9th International Conference on Computing in Economics and Finance, July 11-13, 2003, U of Washington, Seattle

Hedge Fund Classification using K-means Clustering Method

Nandita Das[1]

Abstract

Hedge fund databases vary as to the type of funds to include and in their classification scheme. Investment strategy and/or investment style are the basis for classification. Considerable variation is observed in the definitions, return calculation methodologies, and assumptions. There exists a myriad of classifications, some overlapping and some mutually exclusive. There is a need for an 'alternative approach' to hedge fund classification given the lack of 'pure' hedge fund types. The hedge fund literature shows an almost complete reliance on the existing hedge fund classifications. This means that research on hedge fund performance may produce different results based on the chosen database and the results are difficult to compare, as there are many different ways to classify any hedge fund. The varied classification of hedge funds probably attributes to the disparity in the numbers produced between different organizations measuring hedge fund performance. Asset class, region of investment, the trading strategy used, and the liquidity of the investment strategy can be the basis of hedge fund classification. This study uses cluster analysis approach to classify hedge funds. The classification is based on asset class, size of the hedge fund, incentive fee, risk-level, and liquidity of hedge funds. Nonhierarchical clustering method is used for the classification. The result is compared with the existing classification of US and NON-US hedge funds of ZCM/Hedge database.

I. Introduction

Hedge funds, as an alternative investment vehicle, have enjoyed healthy growth in recent years and continue to increase in popularity. High net worth individuals have dominated the hedge fund industry for a long time. Increasing numbers of institutions are allocating a small portion of assets to alternative investments owing to the long-term success of some hedge funds. Hedge funds became popular for their philosophy of trying to outperform the overall market through individual stock and security selection and by taking market neutral positions in an effort to protect financial capital in times of market volatility. Today, the term 'Hedge Fund' is used to describe a wide range of investment vehicles that can vary substantially in terms of size, strategy, and organizational structures. Work has been done on the benefits of adding hedge funds to the traditional investment portfolio, the performance characteristics of hedge funds and the market impact of hedge funds.

Hedge funds provide very limited information to the investors, mainly periodic (monthly, quarterly, or annual) returns. Sources of data for the industry are the hedge-fund database providers. These databases provide information drawn from the fund-offering documents; such as contractual provisions, descriptions of investments, styles of investment and the periodic return. The contractual provisions include fee structure,

[1] *Nandita Das is a Ph.D. Candidate in Business and Economics at Lehigh University.*
Contact Information: 37 RBC, 621 Taylor Street, Bethlehem, PA 18015. e-mail: nad5@lehigh.edu

The author is thankful to The Foundation for Managed Derivatives Research for providing a research grant, to Mr. Richard E. Oberuc of LaPorte for providing the ZCM/Hedge database and to Lehigh University for the Warren-York Fellowship for this research project. The author is thankful to Dr. D.L. Muething and Dr. L.W. Taylor for their guidance of this work.

minimum investment size, withdrawal provisions, etc. Four main hedge fund databases are used in academics and industry. There is neither legal definition for hedge funds, nor any industry standard for their classification. The databases vary as to the type of funds to include and in their classification scheme. There appears to be a myriad of classifications in existence.

Classifications are based on investment strategy and/or investment styles. Investment strategy consists of the tools and techniques like leverage, short-selling, hedging and arbitrage used by the manager. Investment style is a broader classification based on the strategies used by the managers to invest in different markets using various investment instruments. Investment instruments consist of stocks, bonds and all types of financial derivatives. A ‘market’ refers to the geographic location or an asset class. An asset class could include the stock market or the futures market. Geographic location could include, for example, emerging markets. Among these various classifications, there are some core strategies followed by the hedge fund managers. There is a need for a unified approach to the classification of hedge funds.

In this paper, an attempt is made to classify hedge funds with a unified approach that could be used for all the different databases. This study uses cluster analysis approach to classify hedge funds. The hedge funds are classified based on the asset classes they invest in, the incentive fee, the risk, liquidity of the investment strategy and the size. The results are compared with the existing classification in the ZCM/Hedge database for the US and the Non-US funds.

The paper proceeds as follows: Section II describes the Hedge Fund industry; Section III briefly describes the various classification scheme of the database providers; Section IV discusses the need for an alternative classification scheme; Section V describes the cluster analysis techniques; Section VI describes the methodology and data organization; Section VII discusses the results and Section VIII concludes.

