Transportation And Warehousing	0.0217	0.0348	Transportation
	• Transportation Equipment	0.0190	0.0433	
6	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0256	0.0513	Elaborated transformed manufactures
	• Industrial And Commercial Machinery And Computer Equipment	0.0190	0.0460	
7	• Agricultural Services	0.0160	0.0787	Generic
	• Furniture And Fixtures	0.0166	0.0470	
8	• Electric, Gas, And Sanitary Services	0.0038	0.0219	Utilities
	• Local And Suburban Transit And Interurban Highway Passenger Transportation	-0.0239	0.1801	
9	• Oil And Gas Extraction	0.0161	0.0469	Primary resources
	• Petroleum Refining And Related Industries	0.0007	0.0333	

Table 6-6 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 1992 to 1994 modelling interval

Cluster 1 represents the **Retail** sector. The *Apparel and Accessory stores* Major group is primarily involved in the retailing of apparel goods. Furthermore, analysis of the MDS perceptual map indicates the other Major groups that have been allocated to this cluster include: *General Merchandise Stores, Apparel And Other Finished Products Made From Fabrics And Similar Materials, Food and Kindred Products*. These industries are primarily involved in retail sales. This sector has relatively high levels of risk but low levels of return.

Cluster 2 represents the **Health** sector. Many industries within the *Chemicals and Allied Products* Major group are involved in the manufacture of medical and pharmaceutical products while many of

the industries within the *Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks* are involved in the production of medical instruments, imaging equipment (X-rays, etc.) and other medical apparatus (e.g. *Surgical and Medical Instruments, Electromedical & Electrotherapeutic Apparatus*, etc.). Industries within the *Health services* Major group are involved in the provision of medical and allied health services. This sector has relatively moderate to high levels of risk but low levels of return.

Cluster 3 represents the **metal and paper** sector. It is a combination of the manufacturing and paper clusters from Section 6.3.2. It has moderate levels of risk and moderate levels of return.

Cluster 4 represents the **Apparel and furnishing** sector. The *Leather and Leather Products* Major group is involved in the manufacture of leather clothing, footwear, furniture and so forth. It has moderate levels of risk and moderate levels of return.

Cluster 5 represents the **Transportation** sector. The *Motor Freight Transportation and Warehousing* Major group consists of industries such as: *Trucking & Courier Services, Public Warehousing & Storage, Terminal Maintenance Facilities for Motor Freight Transport* and so forth. The *Transportation Equipment* Major group contain industries that relate to the production and maintenance of various transportation equipment such as motor vehicles, trucks, buses, commercial transportation, aircraft, ship and rail. It has low to moderate levels of risk and moderate to high levels of return. In particular, the *Motor Freight Transportation And Warehousing* Major group exhibits strong performance with low levels of risk coupled with high levels of return.

Cluster 6 represents the **Elaborately Transformed Manufacturing** sector. It is similar to the ETM cluster from previous solutions (see Sections 6.3.4 and 6.3.6). It has moderate to high levels of risk and moderate to high levels of return.

Cluster 7 may represent a **Generic** cluster as it contains disparate industries such as *Agricultural services, Furniture and fixtures* and *Transportation services*.

Clusters 8 and 9 represent the **Utilities** sector and **Primary** resources sectors respectively. They are similar to the **Utilities** and **Primary resources** clusters from previous solutions (see Sections 6.3.2, 6.3.4 and 6.3.6). The **Utilities** sector exhibits mixed performance with strong performance from the *Electric, Gas, And Sanitary Services*, which is consistent with previous results but very poor

performance from the *Local and Suburban Transit and Interurban Highway Passenger Transportation* Major group indicating potential misallocation into the **Utilities** cluster.

6.3.9 1995 to 1997

The next modelling interval is 1995 to 1997. As with the previous results, the Gap statistic test is used to help determine *K**. This is shown in Figure 6-10. The largest statistically significant difference occurs at K = 10 hence K* suggested by this test is 10.

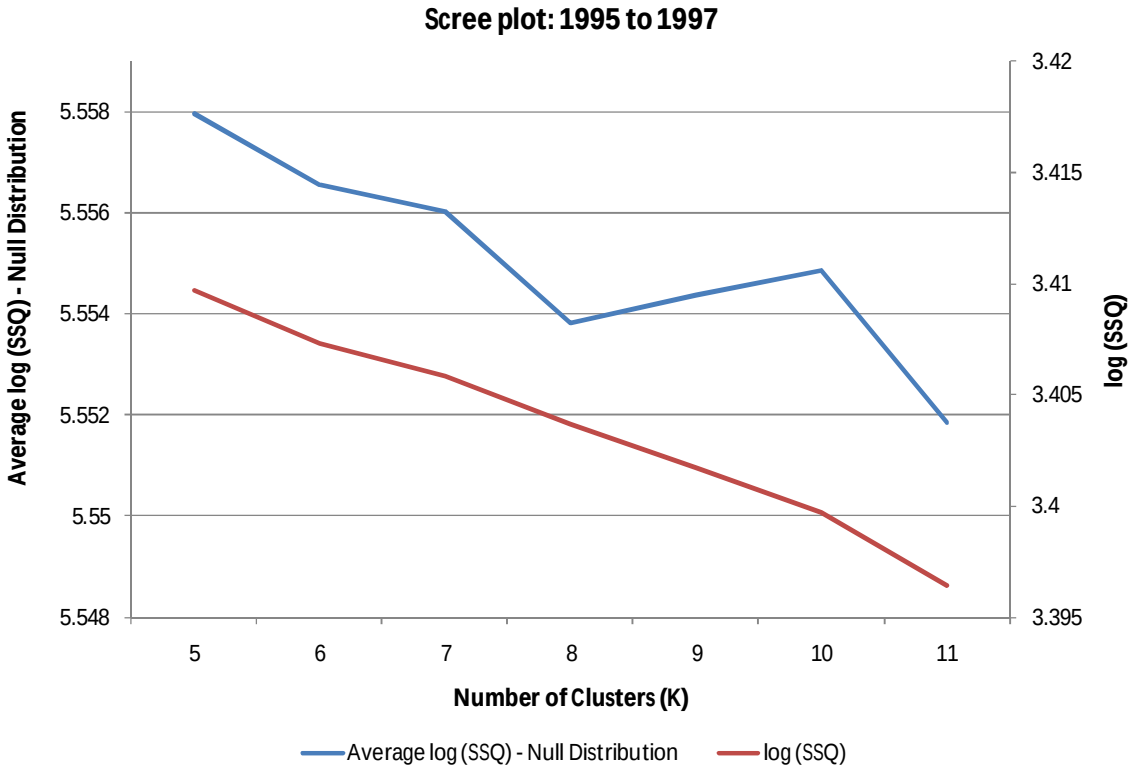


Figure 6-10 The scree plot for the interval 1995 to 1997. The largest statistically significant difference occurs at *K* = 10 hence *K** suggested by this test is 10.

The squared rank correlation matrix is located in the appendices in Table 10-5. Figure 6-11 shows the MDS perceptual map.

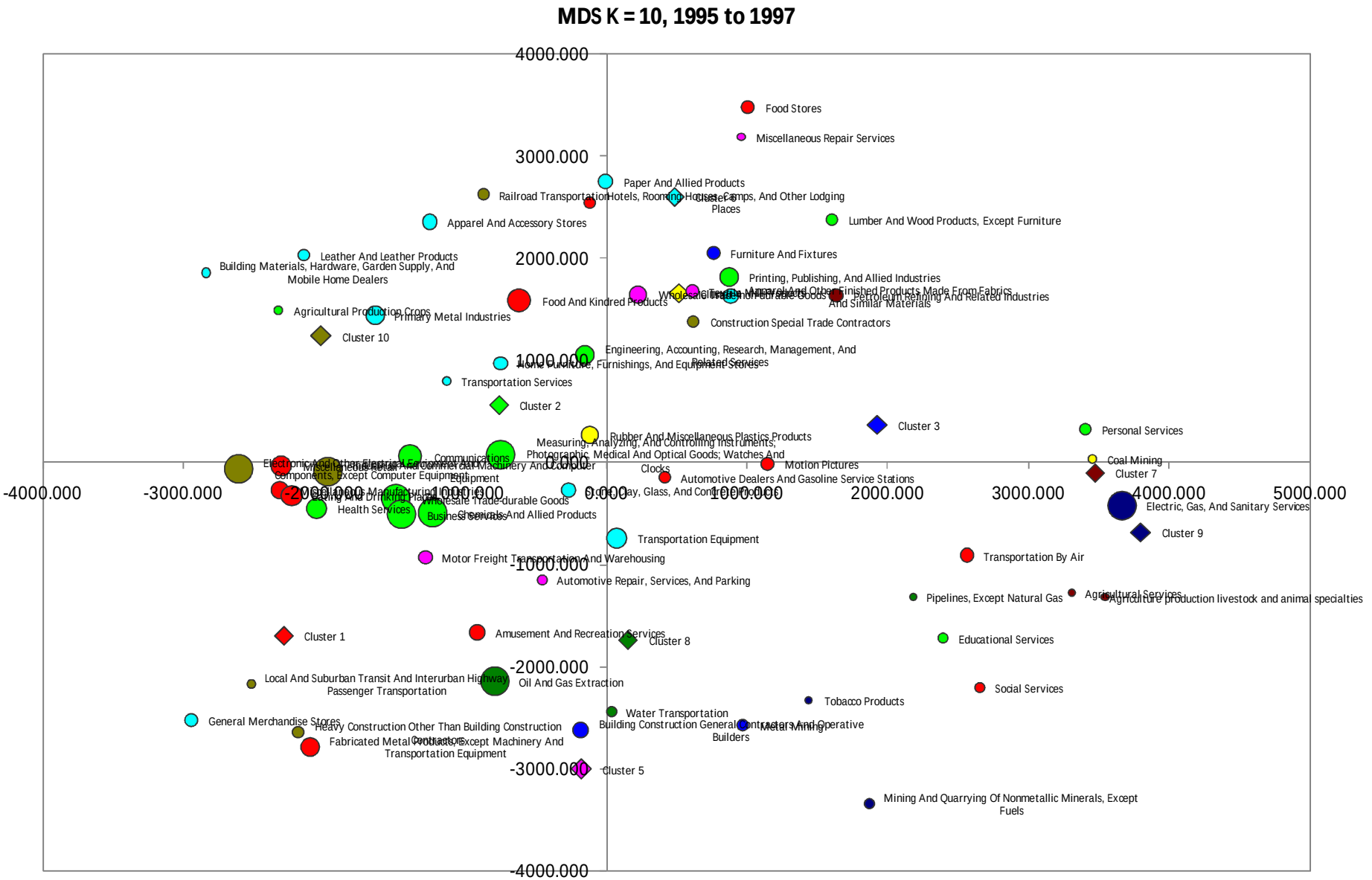


Figure 6-11 MDS chart for the K = 10 solution, 1995 to 1997.

6.3.10 Cluster Interpretations

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-7 shows these industries as well as cluster interpretations, average monthly industry returns and standard deviations across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Amusement And Recreation Services	0.0096	0.0605	Entertainment
	• Motion Pictures	0.0026	0.0528	
	• Eating And Drinking Places	0.0069	0.0464	
2	• Business Services	0.0222	0.0623	Business and information services and specialist equipment
	• Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0232	0.0536	
	• Communications	0.0168	0.0509	
3	• Furniture and Fixtures	0.0230	0.0421	Generic
4	• N/A			Generic
5	• N/A			Generic
6	• Paper And Allied Products	0.0126	0.0399	Manufacturing
	• Primary Metal Industries	0.0138	0.0459	
	• Transportation Equipment	0.0171	0.0346	
7	• N/A			Generic
8	• Oil And Gas Extraction	0.0283	0.0564	Primary resources
	• Water Transportation	0.0182	0.0502	
9	• Electric, Gas, And Sanitary Services	0.0174	0.0244	Utilities
10	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0198	0.0669	Elaborately transformed manufactures
	• Industrial And Commercial Machinery And Computer Equipment	0.0205	0.0650	

Table 6-7 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 1995 to 1997 modelling interval.

Cluster 1 represents the **Entertainment** sector. The *Amusement and Recreation Services* Major group is involved in the provision of various amusement services such as sporting clubs, racing and miscellaneous recreational services. The *Motion Pictures* Major group includes *motion picture distribution, movie theatres, video tape rental and motion picture production*. Lastly, the *Eating and Drinking Places* Major group is involved in food and beverage services such as restaurants, cafes, cafeterias, food courts and so forth. This sector has moderate levels of risk but generates below average levels of return.

Cluster 2 represents the **Business and information services and specialist equipment** sector. This sector deals with professional services such as accounting, advertising, information and

communications technology (ICT) as well as the production of specialist equipment such as laboratory apparatus, optical instruments and lenses, surgical and medical instruments, imaging equipment and so forth. This sector has moderate to high levels of risk and generates moderate to high levels of return.

Cluster 3 represents a **generic** cluster. Even though it contains the *Furniture and Fixtures* Major group, this industry is likely too small to be significant. Similarly, Clusters 4,5 and 7 also represent **generic** clusters.

Cluster 6 represents the **Manufacturing** sector. Many of the processes in the *Primary Metal Industries, Transportation Equipment and Paper* Major groups involve manufacturing or transformation of raw materials into semi-finished and finished products. This sector has low levels of risk and generates low levels of return.

Clusters 8 and 9 represent the **Primary resources** and **Utilities** sectors respectively. They are similar to the **Primary resources** and **Utilities** clusters from earlier solutions. The **Primary resources** sector has moderate to high levels of risk and generates high levels of return. In particular, the *Oil and Gas Extraction* Major group has among the highest levels of return. The performance of this sector has been poor in previous results but has substantially improved indicating that the sector as a whole is a volatile but potentially profitable investment prospect. During poor periods, risk levels are high and returns are low however during good periods, risk levels are moderate and returns are high. This is consistent with the nature of the sector which is characterised by high exploration costs and uncertain payoffs. Volatility in resource prices also add to the complexity of the sector.

The **Utilities** sector on the other hand has low levels of risk but consistently generates moderate levels of return – the reasons for which have been previously discussed.

Cluster 10 represents the **Elaborately transformed manufacturing** sector. It is similar to the ETM clusters found in earlier solutions. See sections 6.3.4, 6.3.6 and 6.3.8. This sector has high levels of risk and generates moderate to high levels of return and is consistent with previous results.

6.3.11 1998 to 2000

The next modelling interval is 1998 to 2000. As with the previous results, the Gap statistic test is used to help determine *K**. This is shown in Figure 6-12. The largest statistically significant difference occurs at K = 8 hence K* suggested by this test is 8.

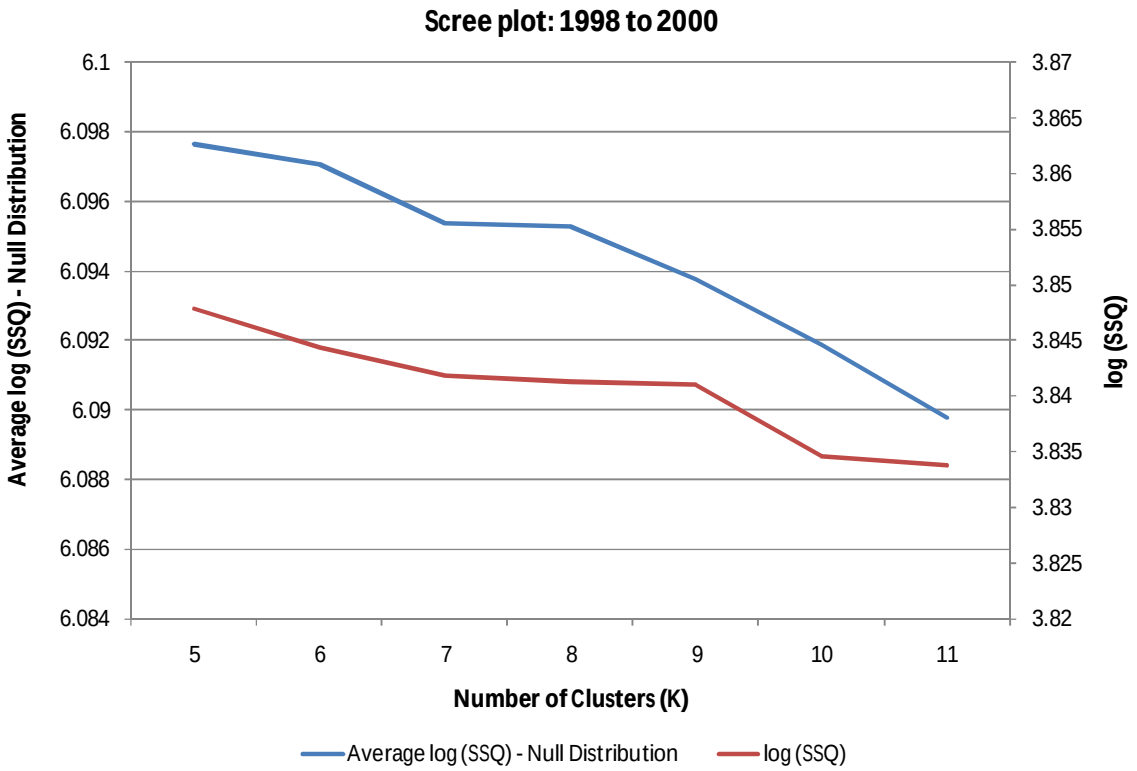


Figure 6-12 The scree plot for the interval 1998 to 2000. The largest statistically significant difference occurs at *K* = 8 hence *K** suggested by this test is 8.

The squared rank correlation matrix is located in the appendices in Table 10-6. Figure 6-13 shows the MDS perceptual map.

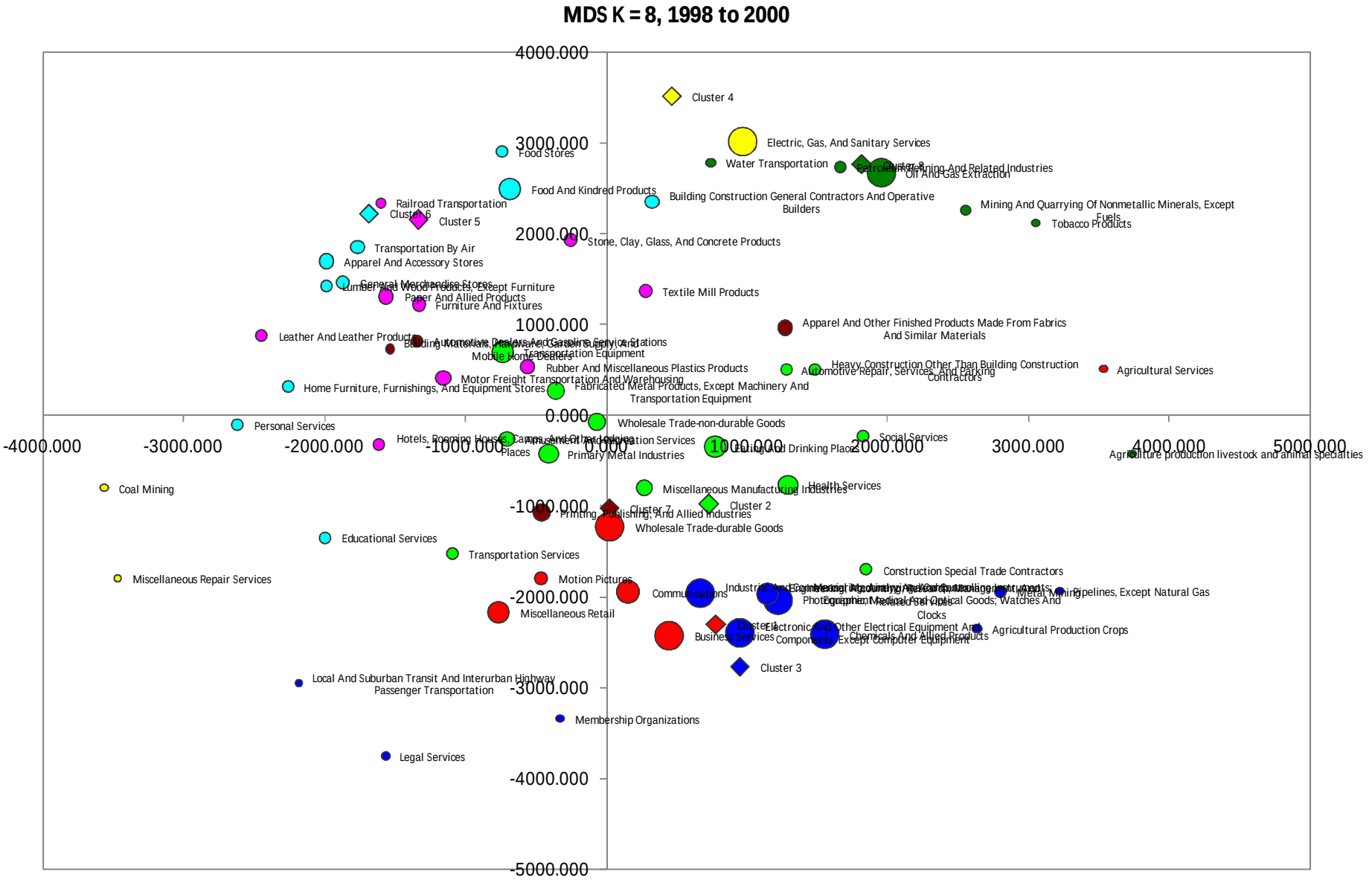


Figure 6-13 MDS chart for the K = 8 solution, 1998 to 2000.

6.3.12 Cluster Interpretation

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-8 shows these industries as well as cluster interpretations, average monthly industry returns and standard deviations across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Business services	0.0125	0.1311	Business and information services
	• Communications	0.0167	0.1047	
2	• Miscellaneous Manufacturing Industries	-0.0052	0.0596	Manufacturing
	• Transportation Equipment	-0.0041	0.0547	
3	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0271	0.1250	Elaborately transformed manufactures
	• Industrial And Commercial Machinery And Computer Equipment	0.0150	0.0976	
	• Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0149	0.0899	
	• Chemicals and Allied Products	0.0250	0.1164	
4	• Electric, Gas, And Sanitary Services	0.0113	0.0362	Utilities
5	• Paper And Allied Products	-0.0025	0.0540	Generic
	• Railroad Transportation	0.0046	0.0530	
	• Furniture and Fixtures	-0.0017	0.0547	
6	• Transportation by Air	0.0049	0.0659	Retail
	• Apparel and Accessory Stores	0.0144	0.0788	
	• Food and Kindred Products	0.0044	0.0410	
	• General Merchandise Stores	-0.0042	0.0648	
7	• N/A			Generic
8	• Oil And Gas Extraction	0.0147	0.1119	Primary resources
	• Water Transportation	0.0009	0.0674	
	• Petroleum Refining And Related Industries	0.0033	0.0628	

Table 6-8 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 1998 to 2000 modelling interval.

Cluster 1 represents the **Business and information services** sector. This sector deals with the provision of professional business services and ICT products. It is similar to the **Business and information services and specialist equipment** from the previous solution. See section 6.3.10. This sector has relatively high levels of risk but generates moderate levels of return.

Cluster 2 represents the **Manufacturing** sector. The *Miscellaneous Manufacturing Industries* Major group is involved in the manufacture of various goods such as sporting and athletic equipment, jewelry, precious metals and so forth. The *Transportation Equipment* Major group is involved in the

production of motor vehicles, aircraft, rail transportation, defence vehicles and ordinance, ships and so forth. This is similar to other **Manufacturing** clusters from previous solutions. This sector has moderate levels of risk but generates negative returns indicating that this period was particularly bad for the manufacturing sector.

Cluster 3 represents the **Elaborately Transformed Manufacturing** sector. This is similar to other ETM clusters found in previous solutions. It has relatively high levels of risk and generates moderate to high levels of return.

Cluster 4 represents the **Utilities** sector. This is similar to other **Utilities** clusters found in previous solutions. As with other **Utilities** clusters, this sector has relatively low levels of risk and generates moderate to high levels of return.

Cluster 5 represents a **Generic** cluster as it contains disparate industries such as: *Paper and Allied Products, Railroad Transportation, Furniture and Fixtures* and so forth.

Cluster 6 represents the **Retail** sector. This sector is primarily involved in the selling of finished goods such as apparel, food and beverage, miscellaneous Retail Services and so forth. The *Transportation by Air* major group consists of industries such as air transportation and courier services. The inclusion of Transportation by Air may at first seem odd however this is logical when one considers the contribution of tourism to the retail sector hence the two sectors are closely linked. As previously discussed, industries such as retail and tourism (which represent non-essential goods and services) have greater exposure to economic factors such as employment and prices (inflation) which can all affect household disposable income. This sector has moderate to high levels of risk but generates low or negative returns. The only exception is the *Apparel and Accessory Stores* Major group which has moderate levels of return.

Cluster 8 represents the **Primary resources** sector. This is similar to other **Primary resources** clusters found in previous solutions. This sector has moderate to high levels of risk but generates low levels of return with the exception of the *Oil and Gas Extraction* Major group which has relatively high levels of return.

Overall, the market performed poorly during this period. Many sectors exhibited high levels of risk but generated low to negative returns. The only exception is the **Utilities** sector which performs well regardless of overall market performance.

6.3.13 2001 to 2003

The next modelling interval is 2001 to 2003. As with the previous results, the Gap statistic test is used to help determine *K**. This is shown in Figure 6-14. The largest statistically significant difference occurs at K = 8 hence K* suggested by this test is 8.

Figure 6-14 The scree plot for the interval 2001 to 2003. The largest statistically significant difference occurs at *K* = 8 hence *K** suggested by this test is 8.

The squared rank correlation matrix is located in the appendices in Table 10-7. Figure 6-15 shows the MDS perceptual map.

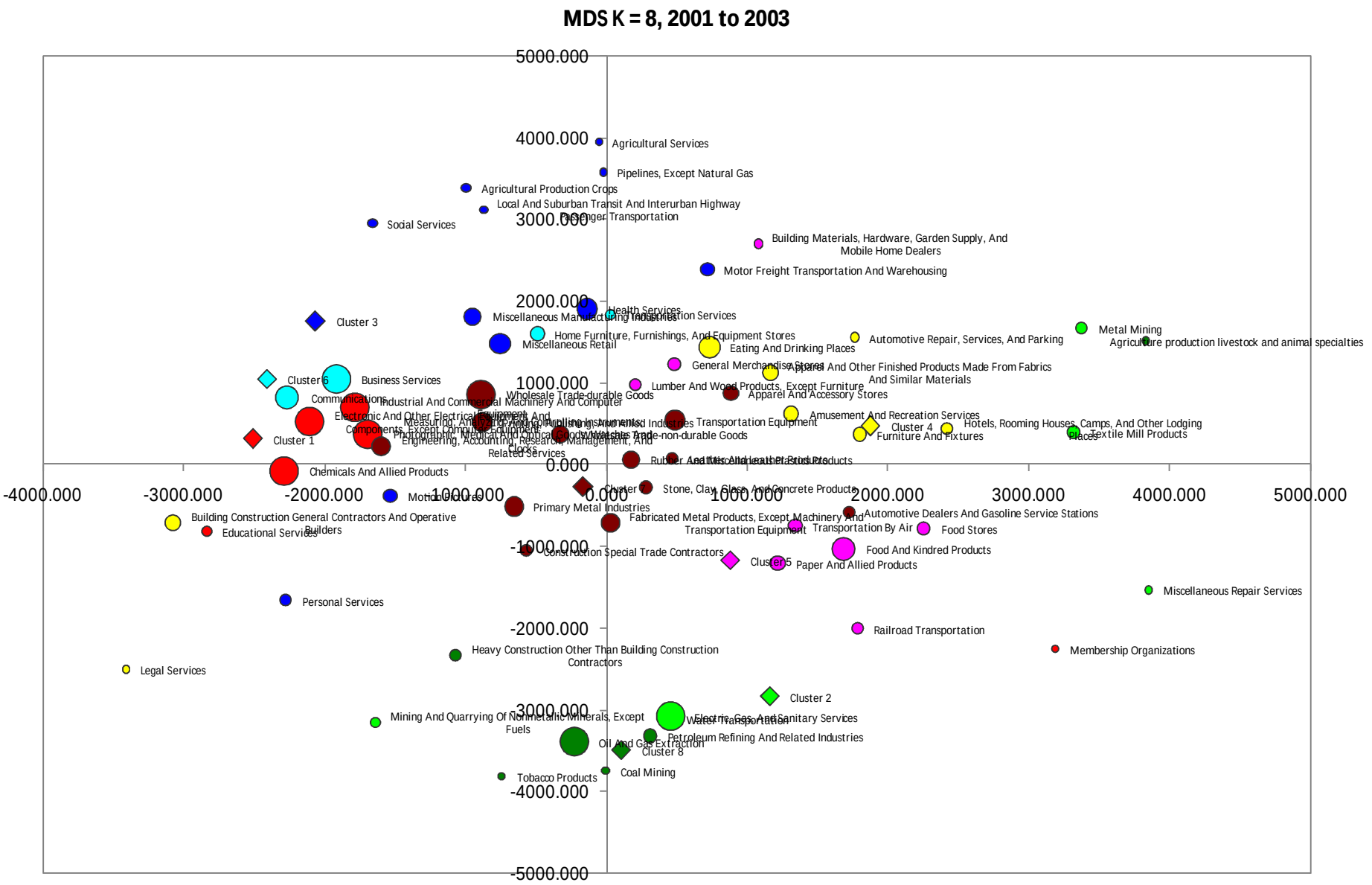


Figure 6-15 MDS chart for the K = 8 solution, 2001 to 2003.

6.3.14 Cluster Interpretation

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-9 shows these industries as well as cluster interpretations, average monthly industry returns and standard deviations across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0201	0.1310	Elaborately Transformed Manufactures
	• Industrial And Commercial Machinery And Computer Equipment	0.0240	0.1121	
2	• Electric, Gas, And Sanitary Services	0.0067	0.0472	Utilities
3	• Local And Suburban Transit And Interurban Highway Passenger Transportation	-0.0621	0.1903	Local transportation services
4	• Eating and Drinking Places	0.0204	0.0619	Entertainment
5	• Paper And Allied Products	0.0125	0.0540	Generic
6	• Business services	0.0279	0.1385	Business and information services
	• Communications	0.0107	0.1207	
7	• Stone, Clay, Glass, And Concrete Products	0.0231	0.0674	Generic
	• Wholesale Trade-durable Goods	0.0288	0.0785	
	• Apparel And Accessory Stores	0.0214	0.0966	
8	• Oil And Gas Extraction	0.0108	0.0803	Primary resources
	• Water Transportation	0.0085	0.0683	
	• Coal Mining	0.0300	0.1175	

Table 6-9 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 2001 to 2003 modelling interval.

Cluster 1 represents the **Elaborately Transformed Manufacturing** sector. This is similar to other ETM clusters found in previous solutions. The sector has relatively high levels of risk and generates high levels of return.

Cluster 2 represents the **Utilities** sector. This is similar to other **Utilities** clusters found in previous solutions. It has relatively low levels of risk and generates low levels of return.

Cluster 3 represents the **Local transportation services** sector. The *Local And Suburban Transit And Interurban Highway Passenger Transportation* Major group consists of industries such as: *Local and Suburban Transit, Local Passenger Transportation, Intercity and Rural Bus Transportation, Taxicabs* and so forth. Note that this sector is not involved with the manufacture or production of transportation vehicles as this would be classified under manufacturing. It also does not include

transportation services such as freight or courier services. This sector has high levels of risk but generates negative return.

Cluster 4 represents the **Entertainment** sector. It is separate (although linked) to food production or retail. It has been classified as **Entertainment** rather than food or retail since the majority of 'added value' in the food and beverage industry has more to do with services such as food preparation and customer service thus making it more similar to the entertainment sector. This sector has high levels of risk and generates high levels of return.

Clusters 5 and 7 have been labelled as **Generic** clusters as they contain disparate industries.

Cluster 6 represents the **Business and Information services** sector and is similar to other **Business and information services** clusters from previous solutions. This sector has moderate levels of risk and generates moderate to high levels of return.

Cluster 8 represents the **Primary resources** sector and is similar to other **Primary resources** clusters from previous solutions. This sector has moderate to high levels of risk and generates moderate levels of return.

6.3.15 2004 to 2006

The final modelling interval is 2004 to 2006. As with the previous results, the Gap statistic test is used to help determine K^*. This is shown in Figure 6-16. The largest statistically significant difference occurs at $K = 9$ however analysis of the cluster solution indicates that $K = 10$ produces slightly better interpretive results so K^* will be set to 10.

Figure 6-16 The scree plot for the interval 2004 to 2006. The largest statistically significant difference occurs at $K = 9$ however K^* will be set to 10.

The squared rank correlation matrix is located in the appendices in Table 10-8. Figure 6-17 shows the MDS perceptual map.

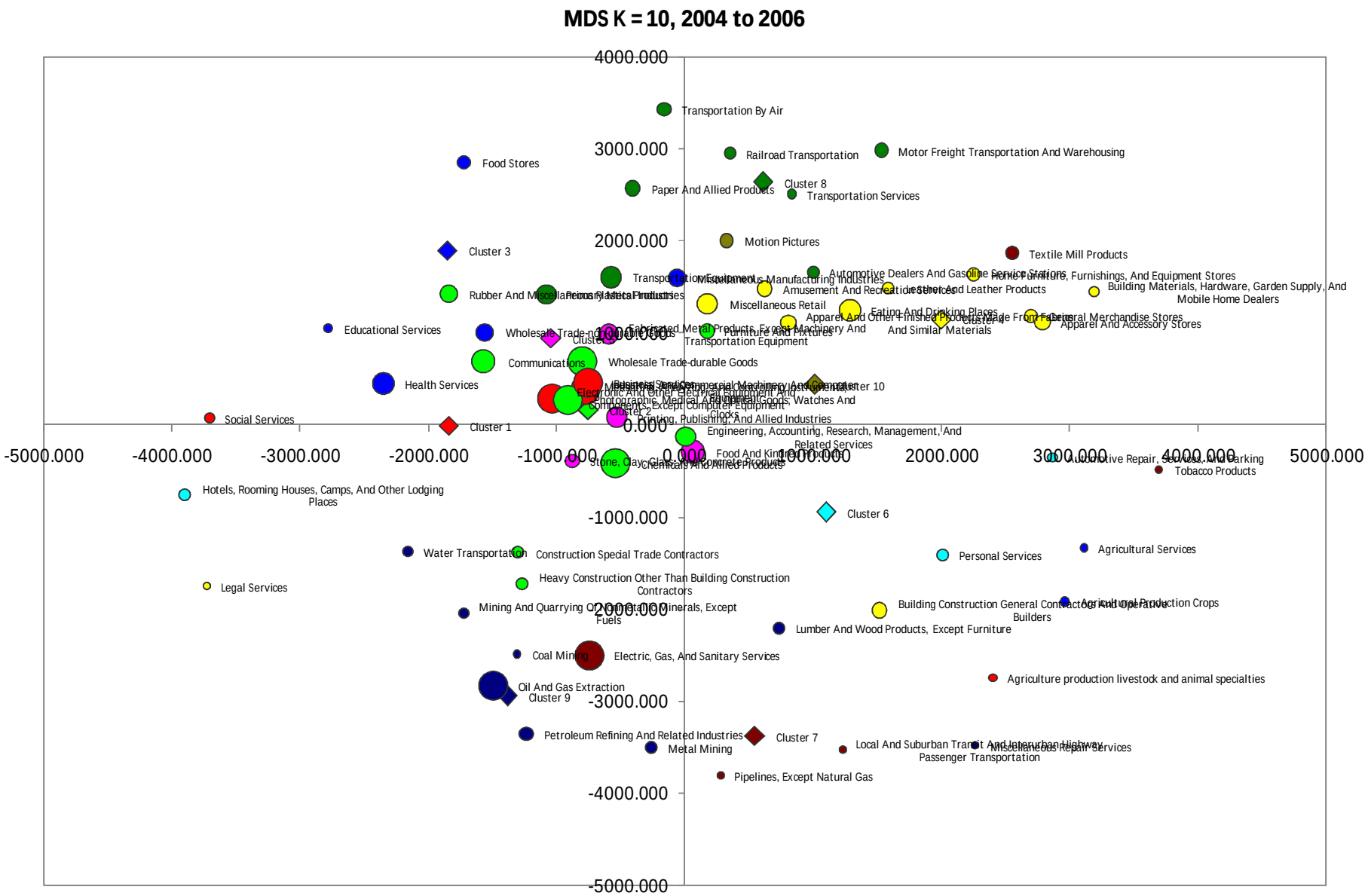


Figure 6-17 MDS chart for the K = 10 solution, 2004 to 2006.

6.3.16 Cluster Interpretation

As before, the squared rank correlation matrix was filtered to remove industries which had lower squared rank correlations than the 95 percent cut-off limit. Table 6-10 shows these industries as well as cluster interpretations, average monthly industry returns and standard deviations across the 36 month modelling interval.

Cluster	Industry	Mean	SD	Interpretation
1	• Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0071	0.0608	Elaborately Transformed Manufactures and Business services
	• Industrial And Commercial Machinery And Computer Equipment	0.0133	0.0532	
	• Business Services	0.0106	0.0503	
2	• Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0108	0.0485	Research and Technology
	• Engineering, Accounting, Research, Management, And Related Services	0.0136	0.0492	
3	• N/A			Generic
4	• Apparel And Accessory Stores	0.0232	0.0525	Retail and Entertainment
	• Miscellaneous Retail	0.0131	0.0439	
	• Eating and Drinking Places	0.0152	0.0377	
	• Leather and Leather Products	0.0161	0.0460	
5	• Fabricated Metal Products, Except Machinery And Transportation Equipment	0.0192	0.0395	Basic metal products
6	• N/A			Generic
7	• Electric, Gas, And Sanitary Services	0.0151	0.0248	Utilities
8	• Railroad Transportation	0.0210	0.0460	Transportation
	• Motor Freight Transportation And Warehousing	0.0145	0.0517	
	• Transportation Services	0.0163	0.0512	
	• Transportation Equipment	0.0119	0.0488	
9	• Oil And Gas Extraction	0.0313	0.0760	Primary resources
	• Petroleum Refining And Related Industries	0.0295	0.0580	
	• Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	0.0243	0.0643	
	• Coal Mining	0.0289	0.1049	
10	• Motion Pictures	0.0023	0.0444	Entertainment

Table 6-10 This table shows the industry major groups that exceed the 95 percent cut-off for the squared rank correlations as well as potential cluster interpretations for the 2004 to 2006 modelling interval.

Cluster 1 represents the **Elaborately transformed manufactures** and **Business services** sector. It is a combination of the **ETMs** and **Business services** clusters from previous solutions. Both sectors

involve the production of highly value added goods and services. This sector is exposed to relatively high levels of risk but generate moderate levels of return.

Cluster 2 represents the **Research and technology** sector. The *Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks* Major group has many industries that deal with scientific and biomedical research as well as the production of scientific equipment and supplies. The *Engineering, Accounting, Research, Management, And Related Services* Major group also has many industries that deal with scientific research although there is a greater emphasis on research and specialty consulting services. However, the grouping of engineering to accounting and management research as prescribed by the SIC scheme appears odd. While both areas are involved in the provision of research and consultancy services, the respective disciplines of engineering, accounting and management are disparate in nature. Such groupings may be the result of the grouping algorithms used by SIC, which emphasize similarity in production processes as opposed to functional outcome groups or even homogeneity in economic markets – a common criticism of the SIC scheme (Clarke, 1989). Nonetheless, the Major groups in cluster 2 share highly similar levels of risk and return, which are both moderate compared to other clusters.

Cluster 4 represents the **Retail and entertainment** sector. This is similar to other **Retail and entertainment** clusters from previous solutions. This sector has moderate to high levels of risk and moderate to high levels of return.

Cluster 5 represents the **Basic metal products** sector. It has low levels of risk and moderate levels of return.

Cluster 7 represents the **Utilities** sector and is similar to other **Utilities** clusters from previous solutions. It has low levels of risk and moderate levels of return. This strong performance is consistent with previous solutions.

Cluster 8 represents the **Transportation** sector. It has moderate to high levels of risk and moderate to high levels of return.

Cluster 9 represents the **Primary resources** sector. It is similar to other **Primary resources** clusters from previous solutions. It has moderate to high levels of risk and generates high levels of return.

This sector has exhibited mixed performance in previous periods. However, it has performed well in the current period generating relatively high returns given its risk exposure.

Cluster 10 represents the **Entertainment** sector. It has low to moderate levels of risk but generates very low return.

Overall, the equity market has performed fairly well in this period with many sectors generating high returns with low risk relative to previous periods.

6.4 Cluster profiles

The objective of this chapter is to explore how stocks may be grouped into returns (and implicitly risk) homogenous clusters and to interpret those clusters in terms of their industry associations. It is not a taxonomy to divide all sectors within the economy. Rather, it is limited only to those sectors that traded on the stock exchange. Public goods and services such as health, education, defence, public transportation, etc. are not considered[42]. Other sectors such as the legal and accounting sectors are underrepresented as ownership is typically privately held. Sectors such as finance, insurance and real estate are also not considered as they are excluded by the filtering conditions (see Section 3.3). Furthermore, many small to medium enterprises, SMEs are also underrepresented as ownership is again often privately held.

Subject to these restrictions, there are several sectors of the economy that are consistently identified as having a unique risk/return profile separate to that of other sectors. The **Primary resources** sector is primarily involved in the extraction and production of raw materials such as mining, oil and gas extraction and petroleum refining (Clarke, 1940). This sector is exposed to many sources of risk such as the heavy capital requirements of exploration operations that have uncertain outcomes. Financial sources of risk include fluctuating commodity prices and since much of these primary resources are exported, exchange rate risk as well. Changes in the legislative and political landscape may also affect the operational environment of the sector. As a result, this sector often experiences periods of turmoil marked by volatility in returns (high risk) and poor returns. During periods of prosperity however, returns are high while risk is kept in check.

[42] There are privatised sectors with close links to the public sector but the bulk of the public sector itself is not considered here.

The secondary sector of the economy manufactures finished and semi-finished goods and includes sectors such as **manufacturing, metal working, construction, chemicals, food and textile production** and so forth. The **manufacturing** sector typically encompasses many manufacturing processes such as automobile production, shipbuilding, aerospace manufacturing, some electronics, etc. This sector has low to moderate levels of risk and low to moderate levels of return. The **metal working** sector is often aligned with the **manufacturing** sector and in some cases placed in the same cluster. This sector involves the production of **metal products** such as fabricated metals, machinery, vehicle components, and building and construction materials. Like the manufacturing sector, the metal working sector also exhibits low to moderate levels of risk with low to moderate levels of return.

The **chemicals** sector is involved in the production of pharmaceutical and medicinal preparations as well as substances that are used in industrial processes. Since the U.S. is a major exporter of pharmaceuticals and chemicals, this sector experiences slightly elevated levels of risk, likely from exposure to external risk (e.g. exchange rate risk, political risk, etc.) but generates moderate to higher levels of return.

The tertiary sector of the economy is primarily comprised of services but may also include some highly value added goods. This sector provides services to businesses and the general population. Commercial services include business services, information and communications technology, research and development, insurance, financial, accounting and legal services, etc. Services to the general population include retailing, transportation and distribution, entertainment, food and beverage, tourism, healthcare and education.

The **Business services, ICT and Research and Development** sectors exhibit high levels of risk but generate high levels of return. The **Retail** sector also has high levels of risk but generally produces relatively low levels of return. Likewise the **Entertainment and Food and Beverage** sectors also exhibit high levels of risk but generally produce low levels of return. The **Transportation** sector on the other hand exhibit moderate levels of risk and return. Note that this sector is not to be confused with the production of vehicles and components as this falls within the manufacturing sector. The transportation sector here is primarily involved in the provision of transportation services such as passenger and freight transportation.

Although **Elaborately Transformed Manufactures** should be classified under the secondary sector (manufacturing), its risk/return profile is more similar to business services and in some cases these sectors are actually grouped into the same cluster. Like the **Business Services** sector, the **ETMs** sector is characterised by high levels of risk and return.

Unfortunately, analysis of accounting and legal services is limited as firms associated with these activities are not commonly traded on the stock exchange. Likewise organisations associated with healthcare and education are not represented on the stock exchange as ownership is typically held by the Government or in private.

The quaternary sector, which consists of intellectual activities such as education and research is underrepresented as these organisations are typically owned by the Government or privately held.

The quinary sector, which includes non-profit activities such as public services, legislation, policy decisions, charities and so forth, is also underrepresented as organisations associated with these activities are not typically traded on the stock exchange. One exception however is the **Utilities** sector, which include activities such as the provision of electric, gas and sanitary services. In the United States many of these sectors are privatised or operate in partnership with the Government. As a result, ownership may be held by the public in the form of equities[43]. According to the clustering results, this sector consistently exhibits low risk and generates moderate returns. This performance is likely due to the proximity of this sector to the Government which may have a stabilizing effect on stock prices; and the necessity of such services regardless of economic conditions.

6.4.1 Stability of the GSC clusters over time

To assess the stability and consistency of the GSC clusters over time, the following table summarises the cluster interpretations across the 8 modelling intervals. Some industry sectors are more consistently observed than others. Primary resources, Utilities, Elaborately Transformed Manufactures (ETMs) and Manufacturing are frequently observed. Furthermore, their risk/return profiles are consistent over the modelling intervals.

[43] Department of Energy (United States) [online] available at: <http://www.oe.energy.gov/information_center/faq.htm#ppl1> [accessed March 2011]

Other sectors such as Retail, Business Services and Transportation are less frequently observed although they do appear in most periods. Their risk/return profiles are also consistent over the modelling intervals.

Sectors such as Research and Technology and Business Information Services (which are largely dominated by technology based firms) are observed with increasing consistency over latter periods (specifically from 1995 onwards) but not earlier. This makes intuitive sense since this sector largely came into prominence in the mid 1990s and onwards. To a large extent the NAICS scheme, which supplanted the SIC scheme in late 1990s was developed with a greater emphasis on the technology sectors. If the modelling period was extended to beyond 2006, it is likely that this sector would again feature prominently.

Clust	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	Generic	Basic metal products	Miscellaneous manufacturing / Generic	Retail	Entertainment	Business and information services	ETMs	ETMs and Business services
2	Manufacturing	Generic	Business services and ETMs	Health	Business and information services and specialist equipment	Manufacturing	Utilities	Research and Technology
3	Utilities	Chemicals	Construction materials	Metal and paper	Generic	ETMs	Local transportation services	Generic
4	Retail	Paper	Generic	Apparel and Furnishing	Generic	Utilities	Entertainment	Retail and Entertainment
5	Basic metal products	ETMs	Retail – Apparel and Furnishing	Transportation	Generic	Generic	Generic	Basic metal products
6	Communications	Generic	Generic / Transportation	ETMs	Manufacturing	Retail	Business and information services	Generic
7	General merchandise	Generic	Food production	Generic	Generic	Generic	Generic	Utilities
8	Primary resources	Primary resources	Utilities	Utilities	Primary resources	Primary resources	Primary resources	Transportation
9	Paper	Utilities	Primary resources	Primary resources	Utilities	-	-	Primary resources
10	Generic	-	-	-	ETMs	-	-	Entertainment

Table 6-11: Cluster interpretations for the 8 (3 year) modelling intervals between 1983 and 2006

6.5 Limitations

This analysis does suffer from several limitations. These include:

- Missing or underrepresented industries
- Only one industry classification scheme is explored
- Industry classifications are explored on a relatively superficial level/Exploring industry classifications at the second level

6.5.1 Missing or underrepresented industries

Only firms (and hence industries) which are traded on the stock exchange are considered in this study. Industries which are comprised of firms which are not commonly traded on the stock exchange are either missing or underrepresented. See Section 6.4 for details. Whether this is a major limitation depends largely on the objective at hand. If the objective is to explore industry risk/return characteristics within an investor's opportunity set, then such missing or underrepresented industries would not represent a significant loss of information since ownership (at least in the form of equities) is not available anyway.

On the other hand, if the objective is to develop an industry classification scheme/taxonomy that groups industries in the entire economy according to similarity in risk and return rather than economic activity, then such missing or underrepresented industries would represent a significant loss of information since these industries represent a significant proportion of an economy.

In either case, including such data would improve the clustering results however it is unlikely that such data would ever become publicly available. For example, legal and accounting firms are under no obligation to disclose their financial data to the general public.

However, it should be noted that this limitation is not a result of a poor research methodology but rather a general limitation that applies to research in this area. If such data did become available, it could easily be accommodated by the GSC procedure.

6.5.2 Industry classification scheme

The Standard Industry Classification (SIC) scheme is used to identify a firm's industry. Stocks are then grouped into clusters via the GSC clustering procedure. The clusters are subsequently 'profiled' based on a stock's industry membership. The accuracy of the profiling process however, depends in part on the ability of the SIC code to correctly match a firm's activities to its corresponding industry. A common criticism of the SIC scheme is that it does not perform this function particularly well (Clarke, 1989). This is due in part to the age of the SIC scheme. Originally established in 1937, it has undergone several revisions to incorporate new and constantly evolving industrial structures but has ultimately been supplanted by the North American Industry Classification System (NAICS) in 1997.

Consider for example the all encompassing and at times, cryptic 'Business Services' SIC Major Group (SIC codes: 7310 to 7389). Under the NAICS scheme there are up to 12 equivalent NAICS 'Sub Sectors' that correspond to the *Business Services* Major Group. These are: *Publishing Industries, Administrative and Support Services, Credit Intermediation and Related Activities, Professional, Scientific, and Technical Services, Internet Service Providers, Web Search Portals, and Data Processing Services, Rental and Leasing Services, Transportation Equipment Manufacturing, Personal and Laundry Services, Computer and Electronic Product Manufacturing, Telecommunications, Other Information Services and Motion Picture* and *Sound Recording* Industries.

This means that under the SIC scheme, a firm may be allocated to the *Business Services* Major Group however under the NAICS scheme, it may be allocated into one of twelve potential Sub Sectors that describe its industry of operation. For example, Yahoo Incorporated (PERMNO: 83435) is allocated to the *Business Services* Major Group under the SIC scheme. However, under the NAICS scheme it is allocated to the *Internet Service Providers, Web Search Portals, and Data Processing Services* Sub Sector. At the same time, the Escala Group Incorporated[44] (PERMNO: 79167) is also allocated to the *Business Services* Major Group under the SIC scheme. However, under the NAICS scheme it is allocated to the *Personal and Laundry Services* Sub Sector. Clearly, Yahoo and the Escala Group have little in common by way of operational activity yet under the SIC scheme would be considered to belong to the same industry thus resulting in some degree of misallocation. In defence of the SIC scheme however, many of the additional NAICS categories and refinements are limited to the information and communications technology sector which has seen substantial change and

[44] Escala are traders of precious metals and other collectables such as coins, stamps and wine.

evolution in the last several decades as opposed to say the Utilities sector, which has undergone less change.

In addition to the SIC and NAICS industry classification schemes, other commonly available classification schemes include the Global Industry Classification Standard (GICS) and the Fama-French industry classification scheme (FF). While these schemes are broadly consistent in the way they allocate firms/stocks to industries, there are some cases where stark differences exist. Consider for example Oracle Incorporated (PERMNO: 10104). Under the SIC, this firm is allocated to *Business Services* while under NAICS, it is allocated to *Publishing Industries*. Under GICS, it is allocated to *Software* and under FF, it is allocated to *Business Services*. Similarly, Research Frontiers Incorporated (PERMNO: 10463) is allocated to *Engineering, Accounting, Research, Management, And Related Services* under SIC, *Professional, Scientific, and Technical Services* under NAICS, *Electronic Equipment, Instruments & Components* under GICS and *Business Services* under FF.

While each industry classification scheme has its respective merits, the cluster profiles may change substantially under different schemes. This represents an area for future research.

6.5.3 Level of detail

In this analysis, stocks are grouped into industrial categories at the second level of detail under the SIC scheme to profile and interpret the GSC clusters. There are 83 industrial categories (known as Major Groups under the SIC scheme) and approximately 8 to 10 clusters are identified in each of the modelling intervals. The second level was chosen as the first level would lead to over-generalisation of the cluster profiles reducing the interpretability of the clusters. The third level of detail, which contains 445 industrial categories (known as Industry Groups) was also not explored as this would provide too much detail. This would again make cluster interpretation difficult. In addition, more clusters would be required thereby increasing the computational burden, especially when performing the Gap statistic test. Depending on the number of clusters, analysis at this level of detail may be prohibitive given the computational resources available at the time of writing.

However, such limitations may be circumvented if the analysis were limited to a given industrial sector of interest. For example, the **Business and Information Services** sector identified by the GSC clustering commonly consists of Major Groups such as **Business Services** and **Communications** among others. Within these two Major Groups, there are a further 29 Industry Groups (the 3rd level).

Cluster analysis may be performed on these 29 Industry Groups to obtain a more comprehensive profile of this sector.

Again, the flexibility of GSC allows it to be applied to any dataset or subset of any data. In theory it would be possible to apply the GSC to various subsets of SIC, at say 3 or even 4-digit SICs (or even under different classification schemes such as NAICS and GICS) to derive a completely new classification system. There are no theoretical limitations on how and in what context the GSC can be applied making it an incredibly flexible clustering algorithm with many potential applications.

Furthermore, since there are no restrictions on what input the GSC requires, users can specify any given variable of choice knowing the GSC will generate the desired groupings based on the specified variable. If say the objective of the study was to classify stocks according to operational characteristics such as R&D expenditure for example, the clustering variable can be replaced with the necessary metric (say standardised R&D expenditure) allowing for a GSC based classification scheme based on such operational characteristics.

Further analysis of the clusters (and the industrial sectors they represent) at deeper levels of SIC (or any other industry classification scheme) will further refine the cluster interpretations providing more detailed insight into the underlying industrial structures within various subsectors of the economy.

6.5.4 The GSC clusters in practice

It is not the intention of this research to question the usefulness of existing industry classification systems. In fact, stocks grouped using classification systems such as SIC, NAICS and GICS have been found to exhibit intra-industry homogeneity in certain accounting ratios such as profitability, liquidity, solvency and asset turnover; as well as other characteristics such as information transfer (Krishnan, Press, 2003; Guenther, Rosman, 1994; Weiner, 2005).

However, where returns are concerned, the evidence presented in this research indicates that stocks grouped via the GSC exhibit greater intra-industry homogeneity (and inter-industry heterogeneity) than equivalent groups formed via existing industry classification schemes. Arranging stocks into homogenous returns groups is not an uncommon practice (in the financial literature or otherwise), nor is attempting to describe these groups in terms of industry affiliation. Even Fama and French

(1997) attach industry names to operationalise their classification scheme. King's (1966) research also found evidence indicating differences in returns across industries. To a large extent, this research is aimed at improving and generalising these studies.

At this stage, the findings from this research are probably unsuitable for application at the practitioner level. Existing industry classification schemes posses varying levels of detail (see Section 6.2.1). Over a thousand distinct industry groups exist at the narrowest level of detail allowing practitioners to investigate securities at very fine levels of 'resolution'. This is a noted limitation of this study (see Section 6.5.3) and represents an area of further work.

6.5.5 Combining the GSC with existing classification schemes

This study implies that a completely new risk adjusted industry classification scheme should be developed to supplement existing industry classification schemes as risk adjusted industry groups can be more appropriate for financial research. However, given the vast number of stocks and the complexity of existing classification schemes[45], developing a completely new classification scheme comparable to existing schemes may be infeasible.

One possible approach would be to combine current industry classification schemes with the GSC. Under this approach, the GSC may be applied within the existing SIC structure. For example, stocks may be first divided into broad industrial groups via say 2-digit SIC. Next, the GSC may be applied within each of these broad groups to derive new risk adjusted sub-industry groups.

This combination approach would be beneficial for a number of reasons. Firstly, it will help achieve some degree of product based homogeneity, from which an economically 'close' set can be extracted. This may ensure more stability in group membership over time thus gaining more 'economic' appeal. Secondly, dealing with smaller datasets (or sub-datasets) reduces computational burden, especially when conducting the GAP statistic test.

Such an approach would retain the essence of the GSC while maintaining the 'economic' structure of existing industry classification schemes thus combining the best elements of each approach. This is outside the scope of this thesis but should be investigated in future research.

[45] recall that over a thousand distinct industry groups exist at the narrowest levels of detail. See section 6.2.1

6.6 Conclusion

Industry classification has been a "long-standing problem in financial research" (Bhojraj et al, 2003). The 'problem' with such traditional classification schemes is that it is not clear how industry groups are formed. Therefore, when researchers divide stocks via these classification schemes, there is no guarantee that such divisions are made in a way that is useful for the research objective at hand. The GSC clustering algorithm is the solution to that problem. Under the GSC, the user may specify the metric of interest allowing a great deal of control over how industry groups are formed. If the objective is to study industrial grouping based on R&D expenditure for example, the user may simply replace the N x T array of returns with an equivalent array of standardised R&D expenditures.

In this chapter, the GSC clustering algorithm has been shown to be an effective means of classifying stocks based on their return (and implicitly risk) profile. Such a classification scheme provides an addition to traditional classification schemes such as SIC, NAICS and GICS that rely on similarities in unrelated characteristics such as production technology or product use, to group firms.

The GSC represents an alternative to these common industry classification schemes. While industry classification schemes allocate firms into predefined categories with little explanation as to how or what these categories mean, the GSC allows users a high degree of flexibility in how these categories can be formed. Although the Gap statistic test provides an objective test for determining K, this decision can ultimately be modified by the user, providing control over the number of industrial categories to be formed. Furthermore, as the clusters are reformed after each modelling interval, stocks are not 'locked in' to a particular industry group allowing stocks to migrate across clusters as their returns profile changes over time. Most importantly, the GSC is entirely data driven and operates on the sole basis of matching firms via proven mathematical principles (such as minimising the total sum of squares). It cannot be criticised for having unclear parameters for making subjective stock groupings purely because it has none but rather allows the data to 'speak for itself' rather than imposing an artificial construct determined *a priori*.

Approximately 10 clusters are identified in each modelling interval indicating that at the broadest level, there are approximately 10 coarse groups of stocks. Each cluster represents a group of stocks that have their own unique risk/return profile. While individual differences exist, an overall trade-off between risk and return is observed, which is consistent with finance theory. There are some minor exceptions such as firms operating in the Utilities sector which have low levels of risk but

consistently generate moderate returns. The proximity of this sector to the Government may have a stabilising effect on stock prices reducing volatility.

Clustering at finer levels of detail is of course possible however this is beyond the scope of this chapter and represents an area for future research. With clustering at finer levels, it is possible that a 'proper' classification scheme with multiple levels of detail may be produced to be used alongside current industry classification schemes.

Such a GSC classification scheme is not intended to replace schemes such as SIC, NAICS or GICS but may be used as a complementary scheme. The GSC classification scheme would be useful for research such as benchmarking studies that require stocks to be divided into homogenous groups based on risk/return rather than economic activity.

The focus of this chapter is to showcase the ability of the GSC to form returns-based industry clusters and suggest potential areas of application. The next chapter however, will provide details of a more 'hands on' application of this technology in obtaining superior cost of capital estimates.

Chapter 7

7 Toward better Industry Cost of Capital estimates

7.1 Introduction

In the previous chapter, the GSC is shown to be an effective method of grouping stocks into risk/return homogenous groups. Such groupings have the potential to improve capital market research in a number of ways including the identification of control firms for benchmarking; as well as describing industrial structure. In addition, further application of the GSC at various 'sub levels' of detail may eventually lead to a new industrial classification schemes based on returns as opposed to the existing SIC, NAICS and GICS classification schemes which rely on economic or business activity.

In this chapter, an altogether different yet highly relevant application of the GSC is presented. This application is in the area of corporate finance. In corporate finance, the cost of capital is an integral component in net present value calculations (NPV) which is used in the capital budget decision making process of all firms. Finance theory (Copeland, 2004) states that wealth maximisation occurs when firms invest in all projects that generate a positive NPV. A crucial component in the NPV calculation however is the choice of discount rate. An overestimated discount rate may lead firms to reject projects that may otherwise be profitable while an underestimated discount rate may lead firms to accept projects that fail to generate the required return. The cost of capital, or more specifically the weighted average cost of capital is commonly used as the discount rate in NPV calculations.

Therefore, the cost of capital if improperly estimated may lead to misallocation of funds and ultimately financial loss. It is in this context that the GSC has a contribution to make as it can improve cost of capital estimates.

7.1.1 Net present value

The NPV provides investors with a single number that indicates the value of an investment after future cash flows are discounted. The choice to investors is therefore simple: projects that have a positive NPV generate an excess in cash flow and should therefore be undertaken while projects that have a negative NPV generate a shortfall in cash flow and should therefore be rejected. Consider the

following example. A project initially costs $1000 but generates a net income of $200 per year for the next 6 years. Figure 7-1 depicts the income stream:

Year	0	1	2	3	4	5	6
Income	-$1000	$200	$200	$200	$200	$200	$200

Figure 7-1 In Year 0, the income is negative $1000 indicating the initial cost of the project. However, Years 1 through 6 generate net income of $200 each year.

For simplicity, assume there is no salvage value. That is, nothing that can be sold at the end of Year 6. *Prima facie*, the total value of future income payments (=$1200) exceeds the initial outlay of $1000. The project generates an excess of $200 and should therefore be undertaken.

This evaluation however does not take into account the *time value of money*. Given a choice between $200 today and $200 in the future, say in one year, an investor will always prefer $200 today. If an investor receives $200 today, that amount can be re-invested in another investment that yields interest. Suppose that over the course of one year, an investor can earn 5% interest by investing in a government security such as a treasury bill. After one year, the investment is worth $210 (=$200 + 0.05 × $200). Therefore, $200 in the future does not represent the same value as $200 today. In fact, the investor has to be offered $210 in one year to make him/her indifferent between receiving the funds today and receiving the funds in the future.

The process of evaluating the future value of funds (by applying some rate of interest) is referred to as *capitalisation*. The opposite of capitalisation, i.e. evaluating the value of future funds in today's terms is referred to as *discounting*. To discount $210 by one year, the following calculation is performed:

$$Present\ Value = \frac{Future\ Value}{(1 + r)^t}$$

$$\$200 = \frac{\$210}{(1 + 0.05)^1}$$

Where r = interest rate

t = number of periods in the future

Note that the effects of inflation can be removed by substituting the nominal interest with the real interest rate via the following: *real interest rate = nominal interest rate – inflation*

Returning to the original example depicted in Figure 7-1, the future stream of payments must be *discounted* to evaluate the value of the project in today's terms, i.e. the present value. Extending the present value calculation shown above to incorporate multiple future periods produces the *Net Present Value*. The NPV is given by the following:

$$NPV = I_0 + \frac{I_1}{1+r} + \frac{I_2}{(1+r)^2} + \cdots + \frac{I_t}{(1+r)^t}$$

Where: I_0 = initial cost of project

I_t = income in period t

r = discount rate/interest rate

In this calculation, the discount rate plays a key role. Suppose another investment, say a government treasury bill generates 5% interest per year. This investment represents an alternative to the investor and the interest rate therefore represents the *opportunity cost*. That is, the income the investor could have earned had they invested in treasury bills. Applying a 5% discount rate to the stream of future payments yields:

Year	0	1	2	3	4	5	6	NPV
Income	-$1000	$200	$200	$200	$200	$200	$200	
Discounted Income	-$1000	$191	$181	$173	$165	$157	$149	$15

Figure 7-2 The income payments are discounted using a discount rate of 5%. Payments further in the future are discounted more representing the increasing opportunity cost as more time passes.

The sum of the discounted income payments, i.e. the NPV is $15. As the project generates a positive NPV, it should still be undertaken. That is, the project generates a net profit of $15 in today's terms. Note that while the project still generates a positive NPV, it is much less attractive than the initial assessment in which the project appears to create an excess cash flow of $200.

Now suppose the interest rate on treasury bills rises to 10% per year. Applying this discount rate yields:

Year	0	1	2	3	4	5	6	NPV
Income	-$1000	$200	$200	$200	$200	$200	$200	
Discounted Income	-$1000	$182	$165	$150	$137	$124	$113	-$129

Figure 7-3 The income payments are now discounted using a discount rate of 10%. Future payments are discounted by a greater amount and the NPV becomes negative indicating that the project should be rejected.

The NPV is now -$129. The future stream of payments is not high enough to compensate the investor for the opportunity cost incurred. Since, the treasury bills now pay higher interest, the

investor would be better off investing in treasury bills instead. As the project generates a negative NPV, it should be rejected.

A closely related concept is the *Internal Rate of Return*. The Internal Rate of Return (IRR) is the discount rate at which the NPV becomes zero. In some sense, it is the maximum allowable opportunity cost before the project becomes a liability. In this case, the IRR is 5.47%. Using this discount rate, the NPV becomes:

Year	0	1	2	3	4	5	6	NPV
Income	-$1000	$200	$200	$200	$200	$200	$200	
Discounted Income	-$1000	$190	$180	$170	$162	$153	$145	$0

Figure 7-4 Using a discount rate of 5.47%, the NPV becomes zero. The discount rate which results in zero NPV is known as the Internal Rate of Return.

The NPV is exactly zero. As the project generates zero NPV, the firm would be indifferent as to whether or not to undertake the project. Note that although the financial gain is zero, there may be other objectives (e.g. social or strategic objectives where the financial gain is difficult to quantify) which the project may achieve. In some cases, organisations may even be willing to undertake projects with negative NPV if such projects achieve strategic or other objectives.

What is clear from this example is that the decision of whether or not to undertake a project depends largely on the choice of discount rate. If the discount rate is high, the project will be rejected and if the discount rate is low, the project will be undertaken. What is not clear is *how* this discount rate should be calculated.

Since the discount rate is meant to represent the opportunity cost of investing, one approach is to use the *risk free rate* as the discount rate. The risk free rate is the theoretical rate of return on an investment that is free from default risk. The return on short term government bonds, e.g. 3 month U.S. Treasury bills is often used as the risk free rate.

Using the risk free rate as the discount rate in NPV calculations however implicitly assumes that investors face only one of two options: invest in the project (which undoubtedly carries some degree of risk) or invest in government securities (which theoretically carry zero risk).

Using the risk free rate in this way is inappropriate. Recall that the discount rate in NPV calculations represents the opportunity cost of investing in the project – that is, it is the rate on return on the **next best alternative** available to the investor. Government securities are not necessarily the next

best alternative available to the investor. In fact, government securities are not an appropriate benchmark because such securities do not belong in the same risk class as private investment projects which do carry some degree of risk.

Many projects carry some degree of risk and the rates of return from such projects would be higher than the risk free rate. Using the risk free rate (which is a low rate of return) as the discount rate in NPV calculations would generate a positive NPV and create a bias in favour of accepting risky projects.

A more appropriate choice of discount rate would be the Weighted Average Cost of Capital (WACC). The WACC is a risk adjusted rate of return. It is the minimum rate of return investors demand from an investment for assuming a given level of risk. As the WACC is risk adjusted, it is a suitable choice for the discount rate in NPV calculations.

7.1.2 Weighted Average Cost of Capital

A firm can raise capital in one of two ways: by assuming debt or issuing equity[46]. The cost of this capital is the return on debt and the return on equity respectively. Suppose a firm borrows $\$X$ (debt) and the interest rate is 5%. The cost of this capital is therefore 5%. Alternatively, the firm may issue $\$X$ in shares (equity) and investors demand a 10% return. The cost of this capital is therefore 10%. The weighted average cost of capital (WACC) measures the average cost of capital after taking into account the relative amounts of debt and equity issued by the firm. In general, the WACC is calculated as follows:

$$WACC = \frac{E}{V}R_e + \frac{D}{V}R_d(1 - T_c)$$

Where R_e = Return/Cost of equity

R_d = Return/Cost of debt

E = Market value of the firm's equity

D = Market value of the firm's debt

V = E + D = Market value of the firm's debt and equity

T_c = Corporate tax rate

[46] These can come from a number of sources including: common equity, preferred equity, straight debt, convertible debt, exchangeable debt, warrants, options and so on.

The ratio E/V is the proportion of capital that is raised from equity and the ratio D/V is the proportion of capital that is raised from debt. These ratios form the 'weights' in the weighted average calculation of the firm's cost of capital. By taking a weighted average, the WACC essentially indicates how much interest the firm has to pay on average for every dollar it finances.

More specifically, the WACC is the minimum rate of return that the firm must earn from its assets in order to satisfy its creditors (in the case of debt) or owners (in the case of equity) before they invest elsewhere. Suppose the firm was only able to generate 9% return on its equity. This falls below the rate required by shareholders and they will invest elsewhere.

In addition, the WACC is assumed to be a risk adjusted rate of return. The return on equity, R_e is the rate of return investors demand after accounting for the risk presented by the stock. For risky stocks, investors require greater return as compensation for assuming greater risk. The return on debt, R_d is the rate of return creditors demand after accounting for the risk presented by the debt. As the WACC is constructed from R_e and R_d, it too is adjusted for risk.

Estimating the return on equity however is a complex process with a variety of approaches but no clearly defined superior methodology (Cochrane, 2001). This stems from the fact that the factors that determine a stock's return are not entirely known. Aside from the ubiquitous risk-return trade-off paradigm, determining the exact sources of risk and how to measure and interpret them remains a key challenge in asset pricing. Estimating or rather measuring the return on debt however, is a less controversial issue. After all, a repayment scheme is established between creditors and debtors before any credit is extended so that creditors know exactly what return they will receive. The only real source of risk therefore is that of default. However, no such certainty exists between shareholders and the firm hence the 'risk' of owning equity. Shareholders are not guaranteed a particular rate of return but rather demand a minimum level of return for assuming various dimensions of risk associated with holding the stock.