II. Hedge Fund Industry

In finance industry terminology, the meaning of hedge is the process of protecting oneself against unfavorable changes in prices. The term ‘hedge fund’ is not defined or separately addressed in any securities or commodity laws. The term has undergone a considerable amount of mutation to represent what it means today compared to what it meant when it first originated in 1949. In 1949, A.W. Jones introduced the concept hedge fund. He combined a leveraged long stock position with a portfolio of short stocks in an investment fund with an incentive fee structure. Hedge fund investment practices and strategies have evolved and expanded since then. Some of today’s hedge funds satisfy all criteria of Jones’ fund; namely long/short positions and incentive-based fees. With no legal definition of a hedge fund, any fund that satisfies two criteria of Jones’ fund is identified as a hedge fund. Some hedge funds do not hedge at all.

While many hedge fund characteristics have changed significantly, many fundamental features have remained the same. Moreover, hedge funds are no longer unique to the U.S. markets, but exist in many areas around the world. In the United States, they normally offer their shares in private placements and have less than 100 high net-worth investors in order to make use of exemptions provided under the Securities Act of 1933, the Securities Exchange Act of 1934, and the Investment Company Act of l940.

In the short history of fifty years, interest in hedge funds and their performance has waxed and waned. In recent years, however, hedge funds have enjoyed healthy growth and appear to have increased in popularity. In particular, the bull market of the late 1980s created more high-net-worth investors. These investors, looking for enhanced returns, started to invest in hedge funds. The renewed interest in hedge funds that began in the late 1980s has not vanished. In 1990, there were about 600 hedge funds worldwide with assets of approximately $38 billion. According to industry publications, at the end of 1998, despite the publicized collapse of Long Term Capital Management (LTCM), there were some 3,300 hedge funds with assets of approximately US$375 billion. The near failure of LTCM in 1998 does not appear to have slowed down the growth of and interest in hedge funds. The LTCM debacle has rightly led to more caution from regulatory authorities and investor interest groups.

All estimates suggest that the hedge fund industry has experienced tremendous growth since mid 1980s, measured either by the number of funds or by assets under management. Additional investments in the hedge fund industry in years 2000 and 2001 were US$40 billion and US$80 billion respectively, and the total industry size today is about US$500 billion.

Hedge funds invest in a variety of liquid assets just like mutual funds, but are quite different from mutual funds. For example, under current federal law, hedge funds do not have any management limitations. There are virtually no limits on the composition of the portfolios and no mandatory disclosure of information about the holdings and performance. Das et al. (2002a) provides an overview of the hedge fund industry.

III. Database Providers and Classification

Four primary databases are popular among researchers and in the investment industry. Providers of these databases offer different services to the industry. Zurich Capital Markets (ZCM/Hedge) database (formerly, MAR/hedge) provides a comprehensive coverage of global hedge funds. Hedge Fund Research (HFR) database contains more equity-based hedge funds. TASS is the information and research subsidiary of Credit Suisse First Boston Tremont Advisers.

Various database providers classify hedge funds, but in different ways. All the four databases have their own indices based on the categories in the database. The index composition is also different for different databases. Hedge fund categories are based on the self-reported style classifications of hedge fund managers that are listed in a particular database. None of the database provides information on the complete hedge fund universe. The databases differ in the definition of the 'hedge fund'. For example, TASS is the only database that includes managed futures fund. Unlike hedge funds, managed futures funds limit their activities to futures market.

The following observations can be made about the performance data for various databases.

- A major limitation of most hedge fund databases is that they typically have data only on funds still in existence or that are new and growing.
- Most hedge fund indices do not include performance of closed funds.
- Only those funds that choose to report are included in the database. Not much can be done with this issue due to the industry structure. ZCM/Hedge and TASS have historical performances of all funds that are included in their database. Historical performances are not included (no backfiling) in index construction, but are available for fund analysis.
- HFR, ZCM/Hedge, and VanHedge have all inclusive selection criteria; they include all funds in their database that classify them as hedge funds. TASS has its own selection criteria.
- The classification method varies across different databases making it difficult to compare.

Hedge fund managers employ a diverse array of strategies. The database providers classify hedge funds based on the voluntary information that they collect from the hedge fund managers. Style definitions and the number of categories of hedge funds differ among the database providers. The classification of hedge funds by various database providers is briefly described here.