The difficulty in estimating the WACC therefore stems directly from the difficulty in estimating the return on equity. Nonetheless, as it is a risk adjusted rate of return, it is an appropriate discount rate to be used for example, in NPV calculations as explained in the previous section.

7.1.3 Industry Weighted Average Cost of Capital

An Industry WACC may be used as a benchmark WACC for firms operating in a given industry. Mining stocks for example are more risky than utilities stocks on average (see Section 6.4). Firms operating in the mining industry require higher rates of return because of the risk inherent to their operations. Those willing to invest in mining firms demand higher rates of return hence the industry WACC for mining firms is higher than for utilities firms, for example, which operate in a lower risk environment. Therefore, one would not expect the WACC for firms operating in the mining sector to be same as the WACC for firms operating in the utilities sector. What is required therefore is a separate WACC – one for the mining sector and one for utilities.

In order to perform such a calculation, it is necessary to first divide the universe of stocks into industry groups and then within those groups, calculate the average rate of return on equity. This can then be combined with the return on debt and the WACC may be estimated. However, as the previous Chapter 6 concludes, existing classification schemes such as SIC, GICS and NAICS do not group firms based on similarity in stock return but rather on the basis of economic activity.

Calculating average rates of return within such groups would therefore be meaningless as there is no guarantee that firms within those groups belong to the same risk/return class. They will essentially be a random collection of elements insofar as risk/returns patterns are concerned. The standard deviation of the returns distribution alone would be so large that one industry group would be virtually indistinguishable from another.

This is where the GSC clustering algorithm can make an essential contribution. By grouping stocks based on shared returns patterns, the GSC clusters and the industry groups they represent will contain stocks that are homogeneous within cluster and heterogenous between clusters in terms of returns. The industry average returns obtained from the GSC clusters will be a better representation of the true rate of return required by investors. Statistically there will be better dispersion in average returns between the clusters allowing better comparison between one industry group and another. This will ultimately lead to better industry WACC estimates.

More precise industry WACC estimates will in turn allow for better discount rates and more accurate NPV calculations. More accurate NPV calculations will result in better allocation of funds (or conversely less misallocation) to their most efficient use and ultimately wealth maximisation, which is the fundamental goal of any profit seeking organisation.

7.1.4 Risk homogeneity

What is clear from the literature (see Chapter 2) is that industry costs of capital are imprecise because industry groups fail to achieve **risk homogeneity**. Although not explicitly articulated in the literature, the following conditions must be satisfied in order to achieve risk homogeneity:

1. Risky assets are partitioned into separate risk classes; and
2. The variation in the returns patterns of risky assets within each risk class is minimised.

Partitioning of risky assets into separate risk classes requires that 'risky' stocks (or at least stocks that share a similar risk profile) be grouped into one group while 'non-risky' stocks be grouped into another group. This process is referred to as creating risk homogenous groups or simply risk homogeneity. Why is risk homogeneity important for estimating the cost of capital?

The cost of capital represents the required rate of return on risky investments. Investments which are higher in risk must have a higher rate of return. Industry costs of capital represent the required rate of return on risky investments for a given industry. However if industry groups are formed using some stocks that are risky and some that are not (risk heterogeneity), then any industry cost of capital estimate for that industry will be imprecise. If on the other hand, industry groups are formed where risky stocks are placed into one group and non-risky stocks are placed into another (risk homogeneity), then industry cost of capital estimates for that industry will be precise.

Given Conditions 1 and 2, a procedure such as cluster analysis is appropriate for achieving risk homogeneity since it operates on the basis of maximising between cluster heterogeneity (Condition 1) while maximising within cluster homogeneity (Condition 2). One way to test whether risk homogeneity is achieved (or more specifically the extent to which risk homogeneity is achieved) is to firstly group the stocks into clusters (using the terminology of cluster analysis) or industry groups (using the terminology of industry classification schemes); and then evaluate the ability of these grouped mean returns to explain the cross section of stock returns. If risk homogeneity is achieved, then the cluster/industry group means will explain the cross section of returns well since the cluster/industry groups are comprised of stocks which are highly similar to each other while being dissimilar to stocks from other groups.

Conversely, if risk homogeneity is not achieved, then the cluster/industry groups will essentially represent a random grouping of stocks and the group means will fail to explain the cross section of returns well. Consider the following hypothetical scenario depicted in Figure 7-5.

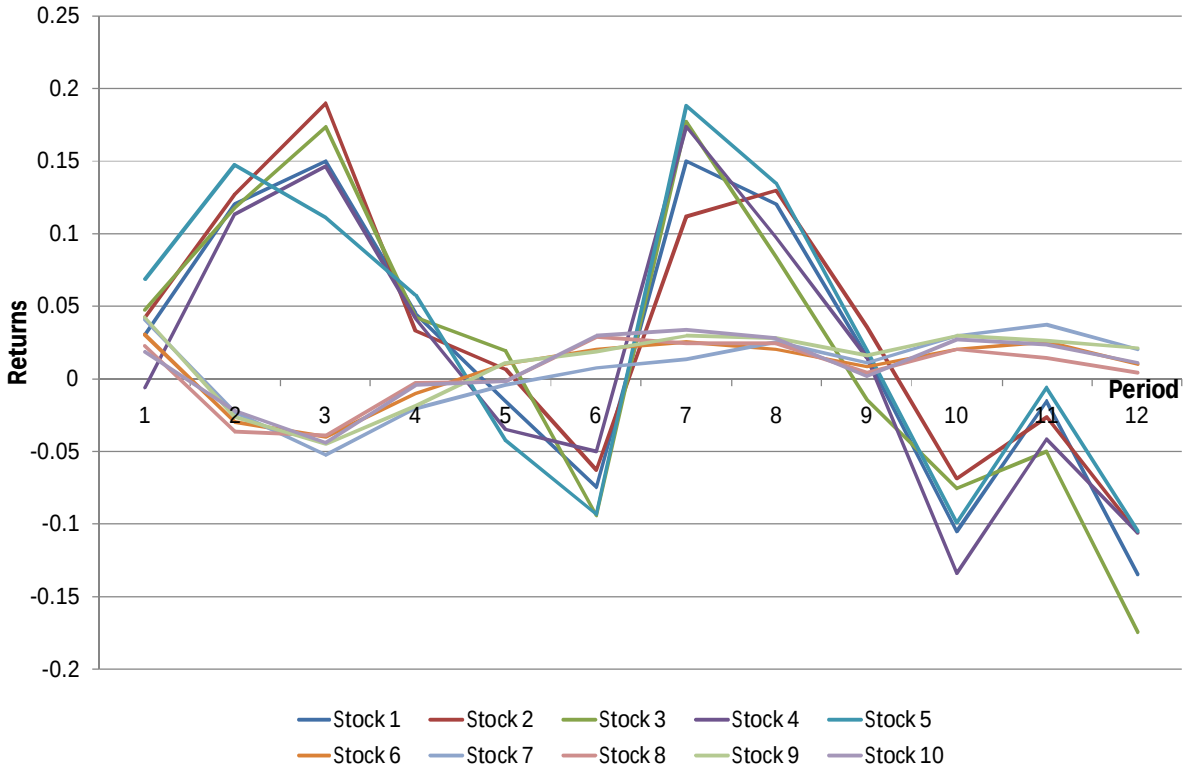


Figure 7-5 Stocks 1 to 5 represent 'high risk' stocks while stocks 6 to 10 represent 'low risk' stocks

Clearly, there are two distinct groups of stocks. Stocks 1 to 5 represent high risk or high volatility stocks characterised by high standard deviation in returns while stocks 6 to 10 represent low risk stocks characterised by low standard deviation in returns. Cluster analysis would group stocks 1 to 5 in one cluster and stocks 6 to 10 in another cluster. The resultant cluster means would be a good representation of their respective stocks. This is depicted in Figure 7-6.

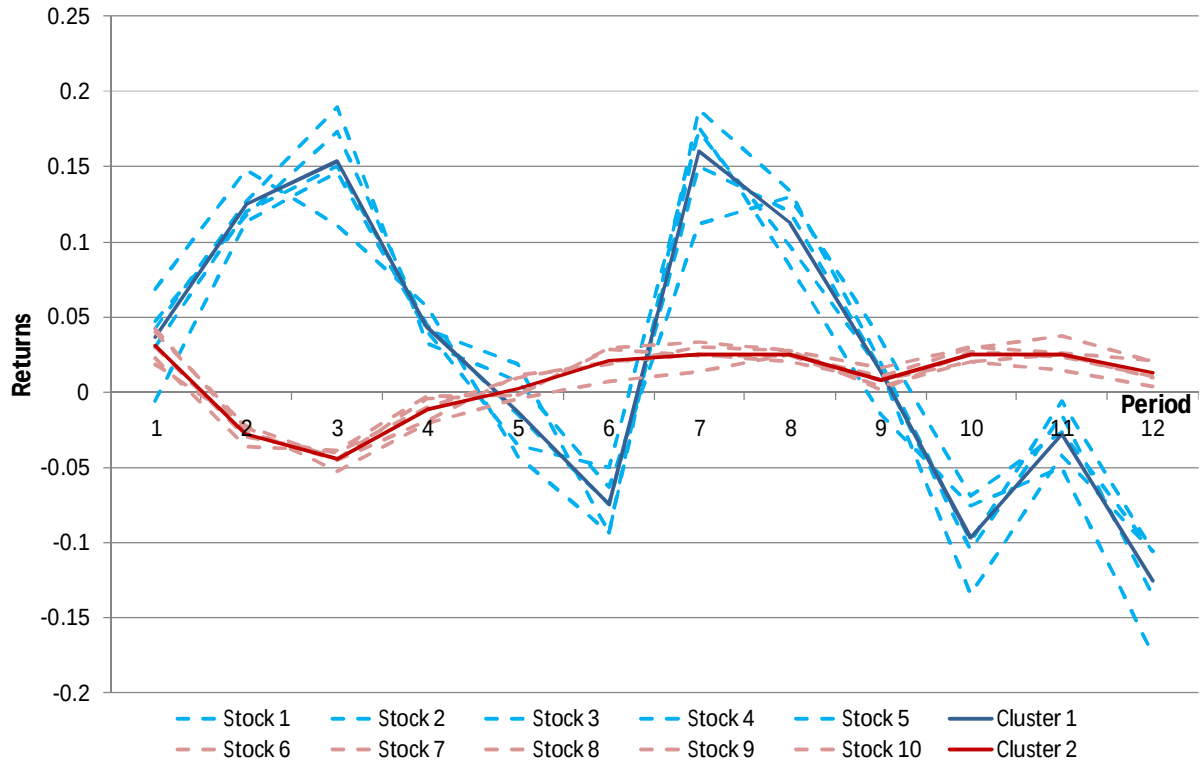


Figure 7-6 The time series of individual stock returns with their cluster mean returns

The time series of individual stock returns are represented by the dotted lines while their cluster mean returns are represented by the solid lines. The cluster mean returns provide an adequate representation of the individual stocks within the cluster; and the cluster mean returns will be able to explain the cross section of stock returns well. In a regression between the individual stock returns and their cluster means (see Equation 7.1, section 7.3.1), the strong correlation between the individual returns and their cluster means will be measured by a high adjusted R^2.

Consider on the other hand if cluster analysis was not used to group the individual stocks; or if the individual stocks were randomly grouped; or if the criteria for grouping stocks was based on characteristics other than similarity in returns. This would be the case if for example, industry classification schemes which are based on similarity in economic activity rather than returns, were used to group the stocks. There is no guarantee therefore, that the resultant industry group means will provide an adequate representation of the individual stocks. This scenario is depicted in Figure 7-7.

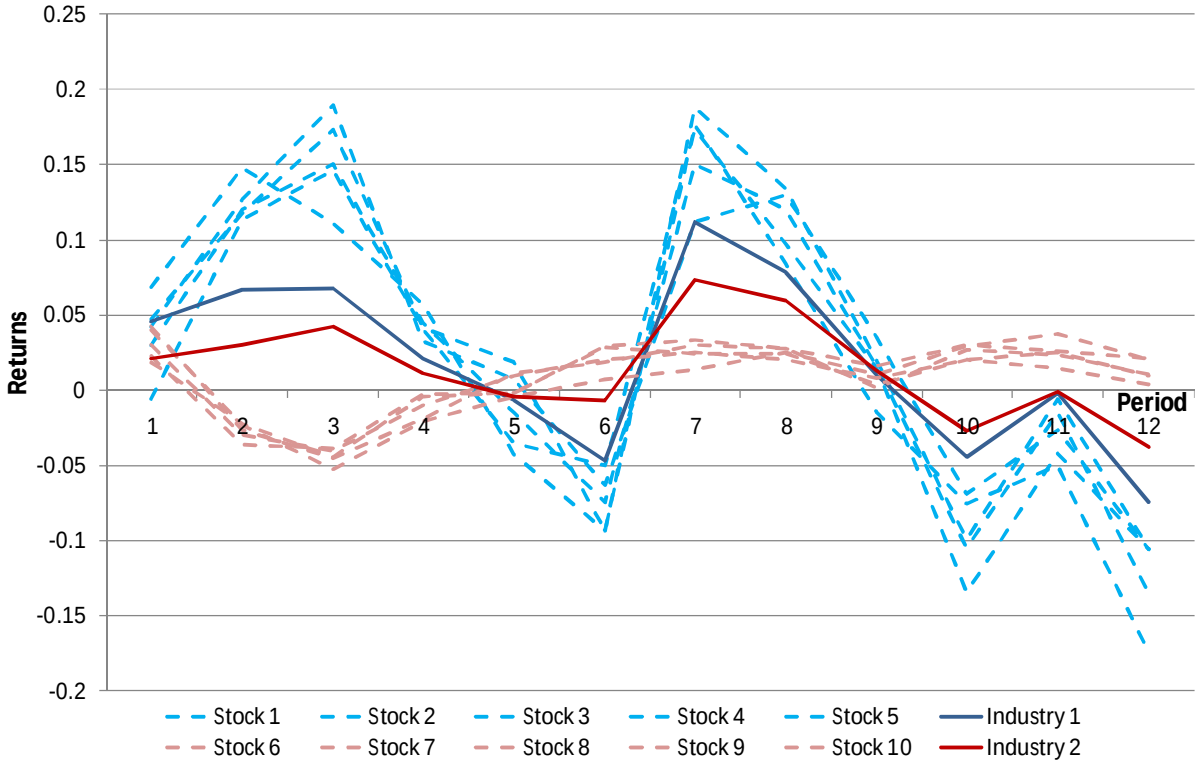


Figure 7-7 Random allocation of stocks into industry groups will result in industry group means that do not necessarily represent individual stock returns well

If stocks were grouped on some basis other than similarity in stock returns, then the industry group means will not represent the individual stocks well and a regression between the individual stock returns and their industry group means will generate low adjusted R^2.

The results in Section 7.4 indicate that the GSC clusters are able to achieve higher adjusted R^2 (approximately 0.4 to 0.6) than similar industry based groups (approximately 0.2 to 0.45). In some cases, the GSC outperforms by up to a factor of 2 to 3.

It is also worth noting that cluster analysis is able to generate greater dispersion in cluster means than random allocation. The blue and red lines are more distinctly separate under cluster analysis (Figure 7-6) than they are under random allocation (Figure 7-7). In cluster analysis terminology, the between cluster heterogeneity is maximised via cluster analysis.

Furthermore, the deviation of the individual stock returns from their cluster means is minimised in cluster analysis than in random allocation. The distance between the dotted lines and their respective solid lines is shorter in Figure 7-6 than in Figure 7-7. Again, using the terminology of cluster analysis, the within cluster homogeneity is maximised via cluster analysis.

7.1.5 Cluster Analysis

Data summarisation techniques such as cluster or factor analysis has been commonly used to summarise the universe of stocks into a more manageable and interpretable set of assets (King, 1966; Brown et al 1997; Elton and Gruber, 1970; Farrell, 1974; Baginski, 1987). In recent work, Ahn, Conrad and Dittman (2005) apply a similar clustering strategy to form what they refer to as *basis assets* (essentially clusters of stocks). The authors cite a potential data snooping bias arising from partitioning based on characteristics and propose the use of clustering as a way to circumvent these issues.

Ahn et al employ a different clustering approach based on the covariance matrix of returns to obtain basis assets. They accordingly use a correlation based distance measure rather than a Euclidean distance measure, which is common in clustering.

A key stage in any clustering algorithm is the calculation of a distance measure. Distance measures are used to determine the degree of 'relatedness' between elements within the cluster variate and also importantly the degree of relatedness between an element and other cluster centroids. One natural condition that the distance measure must satisfy is that similar elements must have a smaller distance than dissimilar elements. In fact, Ormerod and Mounfield (2000) specify three discrete conditions that a Euclidean distance metric must satisfy. These are:

$$d_{ii} = 0$$
$$d_{ij} = d_{ji}$$
$$d_{ij} \leq d_{ik} + d_{kj}$$

The correlation coefficient alone cannot be used as a distance measure since it does not satisfy the conditions outlined by Ormerod and Mounfield (2000). According to the authors, a more appropriate distance measure would be:

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}$$

Where d_{ij} = distance between the i^{th} and j^{th} asset or element

Ahn et al use this distance measure and further elaborate that if two elements are perfectly correlated, this will result in a minimum distance of 0 while if two elements are perfectly negatively

correlated, this will result in a maximum distance of 2. Hierarchical agglomerative clustering was performed using Ward's minimum variance method which seeks to minimise the increase the total sum of squares from the incremental combination (or 'agglomeration') of clusters and/or elements. The total sum of squares was calculated by summing the squared deviations (distance measures) of individual elements within a cluster from its cluster centroid across all clusters.

While the approach used by Ahn, Conrad and Dittman (2005) has its merits, the GSC clustering methodology has potential to produce better clustering outcomes for the following reasons:

Firstly, in computing their distance measure, Ahn et al use the correlation coefficient between one time series of stock returns and another to measure the degree of relatedness between stocks. However doing so condenses the time series of stock returns to one representative cross section loosing much of the dynamic time varying pattern in the returns data resulting in information loss. As such, only one cross section of cluster means will be constructed as clusters are formed. By contrast the GSC clustering methodology forms cluster means at each time period thus incorporating not only the cross sectional but dynamic time varying pattern of the returns data resulting in more effective use of the data.

Secondly, the GSC incorporates a Generalised Least Squares, GLS correction to minimise the effects of heteroskedasticity. As Ahn et al do not apply such a correction, their approach may be vulnerable to periods of high volatility or periods where most stocks perform poorly such as during stock market crashes resulting in potential bias. The GLS correction in the GSC would allocate less weighting to these periods thus minimising the impact of heteroskedasticity.

7.2 Data

As with Chapter 6, data between 1983 and 2006 is used. Although data on CRSP/COMPUSTAT is available from as early as the 1920s, insufficient observations prevent analysis from these earlier periods. Furthermore, as Chapter 5 establishes, the data is divided into 36 month modelling intervals with the first interval being 1983 to 1985 and so forth.

As with Bhojraj, Lee and Oler (2003), industry groups that contain fewer than 5 member elements in any given modelling interval is regarded as 'non functional' and is removed from the analysis.

Additionally, to improve the robustness of the results, the top and bottom 5 percent of the data are 'trimmed', which is consistent with the Least Trimmed Squares method suggested by Knez and Ready (1997). To accomplish this, returns within each month are sorted and observations exceeding the 95th percentile or below the 5th percentile are removed or 'trimmed' using the terminology of Knez and Ready (1997). This is done not only to remove the effects of outliers and to ensure robustness but also to avoid to the problem of 'Heywood cases' in the GSC algorithm. A Heywood case is a scenario in which a cluster contains only one (or very very few) element(s). This can be caused by 'rogue' stocks that exhibit an extreme returns pattern, which is highly dissimilar to any other stock. The GSC will respond to these rogue stocks by placing them in a separate cluster to prevent them from inflating the within cluster SSQ of other clusters. Trimming the top 5 percent significantly reduces the occurrences of such Heywood cases.

7.3 Research Design

7.3.1 In sample estimation

A method similar to Bhojraj, Lee and Oler (2003) who study the ability of industry groups to explain the cross section of stock returns, is used. For each month, industry average returns are computed under the SIC, NAICS, GICS and FF industry classification schemes. Individual returns are then matched with their respective industry average return for estimation via OLS as per the following equation:

$$R_{it} = \alpha + \beta R_{It} + \varepsilon_{it} \qquad (7.1)$$

Where R_{it} = Return of i^{th} stock in period t belonging to I^{th} category

R_{It} = Average return of I^{th} category in period t

To maintain approximate consistency in the number of industry classes across the various industry classification schemes, 2-digit SIC, 3-digit NAICS and 6-digit GICS are used. The SIC scheme has 83 categories at the 2-digit level, the NAICS scheme has 100 categories at the 3-digit level and the GICS scheme has 68 categories at the 6-digit level. The FF industry classification scheme has 48 categories.

Additionally, the number of clusters, K used in the GSC is set to 50. Setting K = 50 is a conservative approach since the NAICS scheme uses twice as many industry classes. Setting K at higher levels will

undoubtedly improve model fit as there are more clusters available for allocating stocks. Consider for example, if there were 1000 stocks in the dataset. At one extreme, setting K = 1000 will allow each stock to be grouped into its own cluster. Estimating Equation 7.1 will yield an R^2 of 1.0 since the dependent variable will be exactly the same as the independent variable. To ensure that the results are not biased toward the GSC in this fashion, K is set to 50, which uses considerably fewer industry classes than the other industry classification schemes, except for the FF scheme which uses 48 classes.

7.3.2 Out of sample testing

To assess the robustness and predictive ability of the various industry classification schemes, cross sectional out of sample testing is performed. For each modelling interval, 10 percent of records are randomly removed from the data set and reserved for validation, henceforth referred to as the validation set (also commonly referred to as a holdout sample). The remaining 90 percent of records, referred to as the estimation set is used to construct industry average returns based on their industry classification. The returns from the validation set are matched to the industry average returns from the estimation set for estimation via OLS as per the following equation:

$$R^*_{it} = \alpha + \beta R_{It} + \varepsilon_{it} \qquad (7.2)$$

Where R^*_{it} = Return of i^{th} stock in the validation set in period t belonging to I^{th} category

R_{It} = Average return of I^{th} category in the estimation set in period t.

Such data splitting procedures are regarded as standard practice for out of sample testing/model validation as explained by Picard and Cook (1984). The authors note however that two important issues must be taken into consideration when performing such splits. The first requires the coefficient estimates, $\hat{\beta}$ to be orthogonal to the dependent variable, Y. To ensure this, the dependent variable must not be taken into consideration when performing the splits. Or in other words, there must be no systematic differences among the Y-values in the estimation and validation sets.

The second requires the error variances to be equal between estimation and validation sets. This may be violated if for example the validation set omitted subgroups of the original data. Such gaps

can occur if these subgroups are small and therefore easily overlooked with reasonable probability when randomly drawing the validation set.

In the current situation, these problems are mitigated since the annual cross section of monthly observations is large. There are approximately 40,000 to 60,000 observations per year (see Figure 5-2). The first condition is met as splits are made on a random basis with no consideration of the dependent values. The second condition is met by virtue of the large number of observations. Therefore data splitting is an appropriate procedure for out of sample testing/model validation.

7.4 Results

7.4.1 In sample estimation

Table 7-1 contains the adjusted R^2 obtained from estimating Equation 7.1.

Period	SIC	NAICS	GICS	FF	GSC (K = 50)	GSC (K = 10)
1983-1985	0.3277	0.3440	0.3383	0.3217	0.4995	0.3971
1986-1988	0.4415	0.4534	0.4515	0.4378	0.5803	0.4958
1989-1991	0.2664	0.2771	0.2758	0.2615	0.4384	0.3487
1992-1994	0.1670	0.1771	0.1776	0.1635	0.3547	0.2472
1995-1997	0.1929	0.2012	0.2058	0.1926	0.3900	0.2931
1998-2000	0.2367	0.2467	0.2627	0.2386	0.4438	0.3498
2001-2003	0.3192	0.3279	0.3458	0.3190	0.5177	0.4261
2004-2006	0.2419	0.2497	0.2601	0.2393	0.4130	0.3156

Table 7-1 In sample adjusted R^2 of common industry classification schemes and their ability to explain the cross section of stock returns via Equation 7.1 as well as the equivalent model under the GSC.

As with Bhojraj et al, GICS exhibits slightly better performance among the common industry classification schemes, particularly towards the latter periods. The model fits here are slightly higher due to the implementation of trimmed least squares (TLS).

However, the GSC based model outperforms other classification schemes by a substantial margin. The GSC (K = 50) achieves adjusted R^2 between 0.4 to 0.6 compared with other industry classification schemes which achieves adjusted R^2 between 0.2 to 0.45. That is, up to 60% of the cross section of stock returns can be explained by cluster mean returns obtained from the GSC.

Reducing the number of clusters to K = 10 still produces impressive results. At K = 10, the GSC still outperforms other classification schemes. Clearly reducing the number of clusters, which is equivalent to having fewer industry classes reduces the model fit but a five-fold reduction in K only worsens model fit by approximately a factor of 1.1 to 1.5[47].

As previously mentioned, K was set to 50 to maintain consistency between the GSC and the number of industry classes. Even at K = 50, the GSC uses less industry classes than the SIC, NAICS and GICS classification schemes. Reducing K to 10 still produces impressive results. At K = 10, the GSC clusters still outperform by a margin of approximately 0.1 in adjusted R^2. Not only is the GSC a superior approach to modelling returns, it can also do this with a more parsimonious classification structure.

[47] For example, in the 1983-1985 modelling interval, the model fit reduces from 0.4995 to 0.3971, which is a factor of 1.26. In fact, the adjusted R^2's reduce on average by a factor of 1.28.

The results indicate that industry clusters formed using the GSC are better able to explain the cross section of returns than equivalent portfolios formed via common industry classification systems. This is expected since the GSC is designed to find common patterns among returns data while industry classification systems such as SIC, NAICS and GICS group firms based on production similarities or demand-side conditions.

In this research, the primary focus is on returns as this variable is of particular interest given its importance in many financial applications. Bhojraj et al study a broad range of accounting and financial variables including valuation multiples, forecasted and realised growth rates, research and development expenditure and various other financial ratios. However, there is no reason to believe the GSC would not perform equally well when applied to such other variables of interest.

For example, if researchers were interested in growth rates, one would simply have to replace the N x T matrix of returns with a matrix of growth rates. The GSC algorithm could then be applied to determine grouping patterns of growth rates. Such groups may even be profiled in a method similar to that outlined in Chapter 6.

This provides researchers, who may previously have relied on industry classifications, with a powerful new way to study the grouping patterns of virtually any variable of interest.

The GSC outperformance becomes more impressive when applied to annual data. Following the methodology of Bhojraj Lee and Oler (2003), who use annual data, a 12-month modelling interval is also tested. The last 10 years of the data set have been used as these periods contain the highest number of observations. The results are summarised in Table 7-2.

Period	SIC	NAICS	GICS	FF	GSC (K = 50)	GSC (K = 10)
1997	0.2226	0.2300	0.2368	0.2226	0.5955	0.4246
1998	0.2955	0.3037	0.3074	0.2937	0.6256	0.4865
1999	0.1568	0.1664	0.1748	0.1570	0.5695	0.3971
2000	0.2381	0.2504	0.2801	0.2440	0.6213	0.4831
2001	0.3350	0.3442	0.3628	0.3354	0.6717	0.5359
2002	0.3012	0.3097	0.3300	0.3003	0.6362	0.4875
2003	0.2052	0.2143	0.2290	0.2044	0.6274	0.4755
2004	0.2448	0.2540	0.2684	0.2423	0.5917	0.4282
2005	0.2421	0.2485	0.2576	0.2383	0.5815	0.4117
2006	0.2296	0.2376	0.2449	0.2283	0.5807	0.4019

Table 7-2 In sample adjusted R^2 of common industry classification schemes and their ability to explain the cross section of stock returns for annual data from 1997 to 2006.

As before with the 36 month modelling interval, the GSC (at 12 month interval) outperforms other industry classification schemes by a substantial factor. The GSC (K = 50) achieves adjusted R^2 between 0.55 to 0.67 compared with other industry classification schemes which achieves adjusted R^2 between 0.15 to 0.36. The GSC achieves an outperformance factor of about 2 to 3.

Reducing the number of clusters to K = 10 still achieves adjusted R^2 between 0.4 and 0.54. At K = 10, the GSC clusters still outperform other industry classification schemes by a considerable margin indicating the GSC is capable not only of achieving superior fit but also parsimony.

7.4.2 Out of sample testing

Table 7-3 contains the adjusted R^2 obtained from estimating Equation 7.2.

Period	SIC	NAICS	GICS	FF	GSC (K = 50)	GSC (K = 10)
1983-1985	0.2475	0.2662	0.2883	0.2754	0.4505	0.3795
1986-1988	0.3944	0.3861	0.4180	0.4042	0.5652	0.5116
1989-1991	0.2299	0.1989	0.2355	0.2352	0.3908	0.3156
1992-1994	0.1255	0.1107	0.1360	0.1448	0.3277	0.2352
1995-1997	0.1718	0.1493	0.1810	0.1772	0.3632	0.2871
1998-2000	0.2062	0.2296	0.2430	0.2272	0.4249	0.3430
2001-2003	0.2917	0.2806	0.3138	0.3040	0.4921	0.4119
2004-2006	0.2124	0.1950	0.2237	0.2270	0.3933	0.3050

Table 7-3 Out of sample adjusted R^2 of common industry classification schemes and their ability to explain the cross section of stock returns via Equation 7.2 as well as the equivalent model under the GSC.

As with the in-sample results, the out of sample results indicate outperformance of the GSC with regard to predicting the out of sample cross section of returns. While there is overall reduction in model fit as would be expected with any form of out of sample testing, the reduction is only minor. Again, as with the out of sample results, the GSC clusters outperform the other classification schemes by a margin of 0.2 to 0.3 or a factor of 2 to 3. Furthermore, the relatively minor reduction in out of sample model fit indicates consistency in the ability of the GSC to predict returns out of sample.

Out of sample testing was also performed for the annual data between 1997 and 2006. The results are summarised in Table 7-4.

Period	SIC	NAICS	GICS	FF	GSC (K = 50)	GSC (K = 10)
1997	0.2012	0.2012	0.2240	0.1901	0.5954	0.4195
1998	0.2755	0.2818	0.2863	0.2851	0.6143	0.4704
1999	0.1415	0.1295	0.1440	0.1462	0.5686	0.3856
2000	0.2131	0.2078	0.2478	0.2209	0.6121	0.4690
2001	0.3110	0.3272	0.3315	0.3038	0.6528	0.5320
2002	0.2942	0.2968	0.2996	0.2923	0.6018	0.4773
2003	0.1735	0.1837	0.1926	0.1751	0.5697	0.4614
2004	0.1969	0.2177	0.2341	0.2178	0.5680	0.4227
2005	0.2151	0.1957	0.2262	0.2136	0.5779	0.3938
2006	0.1965	0.1752	0.2020	0.2016	0.5566	0.3845

Table 7-4 Out of sample adjusted R^2 of common industry classification schemes and their ability to explain the cross section of stock returns for annual data from 1997 to 2006.

As with the 36 month modelling interval, the annual results display a similar pattern of outperformance by the GSC indicating that the GSC clusters are robust to variations in the data and that the same clusters are able to be consistently formed despite these variations.

7.4.3 Better partitioning of risky assets into separate risk classes

As section 7.1.4 explains, the following conditions must be met before any grouped (industry or otherwise) cost of capital may be estimated:

1. Risky assets are partitioned into separate risk classes and
2. The variation in the returns patterns of risky assets within each risk class is minimised.

The GSC is better able to create risk homogenous groupings/classes, i.e. high risk stocks are placed into high risk groups and low risk stocks are placed into low risk groups. One way to measure the riskiness of a group of stocks is the standard deviation of group returns over a given period. If a group of stocks were highly risky, then its average returns would fluctuate greatly over time (i.e. volatility). Such a volatile group of stocks would exhibit a higher standard deviation in average returns. The opposite would be true for a group of stocks that is low in risk.

Industry classification schemes group stocks into coarse industry groups whereas the GSC groups stocks into clusters. The terms ‘industry groups’ and ‘clusters’ may be used interchangeably in this context as they simply refer to different schemes for grouping stocks. If the GSC is superior at

partitioning stocks into risk homogenous groups, then there should be greater dispersion in the standard deviations of cluster/group means via the GSC (see Figure 7-6 vs. Figure 7-7). Recall that the standard deviation commonly measures the volatility of a group of stocks. Since the GSC is better able to partition stocks into separate risk classes, then the standard deviations of the average returns of these separate risk classes will have greater dispersion. Consider the following:

	3 year modelling interval				
Cluster	Month 1	Month 2	...	Month 36	Time series σ
1	$\mu_{1,1}$	$\mu_{1,2}$	...	$\mu_{1,36}$	σ_1
2	$\mu_{2,1}$	$\mu_{2,2}$	...	$\mu_{2,36}$	σ_2
...	...	...	...	...	...
K	$\mu_{K,1}$	$\mu_{K,2}$	...	$\mu_{K,36}$	σ_K

Where $\mu_{K,t}$ = average return of the K[th] cluster in month t.
σ_K = standard deviation of cluster means for the K[th] cluster

For each of the *K* clusters, a time series standard deviation is calculated based on the average cluster returns for that cluster over a 36 month modelling interval. This is repeated for all *T* modelling intervals in the data period (1983 to 2006). If the GSC is better able to partition stocks into separate risk classes, then the standard deviations will have greater dispersion. To measure the dispersion of the standard deviations, the standard deviations of the standard deviations may be used. Furthermore to control for the number of clusters (or industry groups), the standard deviations of the standard deviations is divided by *K* (or number of industry groups). After all, a seemingly high dispersion can be achieved simply by increasing *K*. Thus the appropriate metric becomes:

$$\frac{\sigma(\sigma_K)}{K}$$

This 'standardised'[48] standard deviation of cross sectional standard deviations and standard deviation of cross sectional standard deviations under the GSC is reproduced in Table 10-9. Equivalent cross sectional standard deviations for the SIC, NAICS, GICS and FF industry classification schemes are reproduced in Table 10-10, Table 10-11, Table 10-12 and Table 10-13 respectively.

[48] Meaning standardised by the number of clusters, K or industry groups

The standardised standard deviation of cross sectional standard deviations, $\frac{\sigma(\sigma_K)}{K}$ is reproduced in Table 7-5.

Period	GSC	SIC	NAICS	GICS	FF
1983-1985	0.00032	0.00026	0.00022	0.00026	0.00023
1986-1988	0.00030	0.00021	0.00018	0.00022	0.00021
1989-1991	0.00034	0.00025	0.00022	0.00022	0.00018
1992-1994	0.00031	0.00028	0.00025	0.00023	0.00018
1995-1997	0.00039	0.00022	0.00020	0.00022	0.00022
1998-2000	0.00072	0.00061	0.00040	0.00039	0.00040
2001-2003	0.00086	0.00046	0.00035	0.00050	0.00043
2004-2006	0.00033	0.00024	0.00022	0.00025	0.00028

Table 7-5 The 'standardised' standard deviation of cross sectional standard deviations

As Table 7-5 shows, the GSC produces higher overall dispersion (in some cases, up to twice as much) among the standard deviations indicating better partitioning of risky assets into separate risk classes thus satisfying condition 1 above.

This procedure is repeated for a 12-month modelling interval. The results are reproduced Table 7-6

Year	GSC	SIC	NAICS	GICS	FF
1997	0.00056	0.00022	0.00017	0.00023	0.00027
1998	0.00067	0.00022	0.00023	0.00027	0.00027
1999	0.00066	0.00031	0.00028	0.00030	0.00030
2000	0.00114	0.00076	0.00045	0.00061	0.00070
2001	0.00131	0.00049	0.00038	0.00066	0.00062
2002	0.00088	0.00034	0.00033	0.00041	0.00044
2003	0.00056	0.00033	0.00020	0.00029	0.00037
2004	0.00048	0.00025	0.00024	0.00030	0.00025
2005	0.00042	0.00025	0.00019	0.00024	0.00031
2006	0.00041	0.00025	0.00018	0.00021	0.00036

Table 7-6 The 'standardised' standard deviation of cross sectional standard deviations (12 month interval)

As with the 36 month modelling interval, the GSC at 12 month modelling interval produces higher overall dispersion among the standard deviations indicating better partitioning of risky assets into separate risk classes. Referring to Figure 7-6, it is analogous to having the cluster mean returns time series being well defined and distinct to one another as opposed to Figure 7-7 in which the mean returns series are 'close' and less distinguishable from one another.

7.4.4 Lower returns variation within each risk class

To test whether the GSC is able to minimise returns variation within each cluster/industry group, a simple within cluster variance may be used. The within cluster variance is calculated thus:

$$\sigma^2 = E[(R_{it} - R_{It})^2]$$

Where R_{it} = Return of i^{th} stock in period t belonging to I^{th} category

R_{It} = Average return of I^{th} category in period t

The classification scheme that produces the lowest within cluster variance by definition minimises the returns variations within each risk class. The within cluster variances under the GSC are reproduced in Table 10-14. Equivalent cross sectional standard deviations for the SIC, NAICS, GICS and FF industry classification schemes are reproduced in Table 10-15, Table 10-16, Table 10-17 and Table 10-18 respectively (see Appendix).

The average within cluster variance is reproduced in Table 7-7.

Period	GSC	SIC	NAICS	GICS	FF
1983-1985	0.0036	0.0043	0.0041	0.0043	0.0054
1986-1988	0.0044	0.0051	0.0049	0.0051	0.0064
1989-1991	0.0053	0.0063	0.0063	0.0062	0.0082
1992-1994	0.0056	0.0063	0.0064	0.0061	0.0074
1995-1997	0.0072	0.0081	0.0082	0.0078	0.0100
1998-2000	0.0130	0.0145	0.0149	0.0143	0.0174
2001-2003	0.0108	0.0121	0.0124	0.0124	0.0150
2004-2006	0.0048	0.0050	0.0048	0.0050	0.0055

Table 7-7 Average within cluster variance. The GSC produces lower variances compared to other classification schemes indicating better formation of homogenous risk classes.

It is apparent from Table 7-7 that clusters formed via GSC produce lower within cluster variances than similar industry groups formed via conventional industry classification schemes. Put simply, stocks that share a similar risk profile are placed into the same clusters more often under the GSC based approach than under conventional industry classification. The GSC produces more consistent risk classes by reducing variation within each risk class thereby satisfying condition 2 above.

This procedure is repeated for a 12-month modelling interval. The results are reproduced in Table 7-8.

Year	GSC	SIC	NAICS	GICS	FF
1997	0.0053	0.0088	0.0090	0.0084	0.0091
1998	0.0073	0.0116	0.0118	0.0111	0.0119
1999	0.0086	0.0138	0.0143	0.0138	0.0141
2000	0.0119	0.0189	0.0196	0.0190	0.0197
2001	0.0104	0.0166	0.0167	0.0167	0.0172
2002	0.0073	0.0110	0.0110	0.0109	0.0114
2003	0.0053	0.0080	0.0078	0.0081	0.0083
2004	0.0036	0.0054	0.0054	0.0054	0.0057
2005	0.0033	0.0049	0.0048	0.0050	0.0052
2006	0.0031	0.0048	0.0045	0.0047	0.0052

Table 7-8 Average within cluster variance (12 month interval).

As with the 36 month modelling interval, the GSC at 12 month modelling interval achieves lower within cluster variance across all periods. Referring to Figure 7-6, it is analogous to having the individual time series of returns clustering closely to the time series of cluster mean returns as opposed to Figure 7-7 in which the individual time series of returns are less well represented by their group mean returns.

7.5 Discussion

At the outset of this chapter, issues relating to Net Present Value, NPV calculations and the choice of discount rate were discussed. It was shown how the choice of discount rate could vastly affect the outcome of NPV evaluations. An improperly specified discount rate could lead firms to invest in projects which fail to achieve the required rate of return resulting in financial loss or alternatively not to invest in projects which may otherwise have been profitable resulting lost opportunity.

In order for a discount rate to be accurate, it has to be adjusted for risk. That is, firms operating in risky environments (such as mining) face higher risk and therefore require higher rates of return. The weighted average cost of capital, WACC is a risk adjusted rate of return and commonly used as a discount rate in NPV calculations. Unfortunately previous attempts at calculating industry costs of capital have generated mixed results with the general consensus being that if industry costs of capital are to be properly estimated, then firms operating in an industry must be 'homogenous' in some sense. Unfortunately, there are no clear guidelines in the literature as to how such homogenous industry groups can be formed.

One common albeit less than ideal approach is to group firms using existing industry classification schemes such as SIC, NAICS or GICS. The underlying assumption is that such industry classification schemes are able to group firms into homogenous groups. While such industry classification schemes may be successful at grouping firms into homogenous groups in terms of economic activity, there is no evidence that they are effective at creating homogenous risk/return groups.

The GSC is a clustering algorithm designed specifically to group firms according to similarity in risk/returns. To test whether GSC clusters are superior to industry groups formed via industry classification schemes, the ability to explain the cross section of returns is examined. The results indicate that the GSC clusters are in fact superior to industry groups formed via existing industry classification schemes in their ability to explain the cross section of returns. Furthermore the GSC clusters generate higher dispersion between cluster means while simultaneously achieving lower within cluster variation. This outcome means that cost of equity estimates formed via the GSC are superior to those formed via industry classification schemes.

Better cost of equity estimates in turn lead to better industry cost of capital estimates which in turn lead to more accurate discount rates and NPV calculations, better allocation of funds and ultimately profit maximisation.

7.6 Limitations

7.6.1 Additional measures of homogeneity across firms/stocks

As a concept, homogeneity across firms is multi-faceted. One major facet is the extent to which stock returns are correlated. Given the importance of stock returns in many financial applications, it is only natural for this variable to be the focus of this study. However, as Bhojraj et al explain, “another measure of homogeneity across firms is the extent to which the market ascribes similar valuation multiples to their key accounting measures such as earnings, book value of equity and sales revenue.” Toward this end, they also study the relationship between industry means and variation in firm-level price to book, enterprise value to sales and price to earnings ratio.

They also focus on homogeneity across firms in terms of operational characteristics, for comparison and control purposes. They study the relationship between industry means and return on net operating assets, return on equity, asset turnover ratio, net profit margin and debt to book equity.

Table 7-9 summarises the additional variables studied in Bhojraj et al.

Variable	Description
Price to book value	Market capitalisation divided by total common equity
Enterprise value to sales	Sum of market capitalisation and long term debt divided by net sales
Price to earnings	Market capitalisation divided by net income before extraordinary items
Return on net operating assets	Net operating income after depreciation divided by the sum of property, plant, equipment and current assets, less current liabilities
Return on equity	Net income before extraordinary items divided by total common equity
Asset turnover	Total assets divided by net sales
Profit margin	Net operating income after depreciation divded by net sales
Leverage	Total liabilities divided by total stockholder's equity
Long term analyst growth forecast	
One year ahead realised sales growth	
Scaled research and development expense	R&D research and development expense divided by net sales

Table 7-9 Additional variables of potential interest

Although the GSC was designed to be used with returns data, there is no reason it cannot be used with similar success on other data. It can easily be extended to study any other metric of interest such as those outlined in Table 7-9. These variables represent different facets of firm homogeneity and represent an opportunity for future research.

7.6.2 Equal vs. Value weighted portfolios

Equal weighted portfolios or more specifically, equal weighted portfolio mean returns have been used throughout the asset pricing tests in this chapter. Quite often in the literature, value weighted portfolio mean returns are used to derive a more accurate indication of average returns. This stems from the notion that 'smaller' firms have less of an impact on portfolio performance than 'larger' firms. Their individual returns should therefore be weighted accordingly. The weighting mechanism commonly used is the ratio of the individual firm's market equity (a proxy for 'size') to the overall portfolio market equity, which is derived as the total market equity of all firms constituting the portfolio. Under this approach, portfolio mean returns would be calculated thus:

$$\bar{R}_p = \sum_{i=1}^{n} w_i R_i$$

And $w_i = \frac{ME_i}{\sum_{i=1}^{n} ME_i}$

Where $\bar{R}_p$ is the value weighted return of portfolio *p*

R_i is the return on the i^{th} stock within the portfolio

w_i is the weight of the i^{th} stock within the portfolio

ME_i is the market equity of the i^{th} stock

Computing value weighted returns in this way may lead to more accurate portfolio mean returns thus improving the accuracy of the asset pricing tests in Chapter 7. This is outside the scope of this thesis but should be investigated in future research.

7.7 Conclusion

This study demonstrates how statistical clustering can outperform industry classifications in explaining the cross section of returns. Using returns data between 1983 and 2006 from the merged CRSP/COMPUSTAT database as well as several common industry classification systems, the effect of industry average returns on individual returns at the firm level both in and out of sample is estimated. The statistical clustering procedure known as the Generalised Style Classification, GSC algorithm achieves adjusted R^2 of approximately 0.4 to 0.6 in sample and 0.3 to 0.56 out of sample compared to other classification systems which achieves adjusted R^2 of approximately 0.2 to 0.45 in sample and 0.2 to 0.4 out of sample.

This is a significant finding as it provides researchers with a new method of articulating the relationship between industry and returns. Such a finding has the potential to generate more accurate cost of capital estimates with less error. Current approaches to estimating industry cost of capital rely on industry classification systems such as SIC, NAICS, GICS or FF. The problem with these classification systems is that they group firms based on economic activity not returns. The outcome is inaccurate industry cost of capital estimates that fail to capture the true required rate of return for investors.

This is because current industry classification systems generate industry groups which exhibit a high degree of intra-industry returns heterogeneity (high dispersion in cluster means) and heterogeneity in returns implies heterogeneity in cost of capital thereby making any single estimate meaningless. Statistically such cost of capital estimates generates too much error to be of any practical use (Fama French, 1997).

The GSC circumvents this problem by grouping firms according to returns. Homogeneity in returns implies homogeneity in cost of capital thereby making a single estimate an appropriate and accurate representation of the required rate of return. Additionally this would be estimated with less error since there is less variance in returns within each group.

This study demonstrates the ability of the GSC to form such homogenous returns groups by showing its superior ability in explaining the cross section of returns. Cluster membership from the GSC may then be used as the basis of a new returns based classification system as an alternative to industry classification which is useful as a first step toward better cost of capital estimates.

In addition, the GSC is a highly flexible approach that can be adapted for use with virtually any metric of interest. It has already been successfully used to estimate 'styles' in the mutual funds context and now stock returns at the individual firm level. There is no reason to believe that it could be easily adapted for use with any other metric of interest with equivalent success.

In summary, the GSC is an exciting addition to the finance literature. Its many uses have already been explored in a mutual funds context but this is the first application to individual securities at the stock level. It will help reinvigorate the ongoing discussion on industry effects on stock returns and provide a fresh perspective on the application of cluster analysis to financial data.

Chapter 8

8 Discussion & Conclusion

The overarching objective of this research is to explore how homogeneous groups of stocks based on similarity in returns patterns can be formed. Pursuant to this, there are several sub-objectives that arise. One is to investigate how to the GSC can be used to form a new industry classification scheme based on returns as opposed to economic or business activity. Another is to explore how the GSC based approach can lead to better industry cost of capital estimates. These objectives are in direct response to a number of issues in the finance literature.

In Chapter 2, these issues are explored in detail. The first relates to the current state of industry classification schemes such as SIC, NAICS and GICS. An ongoing issue in the finance literature relates to the formation of homogeneous groupings of stocks. The current standard practice relies on industry classification schemes to make such groupings. But in what way are these stocks meant to be homogeneous? Although common industry classification schemes make these groupings on the basis of economic or business activity, it is not entirely clear what criteria are used to make these groupings. The GSC on the other hand is unambiguous in the way it forms clusters. In this study, stocks are grouped on the basis of returns. Therefore, when two stocks are allocated to a cluster, it is clear that they share the same returns pattern. Furthermore, these clusters have been shown to correlate well to specific industry groups providing an interpretation on the basis of industry. In fact, the GSC can be used with various metrics providing researchers with a powerful way to form homogeneous groups based on any criteria of interest.

The second issue relates to the area of corporate finance. In particular, industry costs of capital. The cost of capital represents the cost to a firm of raising capital or from the shareholder's perspective, the required rate of return from the company's securities. Industry costs of capital therefore represent the cost of capital for firms within a given industry. The cost of capital has two components: the cost of equity and the cost of debt, reflecting the two ways a company can raise capital. Again, the difficulty with estimating industry costs of capital is due to classification scheme used to delineate industries. Previous attempts have failed specifically because industry groups fail to achieve the 'equivalent risk class' assumption of Miller and Modigliani (1966). The GSC circumvents this problem by grouping stocks based on returns. Homogeneity in returns implies homogeneity in cost of capital and since returns are strongly correlated to risk, satisfies the equivalent risk class assumption.

Reliable cost of capital estimates however require reliable cost of equity estimates. By comparison, the cost of debt is easier to calculate. The cost of debt is derived by taking the *risk free rate* (from a security whose duration matches the term structure of corporate debt) and adding a *default premium*. Estimating the cost of equity on the other hand is relatively more challenging as it is not clear what drives returns. The results presented in Chapter 7 indicate that the GSC based approach presents a potential alternative to other approaches to estimating the cost of equity, such as the Fama French approach. The advantage of the GSC based approach is that it does not rely on arbitrary partitioning of the data thereby making it immune to the econometric problems of such practices.

8.1 Homogeneous returns groups

As Chapter 2 explains, capital market research often requires firms to be divided into homogeneous groups. Recent surveys of the finance literature (Kahle and Walking, 1996; Bhojraj, Lee and Oler, 2003), have shown that a large number of studies employ some form of industry classification in their research design. Among these studies, the most common uses for forming industrial groupings include: identifying control firms (over half), describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate.

However the industry classification schemes used to form these homogeneous groups are not clear in the way they make their groupings. There is a fundamental mismatch between the way industrial groups are formed and their eventual application the finance literature. The GSC algorithm however allows researchers to specify the clustering variable thus clusters formed from the GSC will be homogeneous along whatever dimensions specified by the user. This allows for homogeneous groups to be formed in a controlled manner. Furthermore, the number of clusters may be determined *a priori* or objectively via the Gap statistic test.

The GSC solves the problems associated with industry classification schemes precisely because it removes the uncertainty around how the homogeneous groupings are formed and furthermore provides control to the researcher in determining the clustering variable to be used in the formation of such groups.

In Chapter 6, the universe of stocks is clustered via the GSC to form groups that are homogeneous in returns risk. In addition, the Gap statistic test is used to determine the appropriate level of *K* within each modelling interval.

The GSC clusters are then profiled against industry means derived from the SIC scheme at the second digit via squared ranked correlations. In addition, Multidimensional Scaling (MDS) is used to visualise the cluster solution. Visual analysis of the MDS charts is consistent with the profiling suggested by the squared ranked correlation analysis.

Approximately 10 clusters are identified within each modelling interval and these clusters broadly correspond to key sectors within the economy such as the primary, secondary and tertiary sectors.

The success of the clustering procedure depends on a number of factors including whether or not the cluster profiles make intuitive sense, the degree of dispersion between clusters and the theoretical validity of the cluster solution.

The cluster mean returns obtained in this study are well dispersed, the risk associated with each industry cluster are proportional to the mean returns which conforms to the risk-return paradigm in finance theory and these patterns of risk and return are consistent with the operational environment specific to the various industries. This indicates that the clustering solution obtained from the GSC has been successful.

For example, firms operating in the **Primary resources** sector experience relatively higher levels risk (measured by the standard deviation of cluster mean returns) but also generate relatively higher levels return. This sector experiences risk from a number of sources including exploration and extraction which involve large amounts of capital with a high degree of uncertainty. Additionally, fluctuations in commodity prices, exchange rate risk and changes to the political landscape create a complex business environment which contributes to a high degree of risk exposure.

Firms operating in the **Elaborately Transformed Manufactures** (ETMs) sector or **High technology** sector also exhibit a similar pattern of risk and return. Such firms are again exposed to various sources of risk such as high levels of investment into research and development in novel technologies and given that the U.S. is a net exporter of technology based products are also exposed to external sources of risk such as exchange rate risk and changing economic conditions abroad. As a

result, this sector typically experiences relatively higher levels of risk but also generates relatively higher returns.

By contrast, firms operating in the **Basic manufacturing** sector are exposed to relatively fewer sources of risk as many of these products are 'necessity items' or items used in other industrial processes and so exhibit relatively less risk but also generate relatively lower returns. Interestingly, firms in the **Utilities** sector consistently exhibit relatively lower levels of risk but generate moderate returns. The proximity of this sector to the Government and the 'necessity' of their services may have a stabilising effect resulting in lower risk.

Firms operating in the **Retail**, **Apparel and Furnishing** and **Transportation** sectors exhibit relatively moderate levels of risk and generate moderate returns. These sectors do not operate in highly complex business environments (such as the **Primary resources**, **ETMs** or **High technology** sectors) but at the same time are sensitive to changing economic conditions resulting in moderate risk exposure and corresponding levels of return.

In addition to providing taxonomy of the various sectors of an economy in a risk/return context, such risk/return homogeneous groups are also useful for a number of applications. Consider for example, the exercise of identifying control firms for benchmarking purposes (which is the most common application of industry classification schemes). Industry classification schemes are used to identify control firms whose stock performance is compared against that of firms in the same industry who exhibit some form of abnormal phenomenon (Ritter, 1991; Spiess and Affleck-Graves, 1994; Hendricks and Singhal, 2001). In all such studies, SIC codes are used to identify control firms which are then matched against particular firms within the same industry, which exhibit some phenomenon of interest. The intention is to identify, test and estimate the effect of said phenomenon on stock performance. A key stage in this analysis however is the identification of control firms, which ideally, should be identical to the test firms (in terms of characteristics that affect stock performance) with the exception of the phenomenon under study.

The implicit assumption is that firms belonging to the same industry (as identified by the SIC code) also belong to the same returns class. Unfortunately, there is no evidence to suggest this is the case (Clarke, 1989). Nor would there be theoretical grounds to assume this is true given the unclear way in which SIC groupings are formed. Therefore, when control firms are selected in this way, they do not belong to same returns class and so a proper benchmark cannot be established. Industry groups

formed via the GSC by contrast would belong to the same returns class since the clustering is performed specifically with that objective. Therefore identification of control firms via the GSC would be more reliable as they satisfy the equivalent returns class assumption, and selection from these groups will result in more accurate benchmarks.

Chapter 6 not only shows how such equivalent returns clusters are formed via the GSC (and Gap statistic) but also how these clusters may be interpreted in terms of industry. The results indicate that industry groups belonging to the same cluster do in fact exhibit similarity in returns and risk. This finding has the potential to revolutionise studies that involve the identification of control firms for benchmarking purposes as well as various other applications such as describing industrial structure, restricting samples and to categorise acquisitions and divestures as conglomerate or nonconglomerate.

Furthermore the flexibility of the GSC allows the user to specify virtually any variable of interest thereby allowing more reliable benchmarks to be established. Researchers now have a powerful tool to identify control firms that are valid in any context where previously this identification process relied on industry classification schemes, which may only be valid under a set of narrowly defined conditions (such as similarity in economic and business activity).

There are however limitations to this study. Although the GSC is applied to the entire universe of stocks, the profiling is only performed using the second level of detail under the SIC scheme. To derive a comprehensive returns based industry classification scheme, profiling must be performed at deeper levels of detail. This is beyond the scope of this study although it does represent an area for further research. To do so, higher levels of K must be examined, which is prohibitive in practice given the computational requirements of the GSC and in particular the Gap statistic test which requires repeated applications of the GSC to derive the null distribution.

Another limitation is the use of the SIC scheme for cluster profiling. During profiling, SIC industrial groups are used to describe the GSC clusters. The resultant cluster interpretation requires some degree of subjective assessment and this is unavoidable. This study has not considered the use of other industry classification schemes such as NAICS, GICS or FF for profiling and these may result in subtle changes to the cluster interpretations.

For example, under the SIC scheme, the *Petroleum Refining and Related Industries* and *Oil and Gas Extraction* Major Groups may be allocated into one cluster, which can be interpreted as the **Primary Resources** sector however under the GICS scheme the *Oil, Gas & Consumable Fuels* and *Metals & Mining* industries may be allocated to the same cluster, which can be interpreted as the **Primary Fuels and Mining** sector. While there are slight variations, the underlying concept remains the same. Such comparison between different industrial classification schemes is beyond the scope of this study but it does represent an area for further research.

It should be noted however that while common industry classification schemes such as SIC, NAICS, GICS and FF essentially remain static in that once firms are allocated to an industry group, they rarely change membership and that the industry groups and classification structure itself remains mostly constant, the methodology employed in this research allows for periodic dynamic reallocation of firms into different industry groups (i.e. GSC clustering is performed at 36 month intervals) and for those industry groups themselves to change in number[49] and interpretation. This makes the GSC based classification scheme more flexible in adapting to changing industrial structure and accommodating emerging industries in an ever evolving macroeconomic landscape. Note that while existing classification schemes do undergo some periodic revisions, this is usually more to incorporate new and emerging industry groups (such as the technology sector) rather than a dynamic reallocation of existing firms into other industries or a re-evaluation of the existing industrial structures.

8.1.1 Risk adjusted industries

The objectives of this research are to explore ways in which homogenous groups of stocks may be formed based on similarity in risk and return. When one considers the vast universe of stocks, a natural tendency may be to arrange these stocks into groups as there are a number of useful applications for such groupings. The identification of control firms for benchmarking purposes, the pricing of risky assets and the computation of costs of capital are several such applications.

In order to group stocks, a grouping scheme must therefore be established. Such a scheme must possess the ability to group stocks in a way that stocks within a group are similar in terms of some measurable characteristic while being dissimilar to stocks in other groups.

[49] Since the Gap statistic test is used to determine an appropriate level of *K* at each new modelling interval.

Currently a common practice is to use the notion of 'industry' to form such groupings. Researchers commonly use industry classification schemes to create groupings due to their availability and ease of use and interpretation (Kahle and Walking, 1996; Bhojraj, Lee and Oler, 2003; Clarke, 1989). However, as Chapter 2 explains, the problem with such industry classification schemes is that they do not make these groupings on the basis of similarity in risk and return, which is more useful for capital market research but rather on the basis of economic activity. Figure 1-1 depicts how two stocks which belong to the same industry (under the SIC scheme) actually exhibit vastly different patterns of return and risk.

The GSC algorithm utilised throughout this research is a solution this problem as it designed specifically for the purpose of grouping stocks on the basis of risk and return. However, once the GSC clusters are formed we begin to search for useful interpretations of these clusters. The problem here is that it is not immediately clear how these clusters should be interpreted. In Chapter 6, the GSC clusters are once again interpreted on the basis of industry. But since the clusters are not formed on the basis of traditional industry classifications, if an industry interpretation is to be used, what then is the definition of industry?

According to traditional industry classification schemes, industries are groups of firms that share similarity in economic activity – that is, similarity in production process and/or economic output. To some extent the GICS and Fama-French industry classification schemes integrate notions of risk and returns into their grouping algorithms however, the mechanisms by which they do so are unclear. Furthermore, they fail to outperform the GSC clusters in the asset pricing tests in Chapter 7. This notwithstanding, a key question remains as to what exactly constitutes an industry, especially in the context of this research. If the GSC based industry classification scheme is to be accepted, then the definition of industry as it applies to the current research must be re-examined. Therefore, for the purposes of capital market research, we propose that the notion of 'industry' actively incorporate elements of risk and return thus resulting in a **risk adjusted industry**. Such risk adjusted industries are not formed on the basis of economic activity but rather risk and return.

This concept is not dissimilar to the notion of *risk classes* as outlined in Miller and Modigliani (1966) and in more recent research, *basis assets* (Ahn et al., 2005). Whatever the labelling convention, what exists here are groupings of stock that are formed on the basis of risk/return as opposed to economic activity. Furthermore, it is our contention that these groupings are superior for the

purposes of finance research[50]. The superior ability of the GSC clusters in explaining the cross section of stock returns as outlined in Chapter 6 is evidence in support of this contention.

The decision to interpret the GSC clusters in an industry context was made for several reasons. Firstly, evidence does exist in the literature (King, 1966) to suggest that industry effects are present and correlated to a stock's return and secondly, it provides a way to operationalise the GSC clusters giving it some practical meaning. However, this does not preclude the GSC clusters from being interpreted in a different context. The interpretations in Chapter 6 are based on shared correlations between the GSC cluster means and industry portfolios however the industry portfolios may be easily replaced for example with the mimicking portfolios[51] of Fama and French (1992, 1993) giving the GSC clusters a risk pricing based interpretation more consistent with traditional asset pricing theory and further exploration may form the basis for future research.

It is acknowledged that in the context of the current research, traditional concepts of industry may not apply and that the definition of industry must be re-examined. If so, we propose the concept of a **risk adjusted industry** which actively incorporates risk and return. Such risk adjusted industries may be more useful for capital markets research, which often require homogenous groupings of stocks to be made on the basis of risk and return as opposed to economic activity which is what is currently being offered by conventional industry classification schemes.

However, to facilitate interpretation, the clusters have been given an industry interpretation, which is consistent with the work of King (1966). If the link between the GSC clusters and industry groups are too tenuous, then the industry interpretations may at the very least serve as a check for consistency between the clusters and industry groups. For example, a portfolio consisting of mining stock and other firms within the primary resources sector may be compared against the equivalent ***primary resources*** cluster as identified by the GSC.

8.1.2 Risk adjusted industries and asset pricing

As Section 8.1.1 explains, given the approach adopted in Chapter 6, the concept of an industry itself must be re-examined. The traditional notion of industry (in which stocks/firms are grouped on the

[50] As explained in Chapter 8
[51] The SMB and HML portfolios which are meant to proxy for *size* and *value* risk.

basis of economic activity) may be redefined as risk adjusted industries (in which stocks/firms are grouped on the basis of risk and return).

Such risk adjusted industries are superior to traditional industry groupings for the purposes of asset pricing as the results in Chapter 7 indicate. Therefore, rather than using industry benchmarks[52] to gauge individual stock performance, a superior alternative would be to use risk adjusted industry benchmarks. Intuitively, we would expect such risk adjusted industry benchmarks to outperform traditional industry benchmarks since the cluster means more closely follow the individual returns patterns of the stocks that constitute the cluster. This is evident from the lower within group dispersion and higher between group dispersion of the clusters as compared to industry groups.

Furthermore, the risk adjusted benchmarks would outperform traditional industry benchmarks for a number of other reasons. Firstly, firms may drift between industries as their operational activities evolve over time. This would be most prevalent in the case of conglomerates. Secondly, returns are driven by management activity and firm productivity. Portfolios formed on traditional industry groupings do not actively take these into consideration. By contrast, the GSC based approach presumes neither effects but captures both resulting in better grouping of firms and by extension more accurate benchmarks.

8.2 Towards better industry cost of capital estimates

As Section 2.3 explains, the problem with current approaches to industry cost of capital estimates is that industry groups formed via industry classification schemes do not satisfy the equivalent risk class assumption. In order for cost of capital estimates to be valid, firms within the industry group must belong to the same risk class otherwise any cost of capital estimates will be unavoidably imprecise.

While many studies have attempted to reliably estimate industry costs of capital, only a handful (Miller and Modigliani, 1966) have resulted in some degree of success. Furthermore, this finding is only valid under a set of limiting conditions. Miller and Modigliani limited their study to the electric utilities industry which by nature is homogeneous in risk. Even the results of Miller and Modigliani have been questioned by Boness and Frankfurter (1977) who question the equivalent risk class assumption in the “believed-to-be most homogeneous of industries” (p.775), the electric utilities industry studied by Miller and Modigliani.

[52] based on some kind of industry portfolio mean return

The reason why such studies fail to reliably estimate industry cost of capital is because the industry classification scheme used to divide stocks into industry groups fail to create the homogeneous risk classes required by Miller and Modigliani.

In Chapter 7, the GSC algorithm is used to group stocks into clusters. Cluster means are then computed and individual stocks are matched to their respective clusters means. The relationship between the cross section of stock returns and their cluster means is estimated. The procedure is also repeated for common industry classification schemes such as SIC, NAICS, GICS and FF using industry grouped means in a manner similar to that used by Bhojraj, Lee and Oler (2003). To maintain consistency with the number of industrial categories used by the other classification schemes, *K* was set to 50 under the GSC. Setting *K* at this level still uses fewer categories than other classification schemes (with the exception of the FF classification scheme which uses 48 categories). In essence the following equation is estimated:

$$R_{it} = \alpha + \beta R_{It} + \varepsilon_{it}$$

Where R_{it} = Return of i^{th} stock in period t belonging to I^{th} category
R_{It} = Average return of I^{th} category in period *t*

This procedure is performed cross sectionally both in and out of sample. To perform out of sample testing, 10 percent of stocks are randomly removed (referred to as the validation set) and the cluster/industry grouped means are recalculated based on the remaining 90 percent (referred to as the estimation set). The returns in the validation set are matched against the cluster/industry grouped means in the estimation set and the relationship is re-estimated via the following:

$$R_{it}^{*} = \alpha + \beta R_{It} + \varepsilon_{it}$$

Where R_{it}^{*} = Return of i^{th} stock in the validation set in period *t* belonging to I^{th} category
R_{It} = Average return of I^{th} category in the estimation set in period *t*.

Such a validation procedure is considered to be common practice for out of sample testing (Picard and Cook, 1984). As Section 7.3.2 explains, the data used in this study meets the requirements outlined by Picard et al therefore the out of sample testing is valid.

The results from the in-sample estimation indicate that the GSC clusters are able to explain between 40 to 60 percent of a stock's cross sectional variation compared to 20 to 45 percent via other industry classification schemes. Compared to other classification schemes, the GSC achieves substantially superior model fit and utilises fewer industrial categories with the exception of the FF classification scheme.

Reducing *K* to 10 still produces impressive results. At this level, the GSC still explains between 20 to 50 percent of the cross section of stock returns. Therefore, the GSC is able to explain the same amount of cross sectional variation in stock returns as other industry classification schemes but utilising a much more parsimonious classification structure (see Table 7-1 and Table 7-2).

The out of sample performance achieved by the GSC is equally impressive. At *K* = 50, the GSC explains between 33 to 57 percent of the out of sample cross section variation in stock returns compared to 11 to 42 percent achieved by other industry classification schemes. Once again, reducing *K* to 10 still produces overall superior model fit compared to other industry classification schemes with only a modest reduction in model fit relative to *K* = 50 (see Table 7-3 and Table 7-4).

But of what significance are these results to the wider finance community? These impressive results are evidence that the GSC can create groups of stocks that are homogeneous in risk and return where other classification schemes cannot. The creation of such homogeneous groups has a number of useful applications such as estimating industry costs of capital. Previous attempts have failed because the standard errors associated with any point estimate are 'unavoidably imprecise' making them too large to be of any practical use.

Miller and Modigliani (1966) conclude that in order to derive accurate industry costs of capital estimates, firms within the industry must belong to the same 'risk class' (homogeneous in risk). The intuition is simple. From the perspective of the firm, the cost of capital represents the cost to an individual firm of raising capital. From the perspective of the investor however, the cost of capital represents the required rate of return on a risky investment such as stocks. Industry costs of capital therefore represent the required rate of return on stocks in a given industry. If an industry group is comprised of stocks which share a similar risk profile (risk homogeneity), then the resultant industry cost of capital estimate will be a precise representation of the required rate of return for securities within that industry. If on the other hand, an industry group is comprised of stocks which do not

share a similar risk profile (risk heterogeneity), for example if some stocks are ‘high risk’ while others are ‘low risk’, then the resultant industry cost of capital estimate will be an imprecise representation of the required rate of return, that is it will be estimated with too much error. The problem with industry groups formed via existing industry classification schemes is that they exhibit characteristics of the latter.

Currently, industry cost of capital studies rely on industry classification schemes such as SIC, NAICS, GICS or FF to divide firms into coarse industrial groupings. However, these industrial groupings do not contain stocks that are homogeneous in risk thereby violating the equivalent risk class assumption of Miller and Modigliani. As such, industry costs of capital estimates derived from these classification schemes will be imprecise. By contrast, the GSC algorithm does create groupings of stocks that are homogeneous in risk.

While not specifically articulated in the literature, the following conditions must be met in order for groups to be homogeneous in risk (see Section 7.1.4).

1. Risky assets are partitioned into separate risk classes; and
2. The variation in the returns patterns of risky assets within each risk class is minimised.

When risky assets are partitioned into separate risk classes, the cluster/industry grouped means will be well dispersed. Within each cluster/industry grouped mean series, the time series standard deviation may be calculated. Recall that the time series standard deviation measures the riskiness of the cluster/industry group. Groups that contain predominantly risky stocks will exhibit high variability in cluster mean returns (high volatility) while the opposite is true for groups that contain predominantly non-risky stocks. If the cluster/industry grouped means are well dispersed, then the standard deviations of the time series standard deviations will be high. Furthermore, dividing by *K* or the number of industrial groupings will standardise this measure. Thus the appropriate metric to measure the degree of dispersion is $\frac{\sigma(\sigma_K)}{K}$, where σ_K represents the time series standard deviation for the K^{th} cluster.

According to this measure, the GSC achieves higher dispersion compared to other classification schemes (up to twice as much) indicating the GSC is better able to partition risky assets into separate risk classes thereby satisfying condition 1 above (see Table 7-5 and Table 7-6).