A. ZCM/Hedge Classification

The ZCM/Hedge database classifies hedge funds into four general classes and ten broad categories of investment styles, as reported by the managers of the hedge fund. The classes are 'onshore' hedge fund (HF-US), 'offshore' hedge fund (HF-NON), 'onshore' fund-of-funds (FOF-US), and 'offshore' fund-of-funds (FOF-NON). Some of the categories have further sub-classifications. ZCM/Hedge database categories are shown in Figure I.

B. HFR Classification

Hedge Fund Research (HFR) has twenty-six categories of hedge funds. Some of these categories are merely a type of financial instrument or a geographic area for investment. This classification can be reorganized into eleven categories as shown in Figure II. Some of the categories have further sub-classifications.

C. TASS Classification

TASS is the information and research subsidiary of Credit Suisse First Boston Tremont Advisers. It has nine categories of hedge funds, classified based on the investment styles of hedge fund managers. Figure III shows the classification of TASS database.

D. VanHedge Classification

VanHedge maintains an extensive database of hedge funds. It provides consultancy and detailed generic performance data on hedge fund styles. VanHedge database can be organized into thirteen categories and five subcategories. Figure IV shows the classification of VanHedge database.

Figure I. ZCM/Hedge Classification of Hedge Funds.

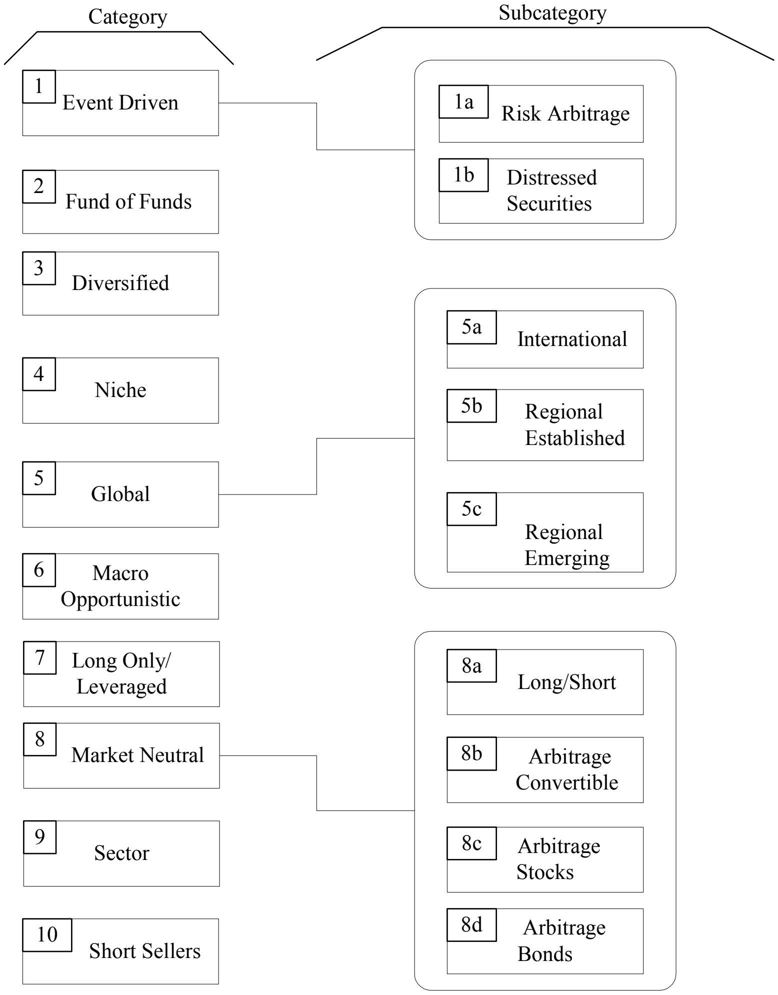


Figure II. HFR Classification of Hedge Funds.