To measure the degree of returns variation within each risk class, a simple within cluster/industry group variance is used. This within cluster variance is measured by:

$$\sigma^2 = E[(R_{it} - R_{It})^2]$$

Where R_{it} = Return of i^{th} stock in period t belonging to I^{th} category

R_{It} = Average return of I^{th} category in period t

Among the various classification schemes, the GSC produces the lowest average within cluster variance (see Table 7-7 and Table 7-8) indicating that it is able to minimise returns variation within each cluster thereby satisfying condition 2 above.

The evidence presented in Chapter 7 is irrefutable and indicates that the GSC is superior to other forms of industry classification at forming groups of stocks that are homogeneous in risk and return thereby satisfying the equivalent risk class assumption of Miller and Modigliani. As such, industry costs of capital estimated by the GSC will be calculated with lower error leading to greater precision and ultimately more accuracy thus providing a solution to the unavoidably imprecise estimates experienced by Fama and French (1997) and non-homogeneous industry groups found by so many others in the literature (Litzenberger and Rao, 1972; Boness and Frankfurter, 1977; Chan et al, 2007; Rapach, Strauss, Tu and Zhou, 2010; Asness, Porter and Stevens, 2000).

By grouping stocks based on returns, the GSC provides a method for researchers to create the homogeneous risk classes so desperately needed in the literature. Furthermore, the innovative statistical features of the GSC means that such homogeneous risk classes are able to be generated with greater precision thereby ensuring more precise industry cost of capital estimates. Such improvements to the industry cost of capital lead to better discount rates used in NPV calculations which will in turn result in better allocation of financial resources and ultimately profit maximisation.

8.3 Comparison to other cost of equity studies

The cost of capital consists of two components: the cost of equity and the cost of debt. Compared to the cost of equity, estimating the cost of debt is relatively simple. Common approaches involve the use of an appropriate *risk free rate* and the incorporation of *default risk*. Estimating the cost of equity however is relatively more complex. This comes from the fact that it is not immediately clear what factors drive returns. While it is widely acknowledged that returns are driven by risk, the specific sources of risk (risk factors) and their interpretation remains an area of wide debate in the literature. This study has examined how the GSC clusters may be used to explain the cross sectional variation of returns in the context of improving industry cost of capital estimates. Implicit to this however, is the estimation of the cost of equity. Therefore, it may be relevant to compare the findings in this study to other cost of equity studies.

In the 1960s the Capital Asset Pricing Model (CAPM) became the dominant paradigm in asset pricing. The ideological implication of the CAPM is that a stock's return is driven entirely by the market. Later, King (1966) successfully identified industry factors as having an effect on a stock's return representing a paradigm shift in asset pricing theory. In recent times however, the Fama French (1992, 1993) model has gained popularity. Under this approach, stocks are thought to be exposed to

three risk factors: the market (as identified by the CAPM), size and value risk. The Fama French model represents an extension of the CAPM in that in addition to the market factor, two more risk factors, size and value are identified as having an impact on stock returns. Ideologically, it represents a shift away from the industry effects paradigm of King.

8.3.1 Industry effects

In 1966, King applied factor analysis to the covariance matrix of monthly changes in closing prices and found that stocks from the same industry tended to have greater co-movement than stocks from different industries. The important ideological paradigm advanced by King was that industry effects do have an impact on the pricing of securities.

However, only 20 percent of a stock's return could be explained by King's procedure. The GSC based approach by contrast can explain up to 60 percent, which represents a threefold increase in explanatory power. The GSC method used here represents a return to the industry-effect paradigm suggested by King. However, unlike King who only studied 63 stocks from 1927 to 1960, the data in this study has a vastly broader range. As Figure 3-1 indicates, there are approximately 40,000 to 60,000 observations per year spanning over the years 1983 to 2006 across all available industries within the macroeconomy.

There are of course several key differences in the methodologies used to arrive at this conclusion. For one, King used **factor analysis** to search for co-movements in the correlation matrix of changes in closing prices. By contrast, the study here uses **cluster analysis** which places individual stocks into groups based on minimising within cluster variation. The fact that these clusters can be profiled in terms of industry indicates that the clusters themselves are in fact a representation of industry groups. Once again, the limiting factor in King's analysis was that he relied on a classification scheme whose method for making industry groupings is unclear[53]. Therefore, even though King had a theoretically valid hypothesis, the industry classification scheme used may not have provided the homogeneous returns groupings required to make the correlation patterns clear. By contrast, the GSC method used here generates its own classification scheme, which is profiled using SIC groups. So while it still relies on the SIC scheme, the extent to which it does so is only for the purposes of cluster profiling not to make the initial groupings themselves.

[53] King used *the Directory of Companies Filing Annual Reports with the Securities and Exchange Commission*, published by the Securities and Exchange Commission in 1961 to make industry groupings from the individual stocks

This notwithstanding, the fact that such different methods were used but the same conclusion is reached (i.e. industry effects exist) is a validation of King's hypothesis.

Although King originated the concept of industry effects, the results here represent a revolution in this area. There can be no doubt that industry membership has a strong effect on the returns generating process. This result is significant not only because it signals a return to the pioneering work of King, but is a substantial leap forward, extending and investigating the vision of the original work.

In addition, King cites several areas where such findings may be beneficial. These include: methods of portfolio selection, the design of index numbers and estimating the cost of capital. This chapter has already explored how the GSC clusters can be beneficial in at least one of these areas: namely the cost of capital. As King explains: "Miller and Modigliani (1958, 1961) assume that firms can be divided into equivalent returns classes, implying that firms in the same class have identical costs of capital"; and "If one can abstract from the unique parts of the variance of individual stock series, the idea of homogeneous risk classes is plausible. Cluster analysis techniques ... should be helpful in showing one where to look for an empirical confirmation of the Modigliani and Miller hypothesis" (p.166 – 167).

The results in Chapter 7 not only show how such homogeneous risk classes can be formed but in doing so provide the exact 'empirical confirmation' King calls for. The GSC clusters formed are not only superior at explaining the cross section of returns both in and out of sample (in some cases outperforming other classification schemes by a factor of 2 – 3), but have better dispersion between cluster means and lower variation within clusters. This is evidence that the GSC clusters have successfully formed the homogeneous risk classes missing in the literature. It is clear that cost of capital estimates derived from the GSC clusters will be superior to those currently available under existing industry classification schemes.

8.3.2 Fama French model

In Fama French (1992, 1993), stock returns are estimated as a direct function of market risk and two additional risk factors: size and value. In a head to head comparison of adjusted R^2, the results in Chapter 7 do not hold up well against studies such as Fama French who achieve R^2 of 0.90 and above

(see Table 2-3). However, this would not be an equivalent comparison as the Fama French approach is based on constructing arbitrary portfolios and estimating the relationship between average value weighted portfolio returns and portfolios constructed to mimic different dimensions of risk (namely *size* and *value* risk) whereas the GSC based approach is applied to all stocks at the individual firm level. The Fama French approach analyses stock returns at the portfolio level whereas the GSC approach used here analyses stock returns at the individual firm level. Under Fama French, it is therefore not possible to obtain cost of equity estimates for individual firms however under the GSC approach, it is.

Therein lays the potential weakness of the Fama French model. Individual stock returns at the firm level are 'noisy'. In such cases, data partitioning may be used to reduce this noise. However, good econometric practice requires as much of the data to be used as this allows the researcher to better explore the myriad of relationships between variables and provides better dispersion within individual data items resulting in improved power for statistical tests. However, when excessive partitioning of the data occurs, several econometric issues arise. These include: selection and truncation bias, loss in power of statistical tests and data snooping bias (see Section 2.4.4). As the Fama French model employs such partitioning of the data, which is done entirely arbitrarily and acknowledged by the authors (Fama French, 1993, p.9), it too may be prone to some of these econometric issues.

It is not the intention of this study to discredit the Fama French approach in any way. The objective however is to demonstrate how an alternative approach based on clustering that does not rely on any partitioning of the data and one that is truly data driven may be used to as an alternative approach to estimating the cost of equity.

The GSC based approach deals with noise in a more constructive way. It is a data driven technique which does not impose any arbitrary structure or limitations on the data. The GSC forms clusters on the sole basis of minimising total sum of squares. This is achieved when within cluster variation is minimised and between cluster variation is maximised (resulting in risk homogeneity). Furthermore, the number of clusters, *K* is estimated via the Gap statistic test. The GSC-Gap combination ensures that the clustering process is entirely data driven and free from the imposition of any extraneous theoretical constructs or prior beliefs as to how to the clusters should be formed. Put simply, the GSC-Gap combination allows the data to speak for itself.

In addition, the integrated GLS correction, which is a well established and proven statistical technique, mitigates the effect of extreme observations. More importantly, as there is no partitioning of the data, the GSC approach suffers from none of the econometric problems described above.

Furthermore The GSC clusters used here can be interpreted and profiled against different industrial sectors (see Chapter 6) making them useful for estimating industry cost of capital. The Fama French portfolios on the other hand have no interpretation in the context of industry making industry cost of capital estimates impossible. In fact, Fama French adapt their approach to estimate industry cost of capital in Fama French (1997) but ultimately find such estimates to be "unavoidably imprecise" and therefore of little practical use. The motivation behind this was to develop industry groups that share common risk characteristics (thereby leading to improved cost of capital estimates). Despite this imprecision, the Fama French industry classification scheme has been used in other research (Gebhardt, Lee and Swaminathan; 2001, Lee, Myers and Swaminathan; 1999) even though their efficacy has never been tested.

The flexibility of the GSC algorithm allows it to be applied to a broad range of variables. The GSC algorithm simply requires an N x T array of numerical data. In the current study returns are used, reflecting its importance in many financial studies. However, there is no reason to believe the GSC algorithm cannot be applied to other variables with equal success (see Section 7.6.1). The GSC has already been used to estimate mutual fund styles in the mutual funds context (Brown and Goetzmann, 1997). Bhojraj, Lee and Oler (2003) study a broad range of financial and accounting variables and the ability of industry groups to explain the cross section of these variables. The GSC may be applied to such variables in a manner similar to that used in the current study to derive GSC industry clusters which may in turn be used to model any given variable of interest whereas the Fama French approach cannot. The size and value sorts used are designed specifically to be applied on returns data.

At the very least, the GSC represents an alternative approach to the established Fama French approach. Those wishing to adopt a truly data driven model that imposes less arbitrary restrictions and partitioning of the data should consider the GSC in their asset pricing models.

8.3.3 Competing paradigms

The implications of this study for financial researchers, practitioners and the wider finance community in general are deep and profound. It raises questions regarding the very underlying nature of the returns generating process. This process as envisaged by King is driven by the firm's operational environment which is captured by its industry. As King explains:

"The very fact that we have averages of industrials, rails and utilities, not to mention indexes founded on narrower classifications of securities implies that many investors think of stocks as falling into groups based on similarity of performance" (p. 139).

Furthermore, practitioners and academics alike have long included industry effects in their discussion of movements in stock returns. Financial information vendors such as Valueline, Morningstar/Ibbotson and Bloomberg regularly publish a range of performance indicators based on industry. Observations such as those made by King underscore a long held sentiment of industry as having an effect on patterns of stock return. Specifically King saw this process as being driven by information about market conditions, some of which have an overall effect and others which affect only a specific portion of the market:

"*There is some news of a monetary nature, for example, which is bound to have a market-wide impact on security price. The magnitude of impact need not, however be the same for all stocks. Other information may affect only certain subgroups of stocks, such as the news of a change in defense policy and its influence on the aircraft industry*" (p.140)

By contrast, Fama French eschew such idiosyncratic industry effects in favour of more generic measures of risk such as size and value, which are captured by their mimicking portfolios, *SMB* and *HML*. The implication of the Fama French model is that the risk faced by any given security comes from only three sources: exposure to the market, size and value. Proponents of the Fama French model cite the high R^2 achieved by the model as evidence of superiority. Ideologically, the philosophy implied by the Fama French approach does reflect certain truths that are difficult to refute. Few would argue that 'small' firms, i.e. those with little market capitalisation do in general pose a greater risk to investors than 'big' firms, which have the ability to operate more profitably under varying economic conditions.

In terms of 'value' risk, undervalued firms (*value* stocks), i.e. those which the market ascribes a lower valuation than its accounting value does pose a greater risk to investors than overvalued firms (*growth* stocks) where the market is optimistic about future performance and hence ascribes a higher valuation than its accounting value.

So which paradigm (industry effects or Fama French factors) is correct? There is no clear answer. It depends on how one views the concept of risk.

Our contention is that any generic measure of risk oversimplifies the various facets of risk. Risk as a concept cannot be captured by a discrete number of generic factors but must be interpreted on a case by case basis. As such, stocks must be grouped on a case by case basis and these 'clusters' of stock must be examined in the context of their operational environment. Some stocks face risk from high capital requirements. Others face risk from external sources such as exchange rate risk or changes in economic conditions both domestic and abroad. Other stocks may be sensitive to changes in the political or legislative infrastructure. And some others may simply be unaffected by these effects.

Perhaps a combination of the two approaches represents the way forward. Generic measures of risk such as the Fama French factors may be responsible for the "market-wide impact on security price(s)" while more specific measures of risk such as industry effects may be responsible for changes to "certain subgroups of stocks" only.

Methodology wise, it may be possible to integrate the matched GSC cluster mean returns along with the Fama French factors in a regression of the cross section of stock returns. However, it would be difficult to divorce the Fama French approach from its arbitrary partitioning of data, which has become a central characteristic of this method. The Fama French approach is applied to carefully selected and heavily partitioned portfolios of stocks whereas the GSC based approach is applied to all stocks at the individual firm level. In a regression of individual stock returns against the Fama French factors and matched GSC cluster mean returns, the Fama French factors lose much of their explanatory power in comparison to the GSC cluster mean returns.

8.4 Concluding remarks

At the outset of this research, it was established that an ongoing problem in the literature is the formation of homogeneous groups of stocks. This research contributes to the literature by providing a way of forming such homogeneous groups.

Methodology wise, it employs an innovative, data driven and objective technology that has already been successfully applied to the mutual funds context – the Generalised Style Classification algorithm. When applied to individual stock returns at the firm level, it provides the homogeneous groupings required but lacking in the literature.

The creation of such homogeneous groups in particular as they relate to industrial sectors has been a "long-standing problem in the financial research" (Bhojraj et al, 2003, p. 746). The current best approach is to use industry classification schemes such as SIC, NAICS and GICS and more recently the FF industry classification scheme developed by the academic community to make such homogeneous groupings. However, these classification schemes fail because of the unclear way groups are formed. In general, these classification schemes are formed on the basis of similarity in economic or business activity however many financial applications require groups to be formed on the basis of similarity in returns (Ritter, 1991; Spiess and Affleck-Graves, 1994; Hendricks and Singhal, 2001). Therefore, there is a fundamental mismatch in the way industry groups are formed and their eventual application in the financial literature. The GSC represents the solution to that problem. As the GSC was designed specifically to construct homogeneous returns groups, it corrects the mismatch between the way groups are formed and their eventual use the literature. The GSC forms groups that are homogeneous in returns which is exactly what the literature requires. Furthermore, as the GSC clusters are shown to correlate well to certain industrial sectors, the GSC clusters themselves can be interpreted on the basis of industry. By correcting this mismatch, studies which require such homogeneous groupings for example in the identification of control firms for benchmarking purposes will improve. Furthermore, the flexibility of the GSC allows it to be applied to various metrics of interest allowing homogeneous groups to be created on virtually any basis as required by the research objective. This makes the GSC an exciting addition to the literature because it represents a potential solution to the "long-standing problem" described by Bhojraj et al.

The second theme of this research is in the area of corporate finance. Estimating industry costs of capital have generally been unsuccessful (Litzenberger and Rao, 1972; Fama French, 1997; Boness

and Frankfurter, 1977; Chan et al, 2007; Rapach, Strauss, Tu and Zhou, 2010; Asness, Porter and Stevens, 2000) because current industry classification schemes used to delineate industries fail to achieve the 'equivalent risk class' assumption of Miller and Modigliani resulting in 'unavoidably imprecise' estimates. The reason why current industry classification schemes fail to achieve this equivalent risk class assumption is because the classification schemes themselves are not constructed on the basis of similarity in risk/return but rather economic and business activity.

Once again, the GSC solves this ongoing problem in the literature. Because of the innovative way in which it creates returns based groupings, the GSC clusters are comprised of stocks which are homogeneous in risk and return. The evidence presented in Chapter 7 indicate how the GSC is able to achieve better between cluster dispersion with lower within cluster dispersion in group means thus satisfying the conditions required for risk homogeneity. As the GSC clusters are better able to achieve risk homogeneity, industry cost of capital estimates generated from the GSC would be more precise, i.e. estimated with lower error. Furthermore, the GSC is able to achieve adjusted R^2 from estimating the cross section of returns both in and out of sample which are up to 3 times greater than current industry classification schemes. This remarkable result is further indication that the GSC is superior to other industry classification schemes at explaining returns, which will again lead to better industry cost of capital estimates. Such a revolution is highly beneficial to the finance literature, which has long suffered from imprecise industry cost of capital estimates. This has a number of highly useful applications both in theory and practice such as improving NPV calculations which is essential for project valuation.

Ideologically, this research cuts to the very core of the founding principles that underpin asset pricing theory. Returns are correlated to risk. That much is undisputed. In today's literature, the dominant paradigm belongs to that of the Fama French model which suggests that risk can be measured along three separate and distinct dimensions: the market, size and value. By contrast, King's study suggests that risk is captured by a firm's industry and the nature of that risk varies from one industry to another. This research extends the work of King and finds evidence in support of industry effects, which is a departure from the paradigm suggested by Fama French. The decision as to which paradigm is superior ultimately depends on one's perception of risk. Proponents of the Fama French paradigm argue in favour of generic dimensions of risk that can be applied (albeit with varying degrees) to all securities. Our contention is that the various dimensions of risk are too heterogeneous in nature to be captured by a handful of generic factors. Each firm is different and while some firms share common sources of risk, these sources need to be evaluated on a case by

case basis and interpreted accordingly. An integrated approach is possible in which the Fama French factors capture the generic component of risk and GSC cluster mean returns capture the 'idiosyncratic' industry specific component of risk. However, in order for such an approach to be developed, differences in the way the data is used must be reconciled. The Fama French method relies on data partitioning to create the necessary portfolios for the regressions whereas the GSC based approach is applied to all stock at the individual firm level.

As previously explained, the aim of this research is not to discredit the Fama French model[54] but to provide an alternative perspective in the area of asset pricing that is based on objectivity and driven by data. We hope that it will reinvigorate discussion about differing paradigms of asset pricing and how the community views concepts of risk and their relation to the returns generating process.

[54] Although any arbitrary partitioning of data needs to be carefully evaluated as these can lead to a number of econometric issues.

9 Bibliography

Ahn, Conrad, & Dittman. (2007). Basis Assets. *Review of Financial Studies* , 22 (12), 5133-5174.

Asness, C., Porter, R., & Stevens, R. (2000). Predicting Stock Returns Using Industry-Relative Firm Characteristics. *Working paper* .

Baginski, S. (1987). Intraindustry Information Transfers Associated with Management Forecasts of Earnings. *Journal of Accounting Research* , 25 (2), 196-216.

Barber, B., & Lyon, J. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics* , 341-372.

Bekaert, & Urias. (1996). Diversification, Integration and Emerging Market Closed-End Funds. *Journal of Finance* , 51 (3).

Berk, J. (2000). Sorting out sorts. *Journal of Finance* , 55, 407-427.

Bhojraj, Lee, & Oler. (2003). What's My Line: A comparison of Industry Classification Schemes for Capital Market Research. *Journal of Accounting Research* , 41 (5).

Boness, J., & Frankfurter, G. (1977). Evidence of Non-Homogeneity of Capital Costs within "Risk Classes". *The Journal of Finance* , 23 (3), 775-787.

Brown, S., & Goetzmann, W. (1997). Mutual Fund Styles. *Journal of Financial Economics* , 43, 373-399.

Chan, L., Lakonishok, J., & Swaminathan, B. (1997). Industry Classifications and Return Comovement. *Financial Analysts Journal* , 63 (6), 56-81.

Chen, N. (1991). Financial Investment Opportunities and the Macroeconomy. *Journal of Finance* , 529-554.

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

Chou, Chou, & Wang. (2004). On the Cross-section of Expected Stock Returns: Fama-French Ten Years Later. *Finance Letters* , 2 (1).

Clarke, C. (1940). *The Conditions of Economic Progress.* Macmillan and co.

Clarke, R. (1989). SICs as Delineators of Economic Markets. *Journal of Business* , 62 (1), 17-31.

Cochrane, J. (2000). *Asset Pricing, 1st Edition.* Princeton University Press.

Copeland, T., Weston, F., & Shastri, K. (2004). *Financial Theory and Corporate Policy (4th Edition).* Addison Wesley.

Daniel, K., & Titman, S. (1999). Sorting out "Sorting out sorts". *Working paper, Northwestern University* .

de Santis, G. (1993). Volatility Bounds for Stochastic Discount Factors: Tests and Implications from International Financial Markets. *Working paper, University of Chicago* .

Dimson, E., & Mussavian, M. (1999). Three centuries of asset pricing. *Journal of Banking and Finance* , 1745-1769.

Elton, E., & Gruber, J. (1970). Homogeneous Groups and the Testing of Economic Hypotheses. *Journal of Financial and Quantitative Analysis* , 4 (5), 581-602.

Elton, E., & Gruber, M. (1971). Homogenous groups and the testing of economic hypotheses. *Journal of Business* , 581-602.

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

Fama, E., & French, K. (1997). Industry Costs of Equity. *Journal of Financial Economics* , 43, 153-93.

Fama, E., & French, K. (1992). The cross-section of Expected Stock Returns. *Journal of Finance* , 47, 427-465.

Farrell, J. (1974). Analyzing Covariation of Returns to Determine Homogeneous Stock Groupings. *Journal of Business* , 47 (2), 186-207.

Fertuck, L. (1975). A Test of Industry Indices Based on SIC codes. *Journal of Financial and Quantitative Analysis* , 37, 837-848.

Gebhart, Lee, & Swaminathan. (2000). Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance* , 51 (1).

Gupta, M., & Huefner, R. (1972). A Cluster Analysis Study of Financial Ratios and Industry Characteristics. *Journal of Accounting Research* , 77-95.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate Data Analysis (7th Edition).* Prentice Hall.

Hendricks, & Singhal. (2001). The Long-Run Stock Price Performance of Firms with Effective TQM Programs. *Management Science* , 47 (3).

Jensen, S. (1971). A Cluster Analysis Study of Financial Performance of Selected Business Firms. *The Accounting Review* , 36-56.

Kahle, K., & Walking, R. (1996). The Impact of Industry Classifications on Financial Research. *Journal of Financial and Quantitative Analysis* , 31 (3), 309-335.

King, B. F. (Jan. 1966). Market and Industry Factors in Stock Price Behaviour. *Journal of Business* , 39, 139-187.

Knez, P., & Ready, M. (1997). On the Robustness of Size and Book-to-Market in Cross-Sectional Regressions. *Journal of Finance* , 52 (4), 1355-1382.

Krishnan, J., & Press, E. (2003). The North American Industry Classification System and Its Implications for Accounting Research. *Contemporary Accounting Research* , 685-717.

Litzenberger, R., & Rao, C. U. (1972). Portfolio Theory and Industry Cost of Capital Estimates. *Journal of Financial and Quantitative Analysis* , 1443-1462.

Livingston, M. (1977). Industry Movements of Common Stocks. *Journal of Finance* , 861-874.

Meyers, S. (1973). A Re-examination of Market and Industry Factors in Stock Price Behavior. *Journal of Finance* , 695-705.

Miller, M., & Modigliani, F. (1966). Some Estimates of the Cost of Capital to the Electric Utility Industry, 1954 - 1957. *American Economic Review* , 333-391.

Milligan, G. W., & Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* , 50 (2), 159-79.

Omerod, P., & Mounfield, C. (2000). Localised Structures in the Temporal Evolution of Asset Prices. *New Approaches to Financial Economics, Santa Fe Conference* .

Pari, R., & Chen, S. (1984). An Empirical Test of the Arbitrage Pricing Theory. *Journal of Financial Research* , 121-130.

Picard, & Cook. (1984). Cross-Validation of Regression Models. *Journal of the American Statistical Association* .

Rapach, D., Strauss, J., Tu, J., & Zhou, G. (2010). How Predictable are Industry Portfolio Returns? *Working Paper* .

Ritter. (1991). The Long-run performance of Initial Public-Offerings. *The Journal of Finance* , 46 (1).

Schwert, G. W. (2002). *Anomalies and Market Efficiency, in Handbook of the Economics of Finance, G. Constantinides.* North Holland, Amsterdam, 939-974: R.M. Stulz and Milton Harris.

Sharpe, W. (1982). Factors in New York Stock Exchange Security Returns. *Journal of Portfolio Management* , 5-19.

Spiess, & Affleck-Graves. (1995). Underperformance in long-run stock returns following seasoned equity offerings. *Journal of Financial Economics* , 38, 243-267.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the Number of Clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society (Series B)* , 411-423.

Weiner, C. (2005). The impact of Industry Classification Schemes on Financial Research. *SFB 649 Discussion Paper* .

Appendix

10 Appendix

10.1 Data Items

Data item	CRSP/ COMPUSTAT code	Description
Standardised Industry Classification	DNUM	This represents the SIC code available on CRSP/COMPUSTAT. It is used to calculate industry average returns and to identify a company's primary operation. To implement condition 1 (see below), companies with SIC codes between 6000 and 7000 were omitted as these represent financial firms.
Global Industry Classification	GIC	This represents the current GIC provided by Standard and Poor's and Morgan and Stanley capital international available on CRSP/COMPUSTAT. It is used to calculate industry average return and to identify a company's primary operation.
North American Industry Classification System	NAICS	This represents the NAICS (1997 version) developed by the Office of Management and Budget available on CRSP/COMPUSTAT. It is used to calculate industry average return and to identify a company's primary operation.
Closing Price	PRCC	This records the closing price for companies on national stock exchanges and bid prices for over the counter issues. The closing price is used in the calculation of Market Equity.
Common Shares Outstanding	CSHOQ	Common shares outstanding are presented on a quarterly basis. The number of common shares outstanding is used in the calculation of market equity.
Share Code	SHRCD	The Share Code is a two-digit code describing the type of shares traded. The first digit describes the type of security traded while the second digit provides additional information. To implement condition 6 (see below), only stocks with share codes 10 or 11 were used as this represents ordinary common shares.
Exchange Code	EXCHD	The Exchange Code indicates the exchange on which a security is listed. To implement condition 2, only stocks with an exchange code of 1, 2, 3, 31, 32 or 33 were included as this represents stocks listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and the NASDAQ stock market respectively.
Return	RET	A return is the change in the total value of an investment in a common stock over some period of time per dollar of initial investment. RET(t) is the return for a sale on day t. It is based on a purchase on the most recent time previous to t when the security had a valid price. Usually, this time is $t-1$. Returns are calculated as follows:

		For time *t* (a holding period), let: *t*' = time of last available price < *t* *r*(*t*) = return on purchase at *t*', sale at *t* *p*(*t*) = last sale price or closing bid/ask average at time *t* *d*(*t*) = cash adjustment for *t* *f*(*t*) = price adjustment factor for *t* *p*(*t*') = last sale price or closing bid/ask average at time of last available price < *t*, then: $$r(t) = \frac{p(t)f(t) + d(t)}{p(t')} - 1$$ *t*' is usually one period before *t*, but *t*' can be up to ten periods before t if there are no valid prices in the interval.
Liquidation value	Data10	This is the estimated amount of money an asset could be sold for when it is allowed insufficient time to sell on the open market
Investment Tax Credit (ITC)	Data51	A tax credit, which is an amount deducted directly from federal income tax otherwise payable, calculated as a fixed percentage of eligible expenditures on scientific research and experimental development.
Redemption value	Data56	The price at which the issuing company may choose to repurchase a security before its maturity date.
Book value (BV)	Data60	Book value or carrying value is the value of an asset or according to its balance sheet account balance. For assets, the value is based on the original cost of the asset less any depreciation, amortization or impairment costs made against the asset. A company's book value is its total assets minus intangible assets and liabilities.
Balance Sheet Deferred Taxes (BSDT)	Data74	The value of assets that is used to reduce the amount of tax that a company will have to pay in a later tax period. Such assets are kept on the balance sheet. For example, a deferred tax asset of $100,000 from the previous year could be applied to before-tax income of $250,000 this year, resulting in taxable income of $150,000 ($250,000 - $100,000).
Par value (PV)	Data130	Par value is a nominal value of a security which is determined by an issuer company at a minimum price.

Table 10-1

10.2 Cluster interpretation

10.2.1 1986 to 1988

	Cluster								
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Agricultural Production Crops	102	30	189	168	65	52	294	36	7
Agricultural Services	18	56	14	90	35	64	66	64	30
Metal Mining	36	12	9	28	15	4	30	40	1
Coal Mining	72	4	64	30	42	18	48	28	3
Oil And Gas Extraction	120	78	27	44	70	14	75	504	5
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	96	261	60	105	120	44	54	26	29
Building Construction General Contractors And Operative Builders	108	224	80	126	189	72	39	22	34
Heavy Construction Other Than Building Construction Contractors	104	70	91	126	72	56	69	42	6
Construction Special Trade Contractors	35	189	70	20	54	80	3	18	80
Food And Kindred Products	225	144	258	329	477	448	75	74	55
Tobacco Products	6	8	30	49	64	81	42	36	32
Textile Mill Products	198	93	432	296	203	195	192	52	16
Apparel And Other Finished Products Made From Fabrics And Similar Materials	222	260	376	140	441	371	84	58	53
Lumber And Wood Products, Except Furniture	162	88	189	288	297	120	90	34	36
Furniture And Fixtures	160	81	486	301	328	300	172	56	27
Paper And Allied Products	280	69	175	513	384	148	318	88	52
Printing, Publishing, And Allied Industries	230	176	330	440	486	399	132	96	51
Chemicals And Allied Products	205	204	504	246	344	378	141	72	46
Petroleum Refining And Related Industries	161	24	75	176	108	80	81	513	9
Rubber And Miscellaneous Plastics Products	252	99	441	320	238	190	184	40	25
Leather And Leather Products	270	100	120	200	133	64	66	64	13
Stone, Clay, Glass, And Concrete Products	234	78	287	432	320	124	260	66	24

Primary Metal Industries	392	128	145	459	224	78	300	106	11
Fabricated Metal Products, Except Machinery And Transportation Equipment	513	330	180	308	448	144	120	104	23
Industrial And Commercial Machinery And Computer Equipment	330	141	399	448	495	172	285	90	22
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	300	212	400	350	450	105	275	102	18
Transportation Equipment	288	200	364	432	459	260	153	80	41
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	364	285	424	270	513	188	147	82	37
Miscellaneous Manufacturing Industries	432	205	266	204	396	128	78	60	42
Railroad Transportation	174	51	231	441	248	170	136	92	39
Motor Freight Transportation And Warehousing	196	204	333	115	184	60	72	30	10
Water Transportation	168	414	51	68	176	115	42	330	15
Transportation By Air	150	150	154	216	243	57	148	50	28
Transportation Services	301	54	204	312	333	145	132	54	43
Communications	155	140	264	294	414	440	108	48	56
Electric, Gas, And Sanitary Services	27	80	44	78	112	264	18	8	513
Wholesale Trade-durable Goods	357	147	276	416	468	164	270	78	38
Wholesale Trade-non-durable Goods	280	215	156	217	351	63	156	98	19
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	70	57	161	232	120	396	128	46	33
General Merchandise Stores	80	448	186	100	182	459	36	28	54
Food Stores	228	100	312	165	342	336	87	68	48
Automotive Dealers And Gasoline Service Stations	95	18	152	114	119	243	57	152	17
Apparel And Accessory Stores	110	228	168	64	200	378	42	44	45
Home Furniture, Furnishings, And Equipment Stores	84	252	256	130	168	360	60	24	50
Eating And Drinking Places	204	160	405	140	288	322	105	84	21
Miscellaneous Retail	448	156	252	322	423	245	93	70	35
Hotels, Rooming Houses, Camps, And Other Lodging Places	99	20	10	72	35	30	30	12	8
Personal Services	156	30	210	342	224	150	164	162	14
Business Services	220	144	408	477	294	180	336	100	20
Automotive Repair, Services, And Parking	50	66	126	32	88	84	30	3	88

Miscellaneous Repair Services	18	48	12	12	28	20	2	171	52
Motion Pictures	60	296	252	48	84	175	48	20	47
Amusement And Recreation Services	376	96	150	224	315	51	225	86	12
Health Services	288	315	90	144	270	88	34	141	49
Educational Services	9	7	5	6	8	2	4	3	4
Social Services	40	63	8	10	12	6	28	6	2
Engineering, Accounting, Research, Management, And Related Services	424	324	280	150	405	112	114	62	31

Table 10-2 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 1986 to 1988 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.2 1989 to 1991

	Cluster								
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Agricultural Production Crops	224	120	60	210	76	243	120	29	72
Agricultural Services	35	4	3	64	20	36	24	342	1
Metal Mining	5	6	72	1	14	4	2	9	40
Coal Mining	12	3	63	10	36	40	96	240	504
Oil And Gas Extraction	56	42	20	2	42	16	15	10	513
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	144	64	30	6	30	12	9	2	210
Building Construction General Contractors And Operative Builders	238	369	304	192	96	125	42	18	3
Heavy Construction Other Than Building Construction Contractors	252	224	144	68	315	90	57	14	48
Construction Special Trade Contractors	99	72	36	4	10	20	21	55	217
Food And Kindred Products	336	312	93	240	132	258	513	112	16
Tobacco Products	12	15	49	72	1	60	99	110	8
Textile Mill Products	66	238	306	148	477	384	245	46	20
Apparel And Other Finished Products Made From Fabrics And Similar Materials	195	376	258	132	495	231	93	16	66
Lumber And Wood Products, Except Furniture	344	174	513	81	120	210	195	74	38
Furniture And Fixtures	72	259	240	140	513	432	235	78	39
Paper And Allied Products	42	135	128	246	448	513	357	42	94
Printing, Publishing, And Allied Industries	250	477	208	138	282	400	385	108	42
Chemicals And Allied Products	378	450	105	408	216	160	176	98	25
Petroleum Refining And Related Industries	234	119	57	76	105	32	150	48	440
Rubber And Miscellaneous Plastics Products	477	264	287	104	204	120	78	88	41
Leather And Leather Products	189	279	162	72	312	105	66	48	21
Stone, Clay, Glass, And Concrete Products	360	210	495	69	162	110	112	70	32
Primary Metal Industries	128	216	448	215	322	459	108	18	98
Fabricated Metal Products, Except Machinery And Transportation Equipment	264	414	378	156	185	376	126	20	70
Industrial And Commercial Machinery And Computer Equipment	255	495	196	392	432	318	114	38	26

Electronic And Other Electrical Equipment And Components, Except Computer Equipment	392	504	176	456	252	220	129	82	28
Transportation Equipment	114	343	250	324	416	504	192	56	23
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	329	468	144	440	240	170	123	68	18
Miscellaneous Manufacturing Industries	513	408	104	252	156	130	90	92	27
Railroad Transportation	16	78	68	112	224	342	100	21	150
Motor Freight Transportation And Warehousing	80	114	273	352	60	315	135	86	10
Water Transportation	108	22	45	70	176	252	147	30	216
Transportation By Air	84	182	125	304	51	261	204	50	15
Transportation Services	51	72	296	238	115	468	222	100	40
Communications	75	115	180	124	224	288	486	106	53
Electric, Gas, And Sanitary Services	60	36	26	126	75	161	448	513	48
Wholesale Trade-durable Goods	230	486	184	343	408	246	99	30	14
Wholesale Trade-non-durable Goods	168	210	99	468	190	368	371	104	34
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	39	125	88	280	450	360	240	20	8
General Merchandise Stores	57	256	112	282	369	273	225	72	13
Food Stores	90	304	235	168	301	495	300	32	90
Automotive Dealers And Gasoline Service Stations	315	168	44	60	175	78	27	7	18
Apparel And Accessory Stores	246	432	210	112	360	280	69	26	88
Home Furniture, Furnishings, And Equipment Stores	217	378	174	116	352	100	48	34	11
Eating And Drinking Places	392	387	270	225	93	259	184	47	104
Miscellaneous Retail	87	240	136	477	336	392	260	90	12
Hotels, Rooming Houses, Camps, And Other Lodging Places	42	90	15	56	18	8	6	1	4
Personal Services	468	176	84	150	36	95	224	102	51
Business Services	440	513	265	300	343	168	105	54	29
Automotive Repair, Services, And Parking	30	84	432	168	44	248	90	66	6
Miscellaneous Repair Services	54	21	126	80	40	45	56	6	38
Motion Pictures	264	405	161	44	108	75	30	8	74
Amusement And Recreation Services	207	128	30	105	84	48	75	31	92
Health Services	333	192	48	120	26	85	203	220	43

Educational Services	60	120	48	78	261	33	119	26	17
Social Services	4	20	8	48	48	63	12	4	154
Engineering, Accounting, Research, Management, And Related Services	320	396	90	154	64	84	39	24	7

Table 10-3 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 1989 to 1991 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.3 1992 to 1994

	Cluster								
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Agricultural Production Crops	287	280	75	168	140	369	252	29	92
Agricultural Services	35	24	70	8	64	36	522	138	11
Metal Mining	4	35	77	112	144	16	66	24	81
Coal Mining	72	1	18	48	10	4	6	56	231
Oil And Gas Extraction	10	66	27	376	48	70	9	275	522
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	63	144	30	42	12	56	21	4	34
Building Construction General Contractors And Operative Builders	162	396	84	108	245	208	135	13	14
Heavy Construction Other Than Building Construction Contractors	136	198	70	33	84	77	72	9	24
Construction Special Trade Contractors	66	171	3	20	20	56	136	1	16
Food And Kindred Products	400	287	108	306	160	378	160	41	82
Tobacco Products	126	4	104	3	4	30	1	364	160
Textile Mill Products	210	39	468	130	238	176	100	39	74
Apparel And Other Finished Products Made From Fabrics And Similar Materials	408	315	99	230	240	405	136	2	8
Lumber And Wood Products, Except Furniture	30	90	360	273	351	56	174	20	88
Furniture And Fixtures	87	186	140	260	376	245	513	32	42
Paper And Allied Products	95	140	504	36	200	96	92	3	4
Printing, Publishing, And Allied Industries	188	322	324	285	416	396	90	35	78
Chemicals And Allied Products	406	522	215	148	222	464	105	15	38
Petroleum Refining And Related Industries	1	8	208	40	54	8	18	399	513
Rubber And Miscellaneous Plastics Products	170	252	294	96	414	256	132	12	68
Leather And Leather Products	80	175	54	522	180	184	140	88	25
Stone, Clay, Glass, And Concrete Products	63	168	200	144	450	248	329	37	60
Primary Metal Industries	112	135	522	324	384	189	78	38	76
Fabricated Metal Products, Except Machinery And Transportation Equipment	215	252	184	87	456	459	371	42	102

Industrial And Commercial Machinery And Computer Equipment	312	408	200	105	371	504	240	31	62
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	324	416	188	147	385	513	255	19	56
Transportation Equipment	69	120	440	160	504	273	294	36	48
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	294	513	148	102	225	424	392	17	36
Miscellaneous Manufacturing Industries	160	360	63	392	108	264	264	27	98
Railroad Transportation	336	210	288	88	155	216	60	25	30
Local And Suburban Transit And Interurban Highway Passenger Transportation	16	3	42	4	56	15	12	504	315
Motor Freight Transportation And Warehousing	240	115	456	180	522	280	117	43	100
Water Transportation	6	48	369	196	126	75	48	96	440
Transportation By Air	96	264	155	385	174	261	63	47	104
Transportation Services	90	105	15	40	176	102	414	11	2
Communications	124	273	424	250	246	423	57	54	106
Electric, Gas, And Sanitary Services	3	16	138	92	100	126	39	522	448
Wholesale Trade-durable Goods	240	486	176	123	258	392	364	23	58
Wholesale Trade-non-durable Goods	216	224	102	215	396	200	148	30	80
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	368	203	68	57	85	324	246	33	46
General Merchandise Stores	495	208	12	78	55	147	48	7	6
Food Stores	185	259	126	192	432	495	270	106	47
Automotive Dealers And Gasoline Service Stations	72	126	8	112	30	104	15	80	14
Apparel And Accessory Stores	513	266	48	105	60	304	144	26	40
Home Furniture, Furnishings, And Equipment Stores	228	344	45	150	76	210	486	16	26
Eating And Drinking Places	315	400	195	93	252	486	144	18	32
Miscellaneous Retail	371	392	90	152	165	450	300	21	12
Hotels, Rooming Houses, Camps, And Other Lodging Places	208	27	162	68	65	180	20	45	378
Personal Services	234	238	96	72	392	432	160	100	35
Business Services	392	440	114	176	306	468	275	8	20
Automotive Repair, Services, And Parking	110	126	408	60	161	333	42	102	42
Miscellaneous Repair Services	105	136	40	54	15	171	32	68	9

Motion Pictures	125	432	88	318	266	224	66	14	72
Amusement And Recreation Services	210	423	87	132	156	272	155	5	52
Health Services	308	504	80	60	120	368	228	6	10
Educational Services	112	40	60	108	162	96	30	10	144
Social Services	104	10	171	9	60	63	3	245	172
Engineering, Accounting, Research, Management, And Related Services	165	477	128	75	252	344	258	22	44

Table 10-4 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 1991 to 1994 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.4 1995 to 1997

	Cluster									
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Agricultural Production Crops	50	360	168	104	48	8	3	315	224	21
Agriculture production livestock and animal specialties	2	5	12	7	1	48	540	3	540	4
Agricultural Services	4	63	78	30	96	220	288	30	104	36
Metal Mining	144	162	540	12	77	66	2	15	2	65
Coal Mining	15	56	8	330	135	28	208	84	39	50
Oil And Gas Extraction	168	168	48	243	102	56	18	600	3	100
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	24	14	6	40	27	1	10	6	128	4
Building Construction General Contractors And Operative Builders	22	260	264	203	200	207	156	16	174	125
Heavy Construction Other Than Building Construction Contractors	430	248	78	112	210	156	36	225	19	468
Construction Special Trade Contractors	168	230	18	24	36	105	186	105	14	270
Food And Kindred Products	470	387	336	352	259	120	138	46	55	205
Tobacco Products	56	27	49	100	4	2	15	36	180	30
Textile Mill Products	45	224	140	477	510	288	86	288	43	130
Apparel And Other Finished Products Made From Fabrics And Similar Materials	216	296	192	170	405	570	82	132	4	329
Lumber And Wood Products, Except Furniture	80	530	99	100	203	468	348	38	50	296
Furniture And Fixtures	16	150	590	105	196	342	87	432	38	88
Paper And Allied Products	120	304	123	48	130	600	413	300	21	522
Printing, Publishing, And Allied Industries	84	480	420	448	468	258	114	220	57	240
Chemicals And Allied Products	513	570	260	287	344	96	40	114	35	234
Petroleum Refining And Related Industries	45	12	288	20	10	290	432	186	51	147
Rubber And Miscellaneous Plastics Products	264	336	138	531	490	329	68	285	31	168
Leather And Leather Products	204	64	34	405	272	460	265	81	12	231
Stone, Clay, Glass, And Concrete Products	69	405	160	400	287	530	100	336	41	190
Primary Metal Industries	210	287	228	288	376	590	110	147	17	504
Fabricated Metal Products, Except Machinery And Transportation	480	312	136	322	432	100	28	90	15	210

Equipment										
Industrial And Commercial Machinery And Computer Equipment	350	495	159	215	456	306	104	212	13	590
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	343	459	188	210	480	294	90	138	5	600
Transportation Equipment	111	240	220	440	504	580	108	406	26	230
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	504	590	204	312	322	240	94	123	33	456
Miscellaneous Manufacturing Industries	530	441	128	216	304	125	70	96	46	280
Railroad Transportation	85	297	60	132	147	136	54	132	40	440
Local And Suburban Transit And Interurban Highway Passenger Transportation	198	80	10	8	12	91	78	55	1	310
Motor Freight Transportation And Warehousing	112	15	20	351	360	248	228	120	7	64
Water Transportation	36	243	360	182	56	108	14	590	9	115
Transportation By Air	270	108	6	35	36	84	60	3	162	144
Pipelines, Except Natural Gas	6	42	133	126	104	12	12	340	25	25
Transportation Services	198	203	215	304	315	400	56	108	37	108
Communications	322	580	168	540	424	234	147	74	49	170
Electric, Gas, And Sanitary Services	12	105	70	144	171	15	264	52	590	4
Wholesale Trade-durable Goods	315	540	232	456	522	185	60	153	22	324
Wholesale Trade-non-durable Goods	287	414	150	240	550	112	84	260	47	392
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	234	112	90	45	112	410	68	16	8	33
General Merchandise Stores	171	36	33	84	144	500	76	24	16	70
Food Stores	351	192	112	150	168	340	72	40	36	95
Automotive Dealers And Gasoline Service Stations	320	171	18	77	60	128	16	68	44	75
Apparel And Accessory Stores	342	120	75	217	264	550	42	112	6	120
Home Furniture, Furnishings, And Equipment Stores	217	85	116	280	390	405	32	54	18	168
Eating And Drinking Places	580	308	93	294	352	486	46	172	29	215
Miscellaneous Retail	550	450	76	306	432	392	153	168	27	265
Hotels, Rooming Houses, Camps, And Other Lodging Places	400	224	42	192	100	297	132	78	42	252
Personal Services	174	470	81	210	110	243	224	8	48	256
Business Services	416	600	147	235	531	264	80	188	28	385

Automotive Repair, Services, And Parking	225	77	132	120	320	32	4	110	159	6
Miscellaneous Repair Services	98	104	24	171	310	45	48	18	11	102
Motion Pictures	590	270	92	138	161	168	66	30	30	145
Amusement And Recreation Services	600	245	54	185	360	210	50	100	10	408
Health Services	459	560	111	464	294	210	74	160	20	300
Educational Services	65	340	210	128	243	40	20	174	24	36
Social Services	350	198	60	119	42	152	33	14	23	32
Engineering, Accounting, Research, Management, And Related Services	486	520	156	378	400	160	72	117	34	270

Table 10-5 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 1995 to 1997 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.5 1998 to 2000

	Cluster							
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Agricultural Production Crops	217	75	392	3	10	15	108	20
Agriculture production livestock and animal specialties	8	3	84	56	48	40	1	160
Agricultural Services	133	88	65	60	4	14	48	28
Metal Mining	156	144	217	18	22	48	60	185
Coal Mining	1	9	30	315	84	24	8	272
Oil And Gas Extraction	7	24	51	250	182	126	14	496
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	80	54	8	15	133	45	40	336
Building Construction General Contractors And Operative Builders	14	208	72	236	155	259	126	70
Heavy Construction Other Than Building Construction Contractors	246	344	180	46	64	108	231	285
Construction Special Trade Contractors	322	392	200	12	68	24	156	84
Food And Kindred Products	9	135	28	244	322	480	180	165
Tobacco Products	3	12	20	198	105	45	10	360
Textile Mill Products	140	280	36	224	294	57	168	10
Apparel And Other Finished Products Made From Fabrics And Similar Materials	150	150	75	96	108	322	328	15
Lumber And Wood Products, Except Furniture	92	115	16	123	343	464	204	50
Furniture And Fixtures	108	165	23	108	480	399	300	120
Paper And Allied Products	63	100	15	94	496	301	144	172
Printing, Publishing, And Allied Industries	250	294	208	38	111	330	464	92
Chemicals And Allied Products	385	276	472	6	12	33	160	84
Petroleum Refining And Related Industries	20	32	8	245	287	162	39	480
Rubber And Miscellaneous Plastics Products	105	357	172	44	456	225	252	108
Leather And Leather Products	54	168	104	28	360	120	150	44
Stone, Clay, Glass, And Concrete Products	20	217	76	240	440	195	60	354
Primary Metal Industries	312	448	200	29	260	87	378	100
Fabricated Metal Products, Except Machinery And Transportation	225	456	138	39	324	192	427	106

Equipment								
Industrial And Commercial Machinery And Computer Equipment	427	330	488	10	72	51	280	62
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	413	288	496	5	26	39	215	76
Transportation Equipment	156	480	105	51	413	250	354	96
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	392	348	480	11	24	42	245	120
Miscellaneous Manufacturing Industries	371	496	235	24	72	140	360	64
Railroad Transportation	12	64	3	348	488	378	48	260
Local And Suburban Transit And Interurban Highway Passenger Transportation	102	84	264	26	21	40	55	4
Motor Freight Transportation And Warehousing	170	308	96	70	424	120	222	18
Water Transportation	11	40	81	285	264	371	28	488
Transportation By Air	26	85	10	126	406	496	234	188
Pipelines, Except Natural Gas	154	42	240	7	9	2	45	32
Transportation Services	228	320	176	54	190	69	280	12
Communications	480	205	378	9	75	128	306	66
Electric, Gas, And Sanitary Services	5	52	4	496	210	125	36	406
Wholesale Trade-durable Goods	456	427	275	21	156	93	372	48
Wholesale Trade-non-durable Goods	220	416	153	25	172	294	336	102
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	144	222	66	64	329	235	368	16
General Merchandise Stores	96	110	9	72	350	472	228	81
Food Stores	24	95	7	220	231	328	138	108
Automotive Dealers And Gasoline Service Stations	128	170	42	30	357	336	440	132
Apparel And Accessory Stores	45	70	11	74	392	488	132	164
Home Furniture, Furnishings, And Equipment Stores	185	192	84	23	136	364	352	26
Eating And Drinking Places	294	400	144	19	84	140	399	28
Miscellaneous Retail	408	180	164	17	60	190	252	12
Hotels, Rooming Houses, Camps, And Other Lodging Places	75	252	240	80	384	136	186	23
Personal Services	116	203	38	156	216	408	85	17
Business Services	496	225	399	1	27	80	318	18

Automotive Repair, Services, And Parking	294	376	58	43	90	220	210	224
Miscellaneous Repair Services	2	30	3	424	112	8	15	6
Motion Pictures	432	228	210	8	69	104	315	22
Amusement And Recreation Services	288	424	111	26	160	200	364	98
Health Services	282	432	371	34	116	66	145	78
Legal Services	56	12	160	4	6	12	10	2
Educational Services	165	224	234	22	63	294	76	76
Social Services	301	312	136	31	66	165	162	58
Membership Organizations	280	126	464	4	1	16	30	6
Engineering, Accounting, Research, Management, And Related Services	406	354	448	16	20	72	235	78

Table 10-6 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 1998 to 2000 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.6 2001 to 2003

	Cluster							
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Agricultural Production Crops	12	36	224	112	21	102	50	3
Agriculture production livestock and animal specialties	4	35	18	3	10	3	8	96
Agricultural Services	70	30	120	72	10	105	36	7
Metal Mining	5	200	72	21	91	16	40	124
Coal Mining	28	196	2	15	66	8	35	480
Oil And Gas Extraction	64	343	9	24	80	18	78	496
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	4	192	33	78	60	10	77	160
Building Construction General Contractors And Operative Builders	130	44	80	336	114	63	147	41
Heavy Construction Other Than Building Construction Contractors	222	150	36	15	155	132	184	399
Construction Special Trade Contractors	136	106	185	246	336	102	360	56
Food And Kindred Products	44	290	96	216	371	72	272	44
Tobacco Products	48	84	10	3	24	18	25	360
Textile Mill Products	9	186	52	147	70	36	128	72
Apparel And Other Finished Products Made From Fabrics And Similar Materials	132	58	205	420	234	105	312	25
Lumber And Wood Products, Except Furniture	114	40	215	258	357	164	352	26
Furniture And Fixtures	120	132	125	385	300	40	280	32
Paper And Allied Products	175	208	14	264	496	81	371	110
Printing, Publishing, And Allied Industries	324	76	208	102	385	250	440	28
Chemicals And Allied Products	456	68	250	51	148	399	282	43
Petroleum Refining And Related Industries	80	329	8	22	102	42	75	464
Rubber And Miscellaneous Plastics Products	204	112	126	324	420	245	464	48
Leather And Leather Products	225	78	160	348	392	111	416	50
Stone, Clay, Glass, And Concrete Products	265	180	78	342	427	188	496	52
Primary Metal Industries	282	86	102	132	378	230	432	53
Fabricated Metal Products, Except Machinery And Transportation	195	171	132	318	399	76	400	54

Equipment								
Industrial And Commercial Machinery And Computer Equipment	488	38	360	78	164	420	255	35
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	496	34	280	72	160	413	294	33
Transportation Equipment	129	70	190	336	413	168	448	42
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	464	72	295	84	172	406	366	46
Miscellaneous Manufacturing Industries	176	22	424	190	66	357	156	15
Railroad Transportation	38	240	68	162	368	48	140	49
Local And Suburban Transit And Interurban Highway Passenger Transportation	66	2	496	24	9	196	20	1
Motor Freight Transportation And Warehousing	87	26	352	180	115	128	168	16
Water Transportation	45	427	10	76	100	26	108	488
Transportation By Air	276	102	81	175	464	144	252	37
Pipelines, Except Natural Gas	52	4	152	50	24	108	84	40
Transportation Services	200	92	144	357	87	416	168	27
Communications	413	30	270	54	144	488	246	29
Electric, Gas, And Sanitary Services	54	496	7	88	125	20	132	413
Wholesale Trade-durable Goods	275	64	330	180	126	385	480	23
Wholesale Trade-non-durable Goods	64	162	276	350	128	215	384	39
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	72	16	210	174	216	120	95	4
General Merchandise Stores	164	52	240	282	329	117	296	17
Food Stores	69	275	32	240	308	92	256	19
Automotive Dealers And Gasoline Service Stations	124	295	69	343	270	44	368	34
Apparel And Accessory Stores	245	54	96	434	294	160	472	22
Home Furniture, Furnishings, And Equipment Stores	250	32	78	288	120	371	304	9
Eating And Drinking Places	75	80	294	488	130	124	231	21
Miscellaneous Retail	260	42	464	148	102	324	294	11
Hotels, Rooming Houses, Camps, And Other Lodging Places	42	180	105	416	168	57	203	38
Personal Services	20	92	154	120	75	33	136	13
Business Services	420	28	366	75	132	496	200	24

Automotive Repair, Services, And Parking	81	60	210	322	105	100	216	10
Miscellaneous Repair Services	9	48	8	8	24	2	21	30
Motion Pictures	240	66	357	93	140	288	248	47
Amusement And Recreation Services	144	82	87	413	312	225	344	30
Health Services	84	74	329	234	96	220	240	8
Legal Services	5	2	2	32	6	4	7	6
Educational Services	294	36	186	69	72	145	200	18
Social Services	85	14	360	56	30	182	84	5
Membership Organizations	84	3	12	72	54	28	30	42
Engineering, Accounting, Research, Management, And Related Services	336	84	285	96	152	392	456	51

Table 10-7 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 2001 to 2003 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

10.2.7 2004 to 2006

	Cluster									
SIC Major Group (2-digit)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Agricultural Production Crops	70	72	198	1	140	90	80	28	18	6
Agriculture production livestock and animal specialties	250	48	70	30	63	9	60	44	30	54
Agricultural Services	24	77	270	180	96	7	54	96	24	50
Metal Mining	153	70	2	24	72	4	124	102	550	65
Coal Mining	65	184	39	36	133	11	56	315	580	96
Oil And Gas Extraction	44	136	5	12	126	16	165	108	610	234
Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	288	154	24	51	378	100	24	186	590	140
Building Construction General Contractors And Operative Builders	133	160	120	520	279	32	294	45	116	11
Heavy Construction Other Than Building Construction Contractors	192	340	69	12	198	150	72	126	308	76
Construction Special Trade Contractors	423	480	165	46	168	140	45	66	400	174
Food And Kindred Products	252	387	352	160	470	148	108	90	32	294
Tobacco Products	4	42	98	171	170	5	296	30	80	60
Textile Mill Products	90	250	63	189	207	272	348	76	14	30
Apparel And Other Finished Products Made From Fabrics And Similar Materials	220	470	148	495	280	114	80	240	35	488
Lumber And Wood Products, Except Furniture	14	216	130	42	300	174	7	140	344	68
Furniture And Fixtures	105	510	156	264	468	378	26	195	82	424
Paper And Allied Products	205	259	424	228	432	28	25	570	108	140
Printing, Publishing, And Allied Industries	300	392	549	205	550	385	102	81	39	128
Chemicals And Allied Products	318	570	357	136	408	549	94	75	33	215
Petroleum Refining And Related Industries	63	25	3	6	88	6	162	52	600	180
Rubber And Miscellaneous Plastics Products	225	500	432	234	396	371	118	87	23	124
Leather And Leather Products	148	224	175	580	272	81	30	405	76	240
Stone, Clay, Glass, And Concrete Products	416	369	60	144	570	255	50	252	159	357
Primary Metal Industries	155	288	93	116	210	38	32	540	513	336
Fabricated Metal Products, Except Machinery And Transportation	204	477	456	235	610	129	53	350	96	294

Equipment										
Industrial And Commercial Machinery And Computer Equipment	600	522	164	240	480	156	42	357	98	348
Electronic And Other Electrical Equipment And Components, Except Computer Equipment	610	495	270	230	448	413	38	129	80	188
Transportation Equipment	195	360	228	60	522	117	34	580	208	399
Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	522	610	260	180	472	406	52	111	84	336
Miscellaneous Manufacturing Industries	196	320	504	306	490	150	82	364	18	270
Railroad Transportation	160	189	116	198	256	36	8	610	90	405
Local And Suburban Transit And Interurban Highway Passenger Transportation	48	56	128	20	160	30	414	1	16	30
Motor Freight Transportation And Warehousing	60	105	14	224	24	39	5	600	222	324
Water Transportation	198	232	40	63	330	20	240	160	504	238
Transportation By Air	130	234	192	104	216	69	38	460	3	231
Pipelines, Except Natural Gas	30	28	1	104	20	2	550	6	306	16
Transportation Services	174	105	84	217	208	52	3	590	188	495
Communications	440	540	522	210	273	336	114	78	16	148
Electric, Gas, And Sanitary Services	115	270	57	14	224	128	610	8	357	150
Wholesale Trade-durable Goods	336	560	215	105	486	184	44	448	108	413
Wholesale Trade-non-durable Goods	276	312	540	185	530	164	48	99	92	322
Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	98	60	68	430	78	248	8	369	1	24
General Merchandise Stores	96	171	175	560	75	51	44	184	7	56
Food Stores	100	279	400	88	232	84	58	168	5	132
Automotive Dealers And Gasoline Service Stations	172	336	329	477	222	225	17	550	50	114
Apparel And Accessory Stores	216	78	33	610	50	84	9	306	20	126
Home Furniture, Furnishings, And Equipment Stores	110	108	30	540	140	66	13	477	112	312
Eating And Drinking Places	90	414	294	590	288	220	14	228	34	120
Miscellaneous Retail	324	468	413	600	400	126	86	240	24	200
Hotels, Rooming Houses, Camps, And Other Lodging Places	12	12	64	100	15	108	20	15	4	49
Personal Services	90	252	72	90	168	470	15	24	63	147
Business Services	590	540	288	250	368	420	70	132	26	208

Automotive Repair, Services, And Parking	224	64	306	120	24	400	1	196	12	144
Miscellaneous Repair Services	8	30	16	27	10	12	6	6	77	18
Motion Pictures	152	231	180	240	369	99	21	376	62	600
Amusement And Recreation Services	240	396	276	570	360	108	23	343	54	164
Health Services	240	380	450	100	304	336	168	28	22	135
Legal Services	36	4	42	160	3	30	4	36	2	40
Educational Services	105	280	550	60	175	513	12	27	38	138
Social Services	340	126	45	77	30	192	33	8	52	1
Engineering, Accounting, Research, Management, And Related Services	513	590	343	294	344	196	39	108	90	220

Table 10-8 This table shows the squared ranked correlations between GSC cluster means and SIC major groups (2-digit) for the 2004 to 2006 modelling interval. Darker shading indicates higher squared rank correlations while lighter shading indicates lower correlations.

Cluster	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	0.00107	0.00156	0.00195	0.00147	0.00191	0.00303	0.00358	0.00138
2	0.00171	0.00213	0.00150	0.00146	0.00155	0.00317	0.00367	0.00146
3	0.00146	0.00158	0.00167	0.00161	0.00161	0.00237	0.00216	0.00153
4	0.00157	0.00169	0.00154	0.00115	0.00042	0.00345	0.00297	0.00181
5	0.00163	0.00160	0.00161	0.00122	0.00160	0.00307	0.00369	0.00101
6	0.00164	0.00162	0.00159	0.00136	0.00154	0.00200	0.00255	0.00158
7	0.00136	0.00085	0.00125	0.00155	0.00147	0.00272	0.00312	0.00142
8	0.00148	0.00162	0.00157	0.00152	0.00153	0.00239	0.00321	0.00127
9	0.00051	0.00190	0.00106	0.00130	0.00183	0.00302	0.00289	0.00142
10	0.00127	0.00163	0.00113	0.00081	0.00160	0.00252	0.00334	0.00106
11	0.00156	0.00184	0.00110	0.00150	0.00101	0.00233	0.00218	0.00115
12	0.00165	0.00133	0.00143	0.00104	0.00178	0.00198	0.00272	0.00128
13	0.00175	0.00102	0.00160	0.00105	0.00199	0.00239	0.00180	0.00130
14	0.00133	0.00174	0.00108	0.00152	0.00177	0.00188	0.00222	0.00096
15	0.00129	0.00125	0.00159	0.00117	0.00165	0.00240	0.00155	0.00159
16	0.00105	0.00163	0.00154	0.00158	0.00170	0.00147	0.00107	0.00121
17	0.00139	0.00170	0.00110	0.00123	0.00174	0.00156	0.00237	0.00114
18	0.00125	0.00115	0.00120	0.00062	0.00102	0.00164	0.00314	0.00103
19	0.00128	0.00182	0.00129	0.00132	0.00109	0.00242	0.00147	0.00110
20	0.00087	0.00176	0.00143	0.00088	0.00140	0.00143	0.00207	0.00107
21	0.00113	0.00172	0.00132	0.00118	0.00128	0.00182	0.00215	0.00118
22	0.00098	0.00160	0.00121	0.00107	0.00136	0.00239	0.00254	0.00048
23	0.00071	0.00171	0.00105	0.00110	0.00121	0.00132	0.00166	0.00099
24	0.00111	0.00080	0.00151	0.00103	0.00120	0.00265	0.00155	0.00041
25	0.00114	0.00169	0.00160	0.00112	0.00089	0.00226	0.00274	0.00078
26	0.00093	0.00085	0.00122	0.00097	0.00089	0.00164	0.00071	0.00108
27	0.00141	0.00171	0.00137	0.00138	0.00114	0.00151	0.00222	0.00152
28	0.00135	0.00149	0.00167	0.00105	0.00119	0.00320	0.00171	0.00090
29	0.00122	0.00159	0.00131	0.00094	0.00092	0.00181	0.00181	0.00091
30	0.00108	0.00128	0.00138	0.00077	0.00124	0.00170	0.00175	0.00115
31	0.00116	0.00143	0.00139	0.00135	0.00139	0.00144	0.00267	0.00045
32	0.00131	0.00186	0.00117	0.00072	0.00106	0.00165	0.00135	0.00142
33	0.00104	0.00185	0.00106	0.00063	0.00050	0.00120	0.00111	0.00115
34	0.00091	0.00192	0.00120	0.00114	0.00165	0.00124	0.00110	0.00115
35	0.00091	0.00192	0.00120	0.00114	0.00165	0.00124	0.00110	0.00115
36	0.00118	0.00170	0.00092	0.00135	0.00088	0.00216	0.00221	0.00151
37	0.00060	0.00167	0.00043	0.00096	0.00101	0.00138	0.00050	0.00071
38	0.00058	0.00155	0.00078	0.00091	0.00082	0.00288	0.00132	0.00083
39	0.00109	0.00152	0.00123	0.00056	0.00069	0.00068	0.00127	0.00096
40	0.00101	0.00127	0.00114	0.00042	0.00113	0.00140	0.00103	0.00097
41	0.00068	0.00150	0.00112	0.00096	0.00153	0.00068	0.00103	0.00108
42	0.00097	0.00143	0.00098	0.00088	0.00120	0.00144	0.00178	0.00055
43	0.00114	0.00186	0.00060	0.00116	0.00068	0.00111	0.00223	0.00095
44	0.00062	0.00147	0.00047	0.00085	0.00070	0.00106	0.00114	0.00064
45	0.00127	0.00156	0.00085	0.00060	0.00087	0.00136	0.00088	0.00073
46	0.00095	0.00148	0.00079	0.00087	0.00076	0.00121	0.00149	0.00066
47	0.00087	0.00131	0.00075	0.00090	0.00117	0.00141	0.00112	0.00067
48	0.00070	0.00079	0.00066	0.00083	0.00110	0.00165	0.00117	0.00156
49	0.00122	0.00130	0.00067	0.00070	0.00100	0.00094	0.00054	0.00090
50	0.00090	0.00117	0.00075	0.00060	0.00066	0.00068	0.00086	0.00070
$\frac{\sigma(\sigma_K)}{K}$	**0.00032**	**0.00030**	**0.00034**	**0.00031**	**0.00039**	**0.00072**	**0.00086**	**0.00033**

Table 10-9 This table contains the standardised cross sectional standard deviations of cluster mean returns for the GSC (K = 50) over the modelling period. As 50 clusters are estimated, there are 50 cross sectional standard deviation measures for each year and each is divided by 50.