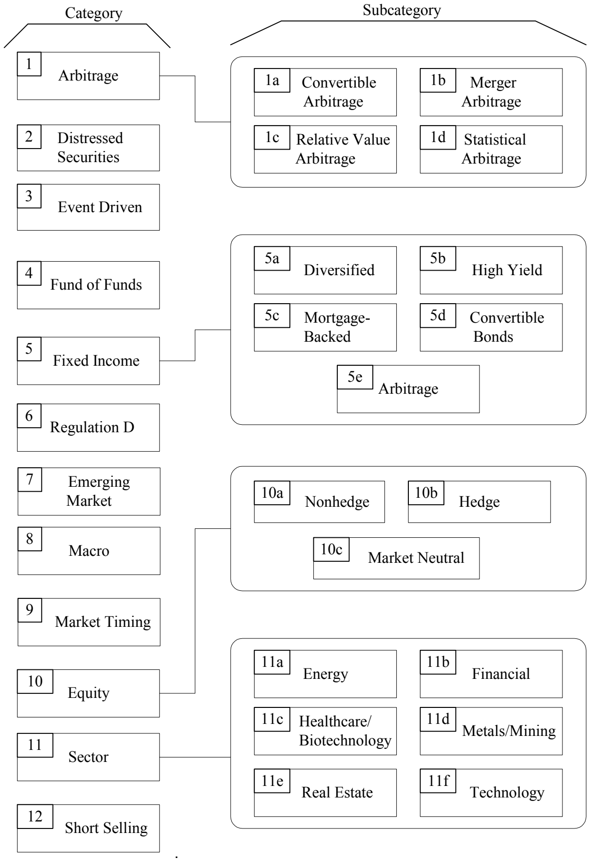


Figure III. TASS Classification of Alternative Investments.

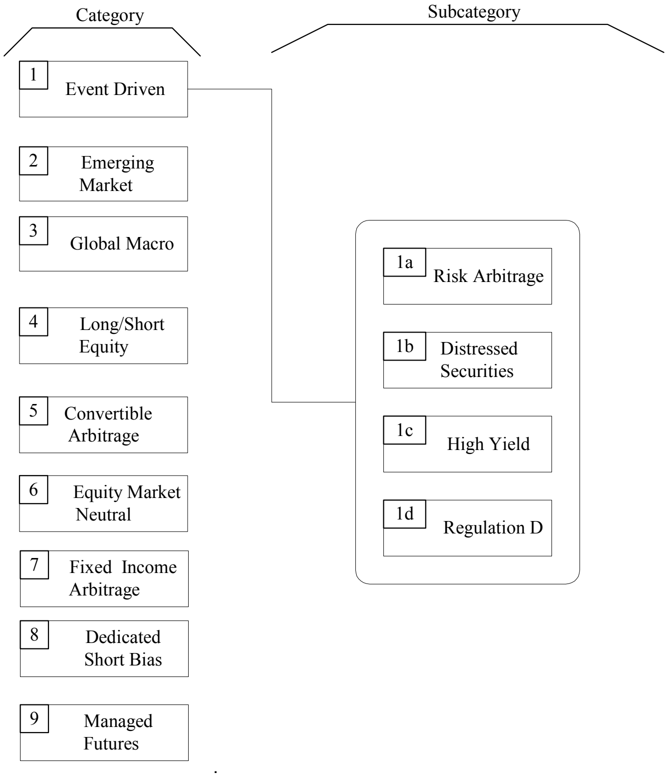
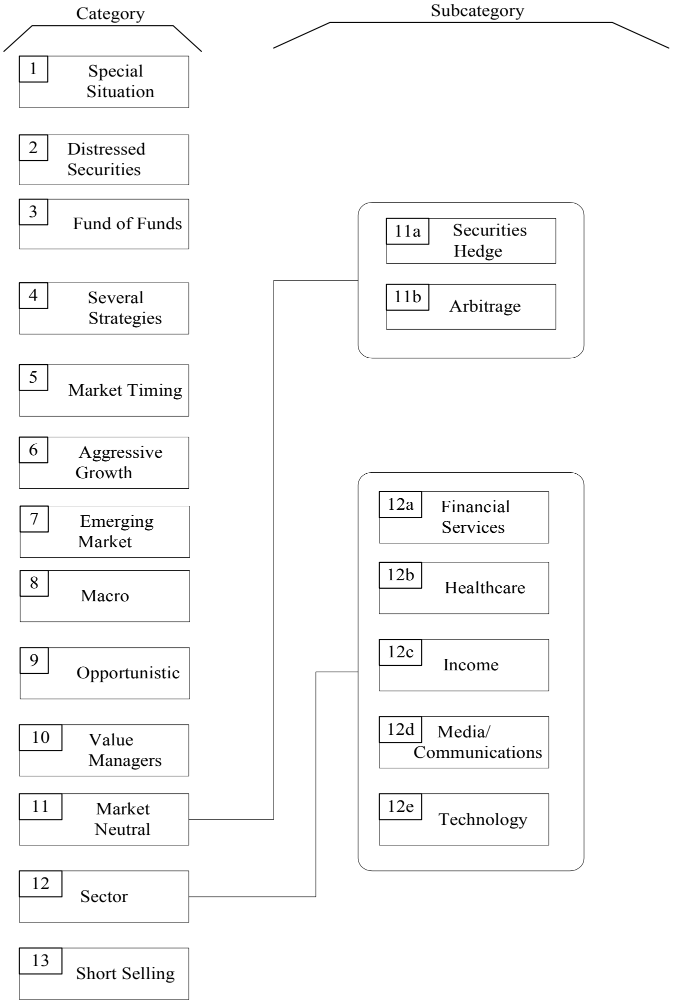


Figure IV. VanHedge Classification of Hedge Funds.