SIC	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	Agricultural Production Crops	0.00093	0.00132	0.00104	0.00057	0.00082	0.00099	0.00125	0.00064
2	Agriculture production livestock and animal specialties					0.00051	0.00165	0.00194	0.00173
7	Agricultural Services	0.00179	0.00157	0.00114	0.00110	0.00123	0.00119	0.00132	0.00085
10	Metal Mining	0.00132	0.00138	0.00086	0.00093	0.00118	0.00156	0.00162	0.00145
12	Coal Mining	0.00131	0.00178	0.00116	0.00095	0.00089	0.00182	0.00190	0.00154
13	Oil And Gas Extraction	0.00093	0.00117	0.00079	0.00062	0.00088	0.00150	0.00126	0.00110
14	Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	0.00066	0.00130	0.00074	0.00060	0.00060	0.00111	0.00124	0.00102
15	Building Construction General Contractors And Operative Builders	0.00106	0.00137	0.00119	0.00104	0.00073	0.00102	0.00134	0.00111
16	Heavy Construction Other Than Building Construction Contractors	0.00113	0.00140	0.00120	0.00073	0.00091	0.00107	0.00147	0.00104
17	Construction Special Trade Contractors	0.00124	0.00186	0.00151	0.00098	0.00094	0.00107	0.00145	0.00100
20	Food And Kindred Products	0.00049	0.00102	0.00072	0.00035	0.00052	0.00072	0.00065	0.00039
21	Tobacco Products	0.00085	0.00129	0.00103	0.00095	0.00113	0.00166	0.00166	0.00071
22	Textile Mill Products	0.00103	0.00135	0.00096	0.00062	0.00061	0.00101	0.00151	0.00075
23	Apparel And Other Finished Products Made From Fabrics And Similar Materials	0.00094	0.00123	0.00105	0.00076	0.00070	0.00100	0.00118	0.00067
24	Lumber And Wood Products, Except Furniture	0.00096	0.00139	0.00101	0.00097	0.00074	0.00088	0.00135	0.00072
25	Furniture And Fixtures	0.00090	0.00116	0.00075	0.00073	0.00066	0.00096	0.00098	0.00064
26	Paper And Allied Products	0.00083	0.00128	0.00098	0.00057	0.00065	0.00090	0.00099	0.00072
27	Printing, Publishing, And Allied Industries	0.00073	0.00118	0.00081	0.00051	0.00056	0.00080	0.00087	0.00051
28	Chemicals And Allied Products	0.00077	0.00126	0.00090	0.00070	0.00087	0.00146	0.00146	0.00078
29	Petroleum Refining And Related Industries	0.00078	0.00112	0.00066	0.00051	0.00052	0.00103	0.00098	0.00091
30	Rubber And Miscellaneous Plastics Products	0.00077	0.00115	0.00090	0.00060	0.00056	0.00085	0.00111	0.00069
31	Leather And Leather Products	0.00099	0.00122	0.00104	0.00069	0.00089	0.00104	0.00130	0.00074
32	Stone, Clay, Glass, And Concrete Products	0.00090	0.00121	0.00090	0.00081	0.00071	0.00097	0.00115	0.00084
33	Primary Metal Industries	0.00089	0.00129	0.00084	0.00079	0.00075	0.00123	0.00146	0.00100
34	Fabricated Metal Products, Except Machinery And Transportation Equipment	0.00074	0.00114	0.00075	0.00057	0.00057	0.00082	0.00101	0.00062
35	Industrial And Commercial Machinery And Computer Equipment	0.00089	0.00126	0.00102	0.00070	0.00094	0.00137	0.00167	0.00084
36	Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.00104	0.00125	0.00091	0.00071	0.00100	0.00176	0.00194	0.00093

37	Transportation Equipment	0.00090	0.00123	0.00093	0.00064	0.00062	0.00094	0.00132	0.00073
38	Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.00098	0.00121	0.00091	0.00071	0.00082	0.00138	0.00145	0.00076
39	Miscellaneous Manufacturing Industries	0.00092	0.00129	0.00110	0.00070	0.00079	0.00095	0.00125	0.00071
40	Railroad Transportation	0.00095	0.00141	0.00095	0.00059	0.00068	0.00097	0.00098	0.00080
41	Local And Suburban Transit And Interurban Highway Passenger Transportation				0.00206	0.00142	0.00347	0.00334	0.00109
42	Motor Freight Transportation And Warehousing	0.00098	0.00127	0.00098	0.00067	0.00067	0.00099	0.00109	0.00075
44	Water Transportation	0.00096	0.00141	0.00101	0.00060	0.00081	0.00109	0.00110	0.00078
45	Transportation By Air	0.00087	0.00119	0.00120	0.00095	0.00082	0.00101	0.00168	0.00090
46	Pipelines, Except Natural Gas					0.00144	0.00252	0.00202	0.00091
47	Transportation Services	0.00105	0.00130	0.00107	0.00073	0.00083	0.00141	0.00155	0.00087
48	Communications	0.00067	0.00099	0.00091	0.00061	0.00081	0.00143	0.00159	0.00059
49	Electric, Gas, And Sanitary Services	0.00043	0.00076	0.00046	0.00039	0.00042	0.00069	0.00070	0.00042
50	Wholesale Trade-durable Goods	0.00088	0.00119	0.00090	0.00069	0.00081	0.00113	0.00131	0.00075
51	Wholesale Trade-non-durable Goods	0.00075	0.00102	0.00076	0.00053	0.00061	0.00098	0.00098	0.00067
52	Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	0.00131	0.00176	0.00134	0.00090	0.00096	0.00116	0.00127	0.00090
53	General Merchandise Stores	0.00100	0.00140	0.00110	0.00078	0.00091	0.00105	0.00133	0.00067
54	Food Stores	0.00064	0.00111	0.00076	0.00059	0.00046	0.00067	0.00098	0.00072
55	Automotive Dealers And Gasoline Service Stations	0.00106	0.00108	0.00109	0.00092	0.00082	0.00105	0.00136	0.00082
56	Apparel And Accessory Stores	0.00122	0.00156	0.00134	0.00091	0.00099	0.00120	0.00160	0.00081
57	Home Furniture, Furnishings, And Equipment Stores	0.00109	0.00154	0.00134	0.00099	0.00099	0.00131	0.00197	0.00080
58	Eating And Drinking Places	0.00091	0.00110	0.00095	0.00081	0.00074	0.00083	0.00099	0.00066
59	Miscellaneous Retail	0.00089	0.00126	0.00103	0.00066	0.00071	0.00120	0.00142	0.00067
70	Hotels, Rooming Houses, Camps, And Other Lodging Places	0.00053	0.00127	0.00098	0.00076	0.00069	0.00094	0.00122	0.00078
72	Personal Services	0.00086	0.00115	0.00081	0.00073	0.00075	0.00103	0.00098	0.00067
73	Business Services	0.00095	0.00123	0.00089	0.00071	0.00093	0.00172	0.00189	0.00079
75	Automotive Repair, Services, And Parking	0.00115	0.00121	0.00129	0.00085	0.00076	0.00117	0.00145	0.00063
76	Miscellaneous Repair Services	0.00161	0.00138	0.00153	0.00144	0.00145	0.00198	0.00208	0.00137
78	Motion Pictures	0.00090	0.00128	0.00086	0.00071	0.00081	0.00125	0.00138	0.00065
79	Amusement And Recreation Services	0.00097	0.00116	0.00110	0.00101	0.00100	0.00103	0.00121	0.00081

80	Health Services	0.00120	0.00121	0.00092	0.00083	0.00080	0.00112	0.00112	0.00058
81	Legal Services						0.00240	0.00216	0.00103
82	Educational Services	0.00185	0.00144	0.00130	0.00104	0.00102	0.00112	0.00112	0.00082
83	Social Services		0.00204	0.00206	0.00162	0.00108	0.00124	0.00136	0.00069
86	Membership Organizations						0.00436	0.00291	
87	Engineering, Accounting, Research, Management, And Related Services	0.00090	0.00117	0.00084	0.00063	0.00077	0.00120	0.00145	0.00070
99	Nonclassifiable Establishments	0.00100	0.00109	0.00095	0.00072	0.00094	0.00153	0.00170	0.00063
	$\frac{\sigma(\sigma_K)}{K}$	**0.00026**	**0.00021**	**0.00025**	**0.00028**	**0.00022**	**0.00061**	**0.00046**	**0.00024**

Table 10-10 This table contains the cross sectional standard deviations of industry group means for the SIC classification scheme at the 2-digit level. There are 63 SIC groups at this level. Categories that contained fewer than 5 stocks per group were regarded as non-functional and removed.

NAICS	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
42		0.00085	0.00110	0.00096	0.00078	0.00068	0.00127	0.00090	0.00075
111	Crop Production	0.00064	0.00093	0.00070	0.00040	0.00055	0.00068	0.00087	0.00045
112	Animal Production					0.00036	0.00115	0.00135	0.00120
211	Oil and Gas Extraction	0.00061	0.00076	0.00047	0.00043	0.00057	0.00095	0.00081	0.00074
212	Mining (except Oil and Gas)	0.00062	0.00085	0.00047	0.00048	0.00058	0.00070	0.00090	0.00081
213	Support Activities for Mining	0.00077	0.00105	0.00084	0.00053	0.00087	0.00154	0.00120	0.00088
221	Utilities	0.00030	0.00050	0.00031	0.00028	0.00029	0.00047	0.00049	0.00030
233		0.00083	0.00093	0.00078	0.00082	0.00051	0.00079	0.00121	0.00059
234				0.00117	0.00082	0.00079	0.00111	0.00133	
235		0.00082	0.00156	0.00115	0.00111	0.00098	0.00123	0.00219	
236	Construction of Buildings	0.00069	0.00096	0.00090	0.00066	0.00054	0.00077	0.00089	0.00077
237	Heavy and Civil Engineering Construction	0.00087	0.00103	0.00080	0.00061	0.00070	0.00075	0.00109	0.00076
238	Specialty Trade Contractors	0.00107	0.00146	0.00115	0.00074	0.00067	0.00078	0.00100	0.00069
311	Food Manufacturing	0.00034	0.00070	0.00049	0.00025	0.00036	0.00053	0.00047	0.00028
312	Beverage and Tobacco Product Manufacturing	0.00051	0.00074	0.00062	0.00029	0.00046	0.00051	0.00051	0.00029
313	Textile Mills	0.00077	0.00095	0.00065	0.00047	0.00044	0.00076	0.00103	0.00053
314	Textile Product Mills	0.00105	0.00109	0.00095	0.00083	0.00063	0.00085	0.00105	0.00071
315	Apparel Manufacturing	0.00063	0.00087	0.00071	0.00049	0.00048	0.00069	0.00084	0.00047
316	Leather and Allied Product Manufacturing	0.00058	0.00084	0.00076	0.00048	0.00061	0.00075	0.00086	0.00049
321	Wood Product Manufacturing	0.00067	0.00097	0.00070	0.00067	0.00053	0.00061	0.00094	0.00051
322	Paper Manufacturing	0.00059	0.00090	0.00069	0.00040	0.00045	0.00062	0.00069	0.00050
323	Printing and Related Support Activities	0.00055	0.00088	0.00062	0.00040	0.00040	0.00062	0.00062	0.00044
324	Petroleum and Coal Products Manufacturing	0.00054	0.00079	0.00045	0.00037	0.00034	0.00073	0.00067	0.00065
325	Chemical Manufacturing	0.00053	0.00087	0.00063	0.00048	0.00060	0.00101	0.00101	0.00054
326	Plastics and Rubber Products Manufacturing	0.00056	0.00080	0.00060	0.00042	0.00038	0.00057	0.00078	0.00048
327	Nonmetallic Mineral Product Manufacturing	0.00063	0.00084	0.00063	0.00056	0.00049	0.00067	0.00080	0.00058
331	Primary Metal Manufacturing	0.00063	0.00089	0.00057	0.00056	0.00052	0.00086	0.00100	0.00075
332	Fabricated Metal Product Manufacturing	0.00050	0.00082	0.00055	0.00038	0.00038	0.00058	0.00066	0.00044
333	Machinery Manufacturing	0.00061	0.00086	0.00064	0.00044	0.00058	0.00083	0.00098	0.00057
334	Computer and Electronic Product Manufacturing	0.00074	0.00088	0.00068	0.00052	0.00070	0.00122	0.00134	0.00063
335	Electrical Equipment, Appliance, and Component Manufacturing	0.00063	0.00082	0.00057	0.00045	0.00053	0.00083	0.00096	0.00049
336	Transportation Equipment Manufacturing	0.00062	0.00085	0.00064	0.00044	0.00044	0.00066	0.00091	0.00050
337	Furniture and Related Product Manufacturing	0.00064	0.00081	0.00056	0.00056	0.00048	0.00067	0.00073	0.00047
339	Miscellaneous Manufacturing	0.00062	0.00084	0.00069	0.00050	0.00053	0.00074	0.00085	0.00047

421		0.00066	0.00083	0.00068	0.00047	0.00057	0.00084	0.00128	
422		0.00058	0.00088	0.00051	0.00047	0.00053	0.00086	0.00109	0.00092
423	Merchant Wholesalers, Durable Goods	0.00062	0.00083	0.00063	0.00051	0.00058	0.00081	0.00092	0.00052
424	Merchant Wholesalers, Nondurable Goods	0.00046	0.00063	0.00053	0.00031	0.00038	0.00061	0.00063	0.00046
425	Wholesale Electronic Markets and Agents and Brokers	0.00151	0.00119	0.00140	0.00112	0.00070	0.00103	0.00099	0.00080
441	Motor Vehicle and Parts Dealers	0.00075	0.00082	0.00076	0.00060	0.00058	0.00076	0.00098	0.00058
442	Furniture and Home Furnishings Stores	0.00098	0.00110	0.00088	0.00071	0.00073	0.00086	0.00136	0.00059
443	Electronics and Appliance Stores	0.00079	0.00111	0.00096	0.00078	0.00069	0.00124	0.00154	0.00065
444	Building Material and Garden Equipment and Supplies Dealers	0.00091	0.00122	0.00093	0.00062	0.00066	0.00081	0.00078	0.00056
445	Food and Beverage Stores	0.00045	0.00077	0.00053	0.00041	0.00032	0.00045	0.00069	0.00051
446	Health and Personal Care Stores	0.00060	0.00091	0.00068	0.00040	0.00050	0.00076	0.00083	0.00048
447	Gasoline Stations	0.00092	0.00105	0.00119	0.00112	0.00085	0.00101	0.00116	0.00092
448	Clothing and Clothing Accessories Stores	0.00085	0.00108	0.00089	0.00060	0.00065	0.00083	0.00109	0.00054
451	Sporting Goods, Hobby, Book, and Music Stores	0.00082	0.00111	0.00086	0.00055	0.00062	0.00079	0.00110	0.00060
452	General Merchandise Stores	0.00070	0.00097	0.00076	0.00054	0.00063	0.00073	0.00092	0.00046
453	Miscellaneous Store Retailers	0.00077	0.00079	0.00084	0.00061	0.00066	0.00103	0.00120	0.00054
454	Nonstore Retailers	0.00105	0.00094	0.00093	0.00059	0.00061	0.00109	0.00116	0.00055
481	Air Transportation	0.00068	0.00089	0.00089	0.00068	0.00059	0.00077	0.00133	0.00070
482	Rail Transportation	0.00066	0.00098	0.00066	0.00041	0.00047	0.00067	0.00068	0.00056
483	Water Transportation	0.00069	0.00096	0.00075	0.00043	0.00055	0.00071	0.00066	0.00053
484	Truck Transportation	0.00072	0.00090	0.00071	0.00048	0.00047	0.00070	0.00084	0.00057
485	Transit and Ground Passenger Transportation				0.00143	0.00149	0.00228		0.00075
486	Pipeline Transportation	0.00058	0.00087	0.00061	0.00045	0.00039	0.00077	0.00099	0.00054
487	Scenic and Sightseeing Transportation				0.00107	0.00111	0.00167	0.00271	
488	Support Activities for Transportation	0.00071	0.00094	0.00073	0.00052	0.00052	0.00086	0.00098	0.00062
492	Couriers and Messengers	0.00076	0.00082	0.00080	0.00064	0.00068	0.00081	0.00079	0.00055
493	Warehousing and Storage	0.00071	0.00099	0.00105	0.00077	0.00076	0.00122	0.00096	0.00060
511	Publishing Industries (except Internet)	0.00059	0.00086	0.00067	0.00048	0.00068	0.00127	0.00138	0.00055
512	Motion Picture and Sound Recording Industries	0.00066	0.00090	0.00062	0.00046	0.00056	0.00092	0.00097	0.00046
513		0.00058	0.00075	0.00074	0.00046	0.00061	0.00115	0.00148	0.00054
514		0.00084	0.00096	0.00070	0.00057	0.00067	0.00150	0.00165	
515	Broadcasting (except Internet)	0.00056	0.00078	0.00055	0.00059	0.00057	0.00089	0.00112	0.00039
516	Internet Publishing and Broadcasting					0.00136	0.00154		0.00153
517	Telecommunications	0.00045	0.00069	0.00060	0.00044	0.00058	0.00099	0.00111	0.00047

518	Internet Service Providers, Web Search Portals, and Data Processing Services	0.00077	0.00103	0.00077	0.00057	0.00069	0.00123	0.00143	0.00057
519	Other Information Services						0.00206	0.00168	0.00125
522	Credit Intermediation and Related Activities	0.00108	0.00120	0.00077	0.00068	0.00045	0.00104	0.00086	0.00060
523	Securities, Commodity Contracts, and Other Financial Investments and Related Activities	0.00129				0.00056	0.00114	0.00098	0.00112
524	Insurance Carriers and Related Activities			0.00099	0.00157	0.00117	0.00214	0.00133	0.00110
525	Funds, Trusts, and Other Financial Vehicles	0.00090	0.00142	0.00117	0.00092	0.00099	0.00139	0.00104	0.00121
532	Rental and Leasing Services	0.00067	0.00085	0.00067	0.00041	0.00044	0.00078	0.00087	0.00048
541	Professional, Scientific, and Technical Services	0.00063	0.00084	0.00056	0.00045	0.00059	0.00101	0.00115	0.00052
561	Administrative and Support Services	0.00064	0.00083	0.00058	0.00048	0.00057	0.00090	0.00101	0.00048
562	Waste Management and Remediation Services	0.00070	0.00102	0.00065	0.00054	0.00071	0.00080	0.00070	0.00034
611	Educational Services	0.00117	0.00109	0.00083	0.00077	0.00071	0.00082	0.00077	0.00055
621	Ambulatory Health Care Services	0.00081	0.00087	0.00062	0.00060	0.00061	0.00081	0.00082	0.00046
622	Hospitals	0.00120	0.00108	0.00098	0.00066	0.00053	0.00094	0.00077	0.00047
623	Nursing and Residential Care Facilities	0.00115	0.00114	0.00111	0.00086	0.00070	0.00085	0.00103	0.00040
624	Social Assistance		0.00142	0.00143	0.00129	0.00099	0.00095	0.00087	0.00065
711	Performing Arts, Spectator Sports, and Related Industries	0.00133	0.00097	0.00147	0.00125	0.00076	0.00098	0.00102	0.00058
713	Amusement, Gambling, and Recreation Industries	0.00050	0.00089	0.00067	0.00090	0.00073	0.00074	0.00089	0.00056
721	Accommodation	0.00053	0.00084	0.00070	0.00061	0.00059	0.00064	0.00085	0.00061
722	Food Services and Drinking Places	0.00062	0.00077	0.00066	0.00058	0.00052	0.00059	0.00069	0.00047
811	Repair and Maintenance	0.00104	0.00076	0.00093	0.00071	0.00069	0.00112	0.00111	0.00061
812	Personal and Laundry Services	0.00059	0.00089	0.00061	0.00051	0.00050	0.00077	0.00075	0.00054
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations						0.00303	0.00202	
	$\frac{\sigma(\sigma_K)}{K}$	**0.00022**	**0.00018**	**0.00022**	**0.00025**	**0.00020**	**0.00040**	**0.00035**	**0.00022**

Table 10-11 This table contains the cross sectional standard deviations of industry group means for the NAICS classification scheme at the 3-digit level. There are 83 NAICS groups at this level. Categories that contained fewer than 5 stocks per group were regarded as non-functional and removed.

GICS	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
101010	Energy Equipment & Services	0.00087	0.00122	0.00100	0.00059	0.00100	0.00162	0.00133	0.00099
101020	Oil, Gas & Consumable Fuels	0.00069	0.00094	0.00057	0.00047	0.00063	0.00109	0.00097	0.00091
151010	Chemicals	0.00068	0.00103	0.00072	0.00041	0.00045	0.00084	0.00080	0.00055
151020	Construction Materials	0.00071	0.00100	0.00080	0.00065	0.00055	0.00092	0.00086	0.00077
151030	Containers & Packaging	0.00070	0.00109	0.00083	0.00047	0.00052	0.00077	0.00089	0.00061
151040	Metals & Mining	0.00081	0.00107	0.00059	0.00064	0.00066	0.00100	0.00120	0.00105
151050	Paper & Forest Products	0.00080	0.00113	0.00090	0.00069	0.00069	0.00088	0.00084	0.00071
201010	Aerospace & Defense	0.00081	0.00106	0.00072	0.00046	0.00063	0.00081	0.00099	0.00057
201020	Building Products	0.00080	0.00112	0.00078	0.00065	0.00049	0.00087	0.00109	0.00071
201030	Construction & Engineering	0.00064	0.00099	0.00083	0.00053	0.00058	0.00070	0.00101	0.00068
201040	Electrical Equipment	0.00077	0.00105	0.00068	0.00055	0.00070	0.00117	0.00134	0.00070
201050	Industrial Conglomerates	0.00074	0.00104	0.00072	0.00043	0.00042	0.00073	0.00090	0.00054
201060	Machinery	0.00070	0.00106	0.00078	0.00049	0.00056	0.00077	0.00085	0.00064
201070	Trading Companies & Distributors	0.00073	0.00093	0.00066	0.00046	0.00061	0.00084	0.00102	0.00071
202010	Commercial Services & Supplies	0.00075	0.00102	0.00071	0.00054	0.00062	0.00090	0.00100	0.00052
203010	Air Freight & Logistics	0.00083	0.00108	0.00093	0.00074	0.00075	0.00105	0.00097	0.00074
203020	Airlines	0.00089	0.00108	0.00105	0.00089	0.00073	0.00100	0.00190	0.00098
203030	Marine	0.00091	0.00123	0.00096	0.00057	0.00053	0.00063	0.00073	0.00068
203040	Road & Rail	0.00081	0.00107	0.00084	0.00054	0.00053	0.00084	0.00094	0.00063
203050	Transportation Infrastructure	0.00092	0.00143	0.00110	0.00108	0.00084	0.00091	0.00166	0.00086
251010	Auto Components	0.00073	0.00103	0.00078	0.00064	0.00053	0.00085	0.00119	0.00064
251020	Automobiles	0.00100	0.00135	0.00099	0.00078	0.00058	0.00101	0.00163	0.00094
252010	Household Durables	0.00083	0.00109	0.00078	0.00067	0.00052	0.00075	0.00093	0.00061
252020	Leisure Equipment & Products	0.00071	0.00110	0.00103	0.00059	0.00063	0.00084	0.00109	0.00069
252030	Textiles, Apparel & Luxury Goods	0.00073	0.00104	0.00081	0.00055	0.00058	0.00081	0.00099	0.00055
253010	Hotels, Restaurants & Leisure	0.00072	0.00096	0.00081	0.00070	0.00069	0.00072	0.00092	0.00058
253020	Diversified Consumer Services	0.00097	0.00100	0.00066	0.00061	0.00064	0.00098	0.00080	0.00055
254010	Media	0.00060	0.00102	0.00069	0.00049	0.00059	0.00099	0.00114	0.00043
255010	Distributors	0.00077	0.00098	0.00082	0.00071	0.00071	0.00103	0.00102	0.00064
255020	Internet & Catalog Retail	0.00126	0.00110	0.00129	0.00071	0.00080	0.00140	0.00173	0.00078
255030	Multiline Retail	0.00087	0.00122	0.00097	0.00069	0.00079	0.00096	0.00125	0.00061
255040	Specialty Retail	0.00084	0.00118	0.00100	0.00065	0.00074	0.00091	0.00127	0.00063
301010	Food & Staples Retailing	0.00059	0.00096	0.00066	0.00045	0.00040	0.00059	0.00079	0.00055

302010	Beverages	0.00052	0.00089	0.00073	0.00032	0.00050	0.00062	0.00056	0.00035
302020	Food Products	0.00042	0.00089	0.00063	0.00032	0.00046	0.00063	0.00060	0.00037
302030	Tobacco	0.00073	0.00082	0.00101	0.00084	0.00082	0.00127	0.00103	0.00062
303010	Household Products	0.00060	0.00096	0.00072	0.00041	0.00043	0.00075	0.00051	0.00046
303020	Personal Products	0.00060	0.00108	0.00079	0.00079	0.00070	0.00099	0.00103	0.00065
351010	Health Care Equipment & Supplies	0.00083	0.00104	0.00082	0.00065	0.00073	0.00113	0.00114	0.00062
351020	Health Care Providers & Services	0.00085	0.00107	0.00082	0.00066	0.00065	0.00097	0.00096	0.00048
351030	Health Care Technology	0.00167	0.00173	0.00123	0.00113	0.00094	0.00123	0.00130	0.00068
352010	Biotechnology	0.00122	0.00138	0.00104	0.00093	0.00107	0.00204	0.00170	0.00087
352020	Pharmaceuticals	0.00073	0.00109	0.00090	0.00064	0.00082	0.00124	0.00125	0.00065
352030	Life Sciences Tools & Services	0.00096	0.00122	0.00094	0.00067	0.00070	0.00150	0.00141	0.00067
401010	Commercial Banks				0.00145	0.00083	0.00210	0.00208	0.00166
402010	Diversified Financial Services	0.00127	0.00129	0.00087	0.00087	0.00072	0.00112	0.00135	0.00068
402020	Consumer Finance		0.00120	0.00128	0.00104	0.00073	0.00121	0.00119	0.00111
402030	Capital Markets	0.00161				0.00070	0.00141	0.00121	0.00140
403010	Insurance	0.00079	0.00100	0.00097	0.00064	0.00066	0.00102	0.00061	0.00062
404010	Real Estate -- Discontinued effective 04/28/2006				0.00091	0.00130	0.00146		
404020	Real Estate Investment Trusts (REITs)	0.00090	0.00146	0.00124	0.00077	0.00073	0.00119	0.00101	0.00120
404030	Real Estate Management & Development		0.00177	0.00139	0.00134	0.00104	0.00166	0.00173	0.00100
451010	Internet Software & Services	0.00096	0.00139	0.00080	0.00084	0.00092	0.00196	0.00205	0.00084
451020	IT Services	0.00091	0.00108	0.00075	0.00056	0.00085	0.00135	0.00142	0.00060
451030	Software	0.00100	0.00116	0.00095	0.00070	0.00093	0.00170	0.00180	0.00071
452010	Communications Equipment	0.00090	0.00109	0.00081	0.00067	0.00096	0.00184	0.00192	0.00094
452020	Computers & Peripherals	0.00095	0.00114	0.00099	0.00076	0.00095	0.00149	0.00175	0.00077
452030	Electronic Equipment, Instruments & Components	0.00091	0.00106	0.00077	0.00059	0.00082	0.00132	0.00151	0.00070
452040	Office Electronics	0.00101	0.00133	0.00120	0.00105	0.00101	0.00134	0.00207	0.00073
452050	Semiconductor Equipment & Products -- Discontinued effective 04/30/2003.	0.00110	0.00125	0.00096	0.00080	0.00114	0.00181	0.00326	
453010	Semiconductors & Semiconductor Equipment	0.00126	0.00139	0.00108	0.00089	0.00133	0.00206	0.00226	0.00102
501010	Diversified Telecommunication Services	0.00058	0.00078	0.00071	0.00047	0.00069	0.00137	0.00127	0.00056
501020	Wireless Telecommunication Services	0.00065	0.00112	0.00118	0.00080	0.00094	0.00159	0.00175	0.00069
551010	Electric Utilities	0.00043	0.00064	0.00043	0.00040	0.00044	0.00064	0.00068	0.00037
551020	Gas Utilities	0.00032	0.00059	0.00034	0.00037	0.00033	0.00062	0.00050	0.00041
551020	Gas Utilities	0.00032	0.00059	0.00034	0.00037	0.00033	0.00062	0.00050	0.00041
551030	Multi-Utilities	0.00046	0.00066	0.00048	0.00041	0.00045	0.00078	0.00073	0.00036
551040	Water Utilities	0.00050	0.00066	0.00040	0.00033	0.00037	0.00061	0.00046	0.00055

551050	Independent Power Producers & Energy Traders	0.00048	0.00091	0.00059	0.00053	0.00046	0.00121	0.00122	0.00065
	$\frac{\sigma(\sigma_K)}{K}$	**0.00026**	**0.00022**	**0.00022**	**0.00023**	**0.00022**	**0.00039**	**0.00050**	**0.00025**

Table 10-12 This table contains the cross sectional standard deviations of industry group means for the GICS classification scheme at the 6-digit level. There are 69 GICS groups at this level. Categories that contained fewer than 5 stocks per group were regarded as non-functional and removed.

FF	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	Agriculture	0.00111	0.00147	0.00116	0.00070	0.00094	0.00102	0.00130	0.00074
2	Food Products	0.00058	0.00122	0.00084	0.00044	0.00061	0.00089	0.00080	0.00051
3	Candy & Soda	0.00099	0.00167	0.00136	0.00084	0.00097	0.00106	0.00093	0.00071
4	Beer & Liquor	0.00086	0.00116	0.00099	0.00050	0.00083	0.00097	0.00087	0.00043
5	Tobacco Products	0.00100	0.00153	0.00122	0.00113	0.00134	0.00197	0.00198	0.00085
6	Recreation	0.00131	0.00161	0.00127	0.00095	0.00102	0.00122	0.00144	0.00093
7	Entertainment	0.00102	0.00141	0.00110	0.00090	0.00106	0.00125	0.00143	0.00088
8	Printing and Publishing	0.00089	0.00136	0.00092	0.00059	0.00067	0.00094	0.00105	0.00058
9	Consumer Goods	0.00089	0.00135	0.00100	0.00081	0.00075	0.00104	0.00119	0.00075
10	Apparel	0.00108	0.00144	0.00124	0.00083	0.00086	0.00115	0.00140	0.00078
11	Healthcare	0.00143	0.00144	0.00109	0.00099	0.00096	0.00133	0.00133	0.00069
12	Medical Equipment	0.00117	0.00150	0.00116	0.00097	0.00098	0.00159	0.00162	0.00090
13	Pharmaceutical Products	0.00101	0.00164	0.00122	0.00104	0.00126	0.00217	0.00199	0.00103
14	Chemicals	0.00100	0.00148	0.00097	0.00058	0.00065	0.00125	0.00123	0.00077
15	Rubber and Plastic Products	0.00093	0.00133	0.00103	0.00073	0.00067	0.00101	0.00131	0.00087
16	Textiles	0.00122	0.00161	0.00114	0.00073	0.00072	0.00120	0.00180	0.00089
17	Construction Materials	0.00094	0.00141	0.00098	0.00082	0.00070	0.00097	0.00128	0.00078
18	Construction	0.00116	0.00156	0.00131	0.00096	0.00090	0.00110	0.00151	0.00113
19	Steel Works Etc	0.00106	0.00153	0.00099	0.00094	0.00089	0.00147	0.00173	0.00118
20	Fabricated Products	0.00113	0.00147	0.00103	0.00080	0.00077	0.00120	0.00146	0.00092
21	Machinery	0.00102	0.00152	0.00113	0.00074	0.00100	0.00133	0.00157	0.00099
22	Electrical Equipment	0.00108	0.00138	0.00098	0.00070	0.00089	0.00139	0.00161	0.00086
23	Automobiles and Trucks	0.00104	0.00147	0.00114	0.00087	0.00074	0.00117	0.00177	0.00092
24	Aircraft	0.00112	0.00152	0.00102	0.00062	0.00078	0.00118	0.00169	0.00084
25	Shipbuilding, Railroad Equipment	0.00121	0.00162	0.00140	0.00074	0.00101	0.00127	0.00121	0.00112
26	Defense	0.00110	0.00133	0.00113	0.00103	0.00096	0.00105	0.00114	0.00090
27	Precious Metals	0.00177	0.00170	0.00110	0.00129	0.00162	0.00215	0.00249	0.00173
28	Non-Metallic and Industrial Metal Mining	0.00080	0.00156	0.00090	0.00067	0.00076	0.00120	0.00156	0.00135
29	Coal	0.00156	0.00211	0.00138	0.00113	0.00106	0.00216	0.00225	0.00183
30	Petroleum and Natural Gas	0.00101	0.00135	0.00087	0.00067	0.00097	0.00167	0.00141	0.00128
31	Utilities	0.00051	0.00085	0.00053	0.00050	0.00050	0.00083	0.00084	0.00050
32	Communication	0.00051	0.00085	0.00053	0.00050	0.00050	0.00083	0.00084	0.00050
33	Personal Services	0.00080	0.00118	0.00108	0.00073	0.00097	0.00170	0.00189	0.00070
34	Business Services	0.00105	0.00131	0.00108	0.00085	0.00088	0.00119	0.00123	0.00072
35	Computers	0.00106	0.00144	0.00103	0.00080	0.00104	0.00190	0.00213	0.00090

36	Electronic Equipment	0.00123	0.00152	0.00128	0.00095	0.00127	0.00212	0.00238	0.00102
37	Measuring and Control Equipment	0.00130	0.00151	0.00114	0.00087	0.00128	0.00230	0.00250	0.00117
38	Business Supplies	0.00128	0.00144	0.00105	0.00078	0.00106	0.00184	0.00198	0.00098
39	Shipping Containers	0.00093	0.00144	0.00106	0.00059	0.00074	0.00097	0.00110	0.00079
40	Transportation	0.00100	0.00169	0.00117	0.00066	0.00088	0.00119	0.00148	0.00092
41	Wholesale	0.00094	0.00141	0.00115	0.00075	0.00076	0.00112	0.00141	0.00083
42	Retail	0.00098	0.00134	0.00101	0.00074	0.00087	0.00125	0.00140	0.00084
43	Restaraunts, Hotels, Motels	0.00101	0.00154	0.00122	0.00082	0.00088	0.00123	0.00164	0.00080
48	Almost Nothing	0.00099	0.00167	0.00136	0.00084	0.00097	0.00106	0.00093	0.00071
	$\frac{\sigma(\sigma_K)}{K}$	**0.00023**	**0.00021**	**0.00018**	**0.00018**	**0.00022**	**0.00040**	**0.00043**	**0.00028**

Table 10-13 This table contains the cross sectional standard deviations of industry group means for the FF classification scheme. There are 48 GICS groups at this level. Categories that contained fewer than 5 stocks per group were regarded as non-functional and removed.

Cluster	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	0.0035	0.0051	0.0080	0.0073	0.0097	0.0186	0.0158	0.0040
2	0.0043	0.0051	0.0063	0.0076	0.0097	0.0193	0.0170	0.0062
3	0.0049	0.0057	0.0074	0.0074	0.0087	0.0177	0.0121	0.0069
4	0.0052	0.0047	0.0078	0.0072	0.0034	0.0178	0.0144	0.0069
5	0.0055	0.0051	0.0076	0.0068	0.0095	0.0205	0.0160	0.0043
6	0.0048	0.0055	0.0065	0.0072	0.0091	0.0160	0.0155	0.0069
7	0.0045	0.0035	0.0064	0.0080	0.0091	0.0184	0.0163	0.0051
8	0.0049	0.0055	0.0074	0.0075	0.0091	0.0164	0.0136	0.0058
9	0.0026	0.0059	0.0058	0.0071	0.0100	0.0182	0.0142	0.0066
10	0.0037	0.0053	0.0047	0.0055	0.0092	0.0164	0.0166	0.0051
11	0.0047	0.0062	0.0067	0.0072	0.0070	0.0170	0.0148	0.0054
12	0.0052	0.0048	0.0055	0.0071	0.0090	0.0152	0.0142	0.0061
13	0.0054	0.0014	0.0064	0.0046	0.0099	0.0106	0.0076	0.0058
14	0.0051	0.0049	0.0027	0.0074	0.0103	0.0139	0.0135	0.0042
15	0.0048	0.0048	0.0073	0.0068	0.0091	0.0177	0.0111	0.0065
16	0.0039	0.0059	0.0075	0.0067	0.0093	0.0122	0.0077	0.0051
17	0.0042	0.0055	0.0068	0.0067	0.0093	0.0138	0.0126	0.0057
18	0.0044	0.0041	0.0056	0.0010	0.0062	0.0134	0.0165	0.0053
19	0.0034	0.0045	0.0062	0.0067	0.0070	0.0171	0.0099	0.0053
20	0.0037	0.0057	0.0075	0.0048	0.0070	0.0090	0.0134	0.0046
21	0.0039	0.0056	0.0054	0.0068	0.0079	0.0146	0.0140	0.0051
22	0.0038	0.0041	0.0052	0.0054	0.0083	0.0164	0.0144	0.0016
23	0.0026	0.0036	0.0058	0.0060	0.0075	0.0104	0.0106	0.0038
24	0.0038	0.0022	0.0046	0.0052	0.0085	0.0166	0.0111	0.0022
25	0.0031	0.0042	0.0064	0.0062	0.0047	0.0143	0.0144	0.0044
26	0.0035	0.0016	0.0064	0.0046	0.0064	0.0126	0.0039	0.0049
27	0.0040	0.0055	0.0045	0.0064	0.0073	0.0128	0.0124	0.0053
28	0.0053	0.0046	0.0082	0.0058	0.0071	0.0169	0.0092	0.0045
29	0.0037	0.0038	0.0068	0.0047	0.0060	0.0131	0.0105	0.0043
30	0.0038	0.0027	0.0062	0.0044	0.0076	0.0128	0.0105	0.0048
31	0.0037	0.0040	0.0057	0.0068	0.0084	0.0089	0.0159	0.0026
32	0.0038	0.0047	0.0058	0.0047	0.0060	0.0112	0.0085	0.0054
33	0.0038	0.0040	0.0045	0.0044	0.0026	0.0102	0.0075	0.0054
34	0.0026	0.0054	0.0044	0.0062	0.0085	0.0113	0.0056	0.0050
35	0.0039	0.0051	0.0052	0.0064	0.0060	0.0154	0.0123	0.0063
36	0.0029	0.0051	0.0035	0.0057	0.0063	0.0105	0.0036	0.0040
37	0.0022	0.0044	0.0046	0.0048	0.0049	0.0172	0.0069	0.0042
38	0.0042	0.0043	0.0058	0.0035	0.0018	0.0076	0.0075	0.0047
39	0.0030	0.0032	0.0042	0.0032	0.0076	0.0116	0.0067	0.0049
40	0.0027	0.0043	0.0045	0.0049	0.0091	0.0052	0.0074	0.0046
41	0.0034	0.0048	0.0062	0.0047	0.0072	0.0124	0.0114	0.0029
42	0.0035	0.0046	0.0026	0.0057	0.0039	0.0041	0.0141	0.0045
43	0.0014	0.0038	0.0022	0.0054	0.0032	0.0097	0.0097	0.0028
44	0.0038	0.0034	0.0036	0.0025	0.0048	0.0086	0.0060	0.0040
45	0.0028	0.0035	0.0033	0.0047	0.0047	0.0112	0.0060	0.0025
46	0.0015	0.0042	0.0030	0.0054	0.0071	0.0097	0.0047	0.0033
47	0.0011	0.0029	0.0009	0.0040	0.0070	0.0078	0.0075	0.0063
48	0.0029	0.0037	0.0018	0.0033	0.0057	0.0058	0.0052	0.0037
49	0.0027	0.0035	0.0009	0.0029	0.0033	0.0069	0.0060	0.0033
50	0.0007	0.0029	0.0027	0.0032	0.0062	0.0079	0.0103	0.0044
Average	**0.0036**	**0.0044**	**0.0053**	**0.0056**	**0.0072**	**0.0130**	**0.0108**	**0.0048**

Table 10-14 This table contains the within cluster variances under the GSC

SIC	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	Agricultural Production Crops	0.0044	0.0035	0.0068	0.0051	0.0075	0.0115	0.0114	0.0044
2	Agriculture production livestock and animal specialties								
7	Agricultural Services		0.0007	0.0028	0.0046	0.0048	0.0072	0.0052	0.0035
10	Metal Mining	0.0034	0.0061	0.0067	0.0073	0.0082	0.0142	0.0131	0.0059
12	Coal Mining			0.0020	0.0029	0.0049	0.0057	0.0062	0.0056
13	Oil And Gas Extraction	0.0056	0.0066	0.0072	0.0072	0.0083	0.0143	0.0098	0.0054
14	Mining And Quarrying Of Nonmetallic Minerals, Except Fuels	0.0029	0.0043	0.0049	0.0050	0.0052	0.0129	0.0107	0.0035
15	Building Construction General Contractors And Operative Builders	0.0050	0.0057	0.0081	0.0078	0.0084	0.0135	0.0102	0.0042
16	Heavy Construction Other Than Building Construction Contractors	0.0041	0.0062	0.0084	0.0072	0.0098	0.0170	0.0144	0.0070
17	Construction Special Trade Contractors	0.0045	0.0060	0.0063	0.0085	0.0101	0.0153	0.0151	0.0062
20	Food And Kindred Products	0.0044	0.0047	0.0054	0.0056	0.0078	0.0116	0.0087	0.0043
21	Tobacco Products	0.0011	0.0002	0.0009	0.0008	0.0035	0.0059	0.0064	0.0034
22	Textile Mill Products	0.0058	0.0053	0.0066	0.0058	0.0068	0.0146	0.0142	0.0056
23	Apparel And Other Finished Products Made From Fabrics And Similar Materials	0.0051	0.0057	0.0084	0.0079	0.0089	0.0166	0.0123	0.0057
24	Lumber And Wood Products, Except Furniture	0.0052	0.0055	0.0068	0.0071	0.0077	0.0115	0.0106	0.0046
25	Furniture And Fixtures	0.0042	0.0048	0.0055	0.0060	0.0066	0.0120	0.0081	0.0048
26	Paper And Allied Products	0.0038	0.0038	0.0045	0.0047	0.0063	0.0111	0.0081	0.0042
27	Printing, Publishing, And Allied Industries	0.0039	0.0048	0.0051	0.0058	0.0069	0.0139	0.0093	0.0040
28	Chemicals And Allied Products	0.0053	0.0064	0.0080	0.0080	0.0105	0.0213	0.0171	0.0072
29	Petroleum Refining And Related Industries	0.0034	0.0046	0.0047	0.0039	0.0040	0.0096	0.0088	0.0046
30	Rubber And Miscellaneous Plastics Products	0.0049	0.0059	0.0065	0.0066	0.0076	0.0146	0.0138	0.0055
31	Leather And Leather Products	0.0046	0.0048	0.0062	0.0060	0.0086	0.0152	0.0105	0.0047
32	Stone, Clay, Glass, And Concrete Products	0.0050	0.0053	0.0068	0.0069	0.0067	0.0140	0.0113	0.0056
33	Primary Metal Industries	0.0049	0.0058	0.0068	0.0067	0.0074	0.0146	0.0155	0.0063
34	Fabricated Metal Products, Except Machinery And Transportation Equipment	0.0047	0.0057	0.0066	0.0065	0.0074	0.0134	0.0115	0.0053
35	Industrial And Commercial Machinery And Computer Equipment	0.0058	0.0066	0.0089	0.0083	0.0106	0.0208	0.0182	0.0069
36	Electronic And Other Electrical Equipment And Components, Except Computer Equipment	0.0061	0.0070	0.0088	0.0088	0.0113	0.0218	0.0185	0.0074

37	Transportation Equipment	0.0050	0.0054	0.0066	0.0064	0.0073	0.0139	0.0129	0.0052
38	Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks	0.0063	0.0071	0.0089	0.0084	0.0105	0.0199	0.0179	0.0070
39	Miscellaneous Manufacturing Industries	0.0064	0.0067	0.0088	0.0075	0.0100	0.0172	0.0158	0.0061
40	Railroad Transportation	0.0023	0.0025	0.0028	0.0038	0.0058	0.0096	0.0054	0.0025
41	Local And Suburban Transit And Interurban Highway Passenger Transportation					0.0070	0.0107		0.0004
42	Motor Freight Transportation And Warehousing	0.0037	0.0058	0.0066	0.0062	0.0071	0.0139	0.0117	0.0050
44	Water Transportation	0.0045	0.0052	0.0050	0.0042	0.0061	0.0121	0.0074	0.0042
45	Transportation By Air	0.0051	0.0045	0.0062	0.0071	0.0092	0.0129	0.0147	0.0070
46	Pipelines, Except Natural Gas							0.0027	0.0022
47	Transportation Services	0.0035	0.0041	0.0044	0.0054	0.0073	0.0185	0.0132	0.0054
48	Communications	0.0041	0.0046	0.0060	0.0064	0.0093	0.0181	0.0167	0.0056
49	Electric, Gas, And Sanitary Services	0.0026	0.0029	0.0028	0.0029	0.0040	0.0082	0.0064	0.0026
50	Wholesale Trade-durable Goods	0.0055	0.0063	0.0080	0.0080	0.0100	0.0179	0.0143	0.0066
51	Wholesale Trade-non-durable Goods	0.0049	0.0054	0.0063	0.0062	0.0085	0.0170	0.0131	0.0061
52	Building Materials, Hardware, Garden Supply, And Mobile Home Dealers	0.0034	0.0042	0.0051	0.0063	0.0083	0.0128	0.0082	0.0023
53	General Merchandise Stores	0.0034	0.0047	0.0058	0.0058	0.0068	0.0118	0.0102	0.0045
54	Food Stores	0.0043	0.0046	0.0050	0.0056	0.0063	0.0110	0.0096	0.0044
55	Automotive Dealers And Gasoline Service Stations	0.0029	0.0055	0.0069	0.0067	0.0083	0.0154	0.0127	0.0043
56	Apparel And Accessory Stores	0.0045	0.0058	0.0084	0.0080	0.0103	0.0177	0.0124	0.0060
57	Home Furniture, Furnishings, And Equipment Stores	0.0045	0.0067	0.0086	0.0087	0.0102	0.0178	0.0144	0.0058
58	Eating And Drinking Places	0.0052	0.0060	0.0082	0.0081	0.0099	0.0147	0.0124	0.0050
59	Miscellaneous Retail	0.0052	0.0055	0.0074	0.0078	0.0096	0.0195	0.0168	0.0064
70	Hotels, Rooming Houses, Camps, And Other Lodging Places	0.0024	0.0041	0.0068	0.0055	0.0066	0.0113	0.0095	0.0036
72	Personal Services	0.0035	0.0051	0.0041	0.0050	0.0074	0.0135	0.0111	0.0044
73	Business Services	0.0059	0.0072	0.0091	0.0089	0.0117	0.0241	0.0207	0.0071
75	Automotive Repair, Services, And Parking	0.0034	0.0042	0.0045	0.0045	0.0079	0.0156	0.0097	0.0039
76	Miscellaneous Repair Services	0.0010	0.0036	0.0060	0.0049	0.0080	0.0033		
78	Motion Pictures	0.0043	0.0061	0.0076	0.0080	0.0097	0.0170	0.0134	0.0053
79	Amusement And Recreation Services	0.0051	0.0059	0.0076	0.0087	0.0111	0.0174	0.0131	0.0055
80	Health Services	0.0054	0.0069	0.0097	0.0092	0.0107	0.0196	0.0149	0.0057

81	Legal Services								
82	Educational Services	0.0032	0.0011	0.0060	0.0076	0.0089	0.0176	0.0122	0.0053
83	Social Services				0.0020	0.0096	0.0156	0.0156	0.0045
86	Membership Organizations								
87	Engineering, Accounting, Research, Management, And Related Services	0.0060	0.0072	0.0086	0.0081	0.0100	0.0203	0.0171	0.0066
99	Nonclassifiable Establishments	0.0051	0.0055	0.0072	0.0083	0.0110	0.0218	0.0199	0.0040
	Average	**0.0043**	**0.0051**	**0.0063**	**0.0063**	**0.0081**	**0.0145**	**0.0121**	**0.0050**

Table 10-15 This table contains the within cluster variances for the SIC industry classification scheme

NAICS	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
42									
111	Crop Production	0.0044	0.0039	0.0073	0.0056	0.0079	0.0115	0.0114	0.0044
112	Animal Production								
211	Oil and Gas Extraction	0.0056	0.0065	0.0069	0.0074	0.0084	0.0147	0.0100	0.0053
212	Mining (except Oil and Gas)	0.0043	0.0064	0.0067	0.0073	0.0081	0.0160	0.0130	0.0066
213	Support Activities for Mining	0.0052	0.0060	0.0067	0.0063	0.0072	0.0102	0.0068	0.0047
221	Utilities	0.0021	0.0023	0.0020	0.0020	0.0026	0.0057	0.0047	0.0021
233		0.0051	0.0045	0.0071	0.0076	0.0087	0.0123	0.0138	
234				0.0042	0.0043	0.0108	0.0183	0.0043	
235					0.0060	0.0096	0.0146	0.0038	
236	Construction of Buildings	0.0046	0.0062	0.0077	0.0075	0.0077	0.0131	0.0096	0.0042
237	Heavy and Civil Engineering Construction	0.0039	0.0056	0.0089	0.0079	0.0091	0.0151	0.0147	0.0076
238	Specialty Trade Contractors	0.0044	0.0041	0.0056	0.0073	0.0086	0.0143	0.0139	0.0062
311	Food Manufacturing	0.0044	0.0047	0.0055	0.0056	0.0075	0.0117	0.0090	0.0044
312	Beverage and Tobacco Product Manufacturing	0.0030	0.0038	0.0035	0.0045	0.0086	0.0104	0.0076	0.0036
313	Textile Mills	0.0057	0.0057	0.0067	0.0056	0.0061	0.0144	0.0149	0.0055
314	Textile Product Mills	0.0039	0.0055	0.0057	0.0061	0.0074	0.0118	0.0093	0.0046
315	Apparel Manufacturing	0.0051	0.0055	0.0083	0.0075	0.0087	0.0167	0.0125	0.0056
316	Leather and Allied Product Manufacturing	0.0052	0.0056	0.0067	0.0065	0.0089	0.0152	0.0107	0.0049
321	Wood Product Manufacturing	0.0053	0.0056	0.0069	0.0072	0.0077	0.0115	0.0108	0.0048
322	Paper Manufacturing	0.0038	0.0038	0.0043	0.0045	0.0063	0.0113	0.0082	0.0042
323	Printing and Related Support Activities	0.0042	0.0050	0.0057	0.0070	0.0076	0.0127	0.0101	0.0039
324	Petroleum and Coal Products Manufacturing	0.0034	0.0045	0.0043	0.0036	0.0037	0.0093	0.0086	0.0046
325	Chemical Manufacturing	0.0053	0.0064	0.0080	0.0079	0.0105	0.0211	0.0171	0.0072
326	Plastics and Rubber Products Manufacturing	0.0045	0.0056	0.0064	0.0065	0.0074	0.0145	0.0142	0.0053
327	Nonmetallic Mineral Product Manufacturing	0.0050	0.0053	0.0068	0.0069	0.0067	0.0140	0.0113	0.0056
331	Primary Metal Manufacturing	0.0047	0.0056	0.0067	0.0065	0.0070	0.0139	0.0149	0.0059
332	Fabricated Metal Product Manufacturing	0.0047	0.0058	0.0065	0.0062	0.0071	0.0131	0.0110	0.0053
333	Machinery Manufacturing	0.0054	0.0063	0.0082	0.0075	0.0094	0.0183	0.0152	0.0063
334	Computer and Electronic Product Manufacturing	0.0065	0.0073	0.0093	0.0091	0.0116	0.0223	0.0196	0.0075
335	Electrical Equipment, Appliance, and Component Manufacturing	0.0050	0.0057	0.0072	0.0075	0.0097	0.0172	0.0141	0.0067
336	Transportation Equipment Manufacturing	0.0051	0.0054	0.0068	0.0063	0.0075	0.0138	0.0129	0.0053

337	Furniture and Related Product Manufacturing	0.0043	0.0050	0.0060	0.0065	0.0070	0.0118	0.0089	0.0049
339	Miscellaneous Manufacturing	0.0060	0.0069	0.0092	0.0082	0.0102	0.0187	0.0166	0.0066
421		0.0057	0.0070	0.0088	0.0084	0.0103	0.0165	0.0204	
422		0.0045	0.0048	0.0060	0.0059	0.0086	0.0185	0.0143	0.0042
423	Merchant Wholesalers, Durable Goods	0.0051	0.0059	0.0073	0.0077	0.0098	0.0180	0.0139	0.0067
424	Merchant Wholesalers, Nondurable Goods	0.0043	0.0046	0.0057	0.0061	0.0078	0.0151	0.0125	0.0058
425	Wholesale Electronic Markets and Agents and Brokers				0.0004	0.0044	0.0137	0.0069	0.0034
441	Motor Vehicle and Parts Dealers	0.0029	0.0053	0.0068	0.0058	0.0082	0.0152	0.0123	0.0041
442	Furniture and Home Furnishings Stores	0.0033	0.0035	0.0063	0.0064	0.0107	0.0141	0.0077	0.0046
443	Electronics and Appliance Stores	0.0052	0.0072	0.0088	0.0087	0.0090	0.0173	0.0150	0.0049
444	Building Material and Garden Equipment and Supplies Dealers	0.0034	0.0042	0.0051	0.0063	0.0087	0.0128	0.0058	0.0014
445	Food and Beverage Stores	0.0043	0.0046	0.0050	0.0054	0.0060	0.0100	0.0095	0.0039
446	Health and Personal Care Stores	0.0040	0.0033	0.0044	0.0064	0.0088	0.0179	0.0146	0.0059
447	Gasoline Stations		0.0011	0.0024	0.0053	0.0064	0.0109	0.0078	0.0023
448	Clothing and Clothing Accessories Stores	0.0045	0.0059	0.0084	0.0082	0.0100	0.0174	0.0127	0.0060
451	Sporting Goods, Hobby, Book, and Music Stores	0.0044	0.0049	0.0063	0.0072	0.0097	0.0165	0.0150	0.0052
452	General Merchandise Stores	0.0034	0.0047	0.0058	0.0058	0.0068	0.0118	0.0102	0.0045
453	Miscellaneous Store Retailers	0.0016	0.0036	0.0075	0.0066	0.0086	0.0176	0.0125	0.0058
454	Nonstore Retailers	0.0041	0.0063	0.0075	0.0084	0.0099	0.0205	0.0189	0.0068
481	Air Transportation	0.0047	0.0042	0.0054	0.0068	0.0094	0.0119	0.0137	0.0069
482	Rail Transportation	0.0023	0.0025	0.0028	0.0038	0.0058	0.0096	0.0054	0.0025
483	Water Transportation	0.0040	0.0052	0.0045	0.0037	0.0055	0.0106	0.0063	0.0038
484	Truck Transportation	0.0035	0.0058	0.0062	0.0062	0.0070	0.0129	0.0119	0.0051
485	Transit and Ground Passenger Transportation					0.0047	0.0033		
486	Pipeline Transportation	0.0036	0.0032	0.0040	0.0038	0.0038	0.0095	0.0101	0.0041
487	Scenic and Sightseeing Transportation					0.0014	0.0053		
488	Support Activities for Transportation	0.0039	0.0041	0.0052	0.0060	0.0075	0.0162	0.0109	0.0055
492	Couriers and Messengers	0.0035	0.0038	0.0064	0.0061	0.0061	0.0166	0.0092	0.0037
493	Warehousing and Storage	0.0017	0.0014	0.0020	0.0031	0.0044	0.0052	0.0038	0.0008
511	Publishing Industries (except Internet)	0.0050	0.0063	0.0082	0.0085	0.0116	0.0243	0.0204	0.0070
512	Motion Picture and Sound Recording Industries	0.0049	0.0061	0.0074	0.0078	0.0092	0.0170	0.0143	0.0048
513		0.0041	0.0045	0.0068	0.0060	0.0096	0.0177	0.0229	0.0020
514		0.0039	0.0054	0.0076	0.0082	0.0111	0.0225	0.0291	
515	Broadcasting (except Internet)	0.0028	0.0034	0.0035	0.0061	0.0083	0.0147	0.0107	0.0039

516	Internet Publishing and Broadcasting								0.0024
517	Telecommunications	0.0035	0.0045	0.0053	0.0063	0.0088	0.0189	0.0182	0.0064
518	Internet Service Providers, Web Search Portals, and Data Processing Services	0.0020	0.0037	0.0073	0.0077	0.0114	0.0251	0.0206	0.0070
519	Other Information Services								0.0033
522	Credit Intermediation and Related Activities		0.0018	0.0058	0.0053	0.0072	0.0166	0.0128	0.0058
523	Securities, Commodity Contracts, and Other Financial Investments and Related Activities								
524	Insurance Carriers and Related Activities								
525	Funds, Trusts, and Other Financial Vehicles								
532	Rental and Leasing Services	0.0048	0.0053	0.0062	0.0058	0.0084	0.0169	0.0150	0.0050
541	Professional, Scientific, and Technical Services	0.0061	0.0075	0.0084	0.0083	0.0108	0.0223	0.0192	0.0071
561	Administrative and Support Services	0.0047	0.0061	0.0078	0.0078	0.0100	0.0203	0.0166	0.0060
562	Waste Management and Remediation Services	0.0057	0.0062	0.0083	0.0076	0.0106	0.0176	0.0142	0.0048
611	Educational Services	0.0042	0.0019	0.0068	0.0082	0.0097	0.0184	0.0125	0.0052
621	Ambulatory Health Care Services	0.0050	0.0068	0.0097	0.0098	0.0113	0.0200	0.0153	0.0063
622	Hospitals	0.0028	0.0046	0.0083	0.0068	0.0082	0.0189	0.0123	0.0031
623	Nursing and Residential Care Facilities	0.0020	0.0034	0.0024	0.0039	0.0081	0.0172	0.0150	0.0045
624	Social Assistance					0.0082	0.0092	0.0067	0.0046
711	Performing Arts, Spectator Sports, and Related Industries	0.0008	0.0024		0.0007	0.0091	0.0168	0.0134	0.0031
713	Amusement, Gambling, and Recreation Industries	0.0014	0.0036	0.0033	0.0086	0.0129	0.0185	0.0132	0.0054
721	Accommodation	0.0039	0.0052	0.0082	0.0071	0.0079	0.0122	0.0101	0.0046
722	Food Services and Drinking Places	0.0052	0.0059	0.0081	0.0080	0.0099	0.0149	0.0124	0.0051
811	Repair and Maintenance	0.0011	0.0053	0.0083	0.0061	0.0088	0.0152	0.0117	0.0029
812	Personal and Laundry Services	0.0034	0.0038	0.0048	0.0052	0.0076	0.0172	0.0110	0.0044
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations								
	Average	**0.0041**	**0.0049**	**0.0063**	**0.0064**	**0.0082**	**0.0149**	**0.0124**	**0.0048**

Table 10-16 This table contains the within cluster variances for the NAICS industry classification scheme

GICS	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
101010	Energy Equipment & Services	0.0056	0.0058	0.0072	0.0071	0.0083	0.0135	0.0101	0.0051
101020	Oil, Gas & Consumable Fuels	0.0050	0.0059	0.0061	0.0066	0.0075	0.0140	0.0104	0.0055
151010	Chemicals	0.0044	0.0049	0.0059	0.0054	0.0064	0.0134	0.0098	0.0049
151020	Construction Materials	0.0041	0.0042	0.0052	0.0048	0.0048	0.0111	0.0081	0.0045
151030	Containers & Packaging	0.0041	0.0053	0.0053	0.0065	0.0068	0.0126	0.0105	0.0047
151040	Metals & Mining	0.0048	0.0064	0.0071	0.0072	0.0078	0.0144	0.0144	0.0063
151050	Paper & Forest Products	0.0036	0.0043	0.0043	0.0040	0.0051	0.0095	0.0079	0.0038
201010	Aerospace & Defense	0.0050	0.0057	0.0072	0.0066	0.0082	0.0145	0.0133	0.0051
201020	Building Products	0.0048	0.0058	0.0075	0.0075	0.0073	0.0138	0.0135	0.0053
201030	Construction & Engineering	0.0047	0.0067	0.0084	0.0068	0.0084	0.0127	0.0125	0.0060
201040	Electrical Equipment	0.0052	0.0063	0.0076	0.0079	0.0103	0.0181	0.0156	0.0069
201050	Industrial Conglomerates	0.0029	0.0027	0.0032	0.0033	0.0034	0.0065	0.0057	0.0036
201060	Machinery	0.0049	0.0056	0.0071	0.0063	0.0075	0.0139	0.0104	0.0051
201070	Trading Companies & Distributors	0.0040	0.0038	0.0040	0.0044	0.0065	0.0129	0.0107	0.0055
202010	Commercial Services & Supplies	0.0053	0.0060	0.0074	0.0075	0.0097	0.0181	0.0147	0.0056
203010	Air Freight & Logistics	0.0045	0.0049	0.0062	0.0067	0.0077	0.0156	0.0101	0.0052
203020	Airlines	0.0042	0.0043	0.0055	0.0069	0.0092	0.0116	0.0137	0.0068
203030	Marine	0.0039	0.0058	0.0046	0.0037	0.0037	0.0066	0.0037	0.0023
203040	Road & Rail	0.0036	0.0049	0.0055	0.0058	0.0071	0.0135	0.0106	0.0045
203050	Transportation Infrastructure	0.0039	0.0046	0.0031	0.0032	0.0052	0.0092	0.0133	0.0016
251010	Auto Components	0.0049	0.0051	0.0071	0.0067	0.0076	0.0151	0.0146	0.0058
251020	Automobiles	0.0036	0.0045	0.0065	0.0057	0.0052	0.0097	0.0113	0.0041
252010	Household Durables	0.0057	0.0061	0.0071	0.0073	0.0082	0.0137	0.0108	0.0053
252020	Leisure Equipment & Products	0.0057	0.0062	0.0084	0.0069	0.0098	0.0169	0.0148	0.0055
252030	Textiles, Apparel & Luxury Goods	0.0056	0.0060	0.0080	0.0074	0.0084	0.0162	0.0124	0.0055
253010	Hotels, Restaurants & Leisure	0.0053	0.0062	0.0083	0.0082	0.0102	0.0154	0.0126	0.0054
253020	Diversified Consumer Services	0.0035	0.0054	0.0047	0.0053	0.0080	0.0161	0.0121	0.0049
254010	Media	0.0045	0.0054	0.0064	0.0069	0.0090	0.0173	0.0135	0.0046
255010	Distributors	0.0057	0.0061	0.0084	0.0087	0.0098	0.0181	0.0154	0.0061
255020	Internet & Catalog Retail	0.0053	0.0064	0.0061	0.0072	0.0100	0.0225	0.0200	0.0070
255030	Multiline Retail	0.0033	0.0044	0.0056	0.0061	0.0073	0.0121	0.0105	0.0047
255040	Specialty Retail	0.0052	0.0066	0.0085	0.0083	0.0101	0.0174	0.0140	0.0058
301010	Food & Staples Retailing	0.0042	0.0044	0.0050	0.0054	0.0070	0.0128	0.0106	0.0047

302010	Beverages	0.0031	0.0033	0.0035	0.0040	0.0078	0.0094	0.0057	0.0033
302020	Food Products	0.0046	0.0047	0.0058	0.0058	0.0078	0.0120	0.0099	0.0046
302030	Tobacco	0.0027	0.0025	0.0040	0.0032	0.0039	0.0080	0.0077	0.0037
303010	Household Products	0.0047	0.0040	0.0036	0.0041	0.0064	0.0116	0.0065	0.0036
303020	Personal Products	0.0052	0.0075	0.0091	0.0082	0.0109	0.0170	0.0150	0.0070
351010	Health Care Equipment & Supplies	0.0066	0.0074	0.0093	0.0088	0.0107	0.0197	0.0175	0.0069
351020	Health Care Providers & Services	0.0054	0.0064	0.0087	0.0088	0.0105	0.0205	0.0154	0.0057
351030	Health Care Technology		0.0052	0.0058	0.0082	0.0126	0.0224	0.0165	0.0068
352010	Biotechnology	0.0066	0.0074	0.0106	0.0091	0.0123	0.0222	0.0185	0.0080
352020	Pharmaceuticals	0.0050	0.0060	0.0081	0.0077	0.0110	0.0210	0.0180	0.0074
352030	Life Sciences Tools & Services	0.0041	0.0059	0.0073	0.0072	0.0083	0.0209	0.0161	0.0068
401010	Commercial Banks				0.0011	0.0036	0.0016		
402010	Diversified Financial Services	0.0017	0.0035	0.0053	0.0051	0.0078	0.0132	0.0147	0.0035
402020	Consumer Finance		0.0006	0.0054	0.0055	0.0067	0.0150	0.0094	0.0045
402030	Capital Markets								
403010	Insurance	0.0011	0.0017	0.0023	0.0034	0.0049	0.0094	0.0033	0.0018
404010	Real Estate -- Discontinued effective 04/28/2006					0.0047	0.0061		
404020	Real Estate Investment Trusts (REITs)	0.0020	0.0019	0.0030	0.0015	0.0017	0.0027	0.0016	0.0012
404030	Real Estate Management & Development				0.0016	0.0044	0.0047	0.0094	0.0060
451010	Internet Software & Services	0.0042	0.0058	0.0081	0.0098	0.0126	0.0277	0.0240	0.0081
451020	IT Services	0.0062	0.0068	0.0077	0.0075	0.0106	0.0224	0.0167	0.0062
451030	Software	0.0064	0.0081	0.0104	0.0096	0.0126	0.0243	0.0208	0.0074
452010	Communications Equipment	0.0066	0.0072	0.0092	0.0092	0.0117	0.0227	0.0201	0.0074
452020	Computers & Peripherals	0.0066	0.0075	0.0100	0.0102	0.0122	0.0218	0.0205	0.0074
452030	Electronic Equipment, Instruments & Components	0.0063	0.0070	0.0089	0.0086	0.0107	0.0210	0.0173	0.0070
452040	Office Electronics	0.0025	0.0029	0.0043	0.0055	0.0099	0.0100	0.0087	0.0014
452050	Semiconductor Equipment & Products -- Discontinued effective 04/30/2003.	0.0055	0.0071	0.0087	0.0080	0.0109	0.0188	0.0197	
453010	Semiconductors & Semiconductor Equipment	0.0056	0.0070	0.0096	0.0090	0.0113	0.0223	0.0172	0.0071
501010	Diversified Telecommunication Services	0.0038	0.0047	0.0059	0.0063	0.0094	0.0196	0.0182	0.0060
501020	Wireless Telecommunication Services	0.0015	0.0039	0.0069	0.0060	0.0092	0.0195	0.0180	0.0061
551010	Electric Utilities	0.0017	0.0018	0.0015	0.0014	0.0022	0.0051	0.0048	0.0015
551020	Gas Utilities	0.0023	0.0026	0.0022	0.0024	0.0022	0.0039	0.0030	0.0015
551020	Gas Utilities	0.0023	0.0026	0.0022	0.0024	0.0022	0.0039	0.0030	0.0015
551030	Multi-Utilities	0.0017	0.0018	0.0016	0.0013	0.0014	0.0031	0.0037	0.0012
551040	Water Utilities	0.0025	0.0028	0.0024	0.0022	0.0028	0.0064	0.0034	0.0028

551050	Independent Power Producers & Energy Traders	0.0019	0.0032	0.0026	0.0033	0.0064	0.0135	0.0094	0.0051
	Average	**0.0043**	**0.0051**	**0.0062**	**0.0061**	**0.0078**	**0.0143**	**0.0124**	**0.0050**

Table 10-17 This table contains the within cluster variances for the GICS industry classification scheme

FF	Description	1983-1985	1986-1988	1989-1991	1992-1994	1995-1997	1998-2000	2001-2003	2004-2006
1	Agriculture	0.0049	0.0039	0.0066	0.0055	0.0076	0.0113	0.0111	0.0048
2	Food Products	0.0043	0.0047	0.0057	0.0057	0.0076	0.0117	0.0094	0.0047
3	Candy & Soda	0.0019	0.0026	0.0019	0.0039	0.0076	0.0110	0.0053	0.0029
4	Beer & Liquor	0.0048	0.0041	0.0038	0.0040	0.0080	0.0103	0.0059	0.0028
5	Tobacco Products	0.0011	0.0002	0.0009	0.0008	0.0035	0.0059	0.0064	0.0034
6	Recreation	0.0067	0.0066	0.0089	0.0086	0.0104	0.0184	0.0160	0.0066
7	Entertainment	0.0051	0.0063	0.0078	0.0087	0.0106	0.0177	0.0137	0.0056
8	Printing and Publishing	0.0034	0.0047	0.0046	0.0052	0.0064	0.0136	0.0084	0.0040
9	Consumer Goods	0.0049	0.0057	0.0069	0.0073	0.0085	0.0149	0.0113	0.0055
10	Apparel	0.0053	0.0059	0.0080	0.0076	0.0091	0.0164	0.0118	0.0054
11	Healthcare	0.0054	0.0069	0.0097	0.0092	0.0107	0.0196	0.0149	0.0057
12	Medical Equipment	0.0066	0.0076	0.0097	0.0089	0.0107	0.0195	0.0180	0.0072
13	Pharmaceutical Products	0.0061	0.0071	0.0093	0.0088	0.0117	0.0224	0.0188	0.0077
14	Chemicals	0.0044	0.0049	0.0056	0.0053	0.0062	0.0130	0.0096	0.0051
15	Rubber and Plastic Products	0.0050	0.0059	0.0062	0.0066	0.0075	0.0149	0.0144	0.0057
16	Textiles	0.0058	0.0053	0.0066	0.0058	0.0068	0.0146	0.0142	0.0056
17	Construction Materials	0.0050	0.0057	0.0068	0.0068	0.0071	0.0135	0.0107	0.0052
18	Construction	0.0054	0.0069	0.0089	0.0084	0.0094	0.0154	0.0135	0.0061
19	Steel Works Etc	0.0049	0.0058	0.0068	0.0067	0.0074	0.0146	0.0155	0.0063
20	Fabricated Products	0.0050	0.0058	0.0069	0.0067	0.0085	0.0144	0.0132	0.0062
21	Machinery	0.0055	0.0062	0.0081	0.0073	0.0092	0.0173	0.0139	0.0061
22	Electrical Equipment	0.0051	0.0058	0.0076	0.0076	0.0097	0.0170	0.0143	0.0066
23	Automobiles and Trucks	0.0049	0.0053	0.0070	0.0068	0.0070	0.0143	0.0141	0.0054
24	Aircraft	0.0045	0.0049	0.0044	0.0043	0.0067	0.0121	0.0099	0.0039
25	Shipbuilding, Railroad Equipment	0.0034	0.0039	0.0057	0.0049	0.0058	0.0118	0.0077	0.0052
26	Defense	0.0035	0.0039	0.0043	0.0062	0.0067	0.0085	0.0090	0.0051
27	Precious Metals	0.0030	0.0058	0.0064	0.0070	0.0077	0.0141	0.0117	0.0053
28	Non-Metallic and Industrial Metal Mining	0.0031	0.0049	0.0055	0.0056	0.0064	0.0140	0.0112	0.0053
29	Coal			0.0020	0.0029	0.0049	0.0057	0.0062	0.0056
30	Petroleum and Natural Gas	0.0051	0.0062	0.0067	0.0066	0.0078	0.0138	0.0099	0.0053
31	Utilities	0.0022	0.0024	0.0019	0.0019	0.0023	0.0052	0.0045	0.0020
32	Communication	0.0022	0.0024	0.0019	0.0019	0.0023	0.0052	0.0045	0.0020
33	Personal Services	0.0041	0.0046	0.0060	0.0064	0.0093	0.0181	0.0167	0.0056
34	Business Services	0.0050	0.0056	0.0067	0.0066	0.0093	0.0166	0.0133	0.0050

35	Computers	0.0059	0.0072	0.0090	0.0088	0.0112	0.0235	0.0203	0.0069
36	Electronic Equipment	0.0064	0.0070	0.0096	0.0094	0.0121	0.0240	0.0209	0.0076
37	Measuring and Control Equipment	0.0062	0.0072	0.0091	0.0089	0.0115	0.0223	0.0191	0.0074
38	Business Supplies	0.0061	0.0069	0.0082	0.0078	0.0103	0.0207	0.0177	0.0067
39	Shipping Containers	0.0041	0.0041	0.0048	0.0050	0.0068	0.0118	0.0083	0.0045
40	Transportation	0.0026	0.0043	0.0039	0.0051	0.0059	0.0118	0.0115	0.0036
41	Wholesale	0.0049	0.0054	0.0062	0.0064	0.0081	0.0151	0.0126	0.0057
42	Retail	0.0054	0.0061	0.0076	0.0076	0.0096	0.0178	0.0140	0.0065
43	Restaraunts, Hotels, Motels	0.0049	0.0059	0.0075	0.0076	0.0093	0.0175	0.0142	0.0059
48	Almost Nothing	0.0050	0.0059	0.0083	0.0080	0.0095	0.0142	0.0122	0.0049
	Average	**0.0054**	**0.0064**	**0.0082**	**0.0074**	**0.0100**	**0.0174**	**0.0150**	**0.0055**

Table 10-18 This table contains the within cluster variances for the FF industry classification scheme