IV. Alternative Classification Requirement

There exists a lot of variation in the definitions, calculation methodologies, assumptions, and data employed by the different managers and databases. It is necessary to benchmark hedge fund manager practices relative to their peers as hedge funds follow diverse strategies. 'One size fits all,' is not true for the various classification schemes and multiple peer groups may be relevant depending on the strategies employed by the manager. It is important to clearly identify a peer for the various hedge fund strategies. This may not be an easy task since hedge fund managers refrain from disclosure.

Hedge funds are primarily distinguished by their use of short-selling, leverage, derivatives and portfolio concentration. Hedge fund manager refrains from disclosure for two reasons. They are not permitted by regulation to advertise to the public. Secondly, the proprietary nature of the trades may result in herding. Hedge fund managers profit by identifying arbitrage opportunities. These opportunities are based on very slim price differentials, but the manager hopes to profit by properly timing his trade and through portfolio concentration.

There is a need for an 'alternative approach' to hedge fund classifications given the lack of 'pure' hedge fund types that exist in the industry. The hedge fund literature shows an almost complete reliance on the existing hedge fund classifications. Performance comparison of various hedge funds with the existing hedge fund indices return data is not appropriate as a particular hedge fund could be classified in two or more classes depending on the database. Table I compares the classifications of ZCM/Hedge, HFR, TASS and VanHedge databases.

Table I. Comparison of ZCM/Hedge, HFR, TASS, and VanHedge Classifications.

Item	ZCM/Hedge Strategies	HFR Strategies	TASS Strategies	VanHedge Strategies
1a	Event Driven: Risk Arbitrage	Event Driven Merger Arbitrage	Event Driven: Risk Arbitrage	Special Situation
1b	Event Driven: Distressed Securities	Distressed Securities	Event Driven: Distressed Securities	Distressed Securities
2	Fund of Funds	Fund of Funds	None	Fund of Funds
3	Diversified	Fixed Income Diversified	None	Several Strategies
4	Niche	Fixed Income: High Yield Regulation D	Event Driven: Regulation D Event Driven: High Yield	None
5	Global	Emerging Markets	Emerging Markets	Emerging Markets
6	Macro Opportunistic	Macro Market Timing Relative Value Arbitrage Statistical Arbitrage	Global Macro	Opportunistic Value Managers
7	Long Only / Leveraged	Equity Nonhedge	None	None
8a	Market Neutral: Long/Short	Equity Hedge	Long/Short Equity	Market Neutral: Securities Hedge
8b	Market Neutral: Arbitrage Convertible	Convertible Arbitrage	Convertible Arbitrage	Market Neutral: Arbitrage
8c	Market Neutral: Arbitrage Stock	Equity Market Neutral	Equity Market Neutral	Market Neutral: Arbitrage
8d	Market Neutral: Arbitrage Bond	Fixed Income Arbitrage	Fixed Income Arbitrage	Market Neutral: Arbitrage
9	Sector	Sector: Energy Sector: Financial Sector: Health Care/ Biotechnology Sector: Metals/Mining Sector: Real Estate Sector: Technology	None	Financial Services Health Care Income Media/ Communications Technology
10	Short Selling	Short Selling	Dedicated Short Bias	Short Selling

It appears from Table I that research on hedge fund performance may produce different results, based on the database used. There seems to be no common comparison basis for the existing literature on hedge funds. The disparity that is observed in the numbers produced between different organizations measuring hedge fund performance could be attributed to the varied classification of hedge funds. Goldman Sachs & Co. & FRM (1998) describe various methods used by hedge fund managers. The description of various hedge fund styles certainly does not cover all the permutations, but provides an overall idea of the various strategies used by the managers. Table II compares the different segments of hedge fund in terms of investment strategy, use of leverage, and risk control.

Table II. Comparison of Different Core Segments of Hedge Fund Investments.

Segment	Investment Strategy	Use of Leverage	Risk Control
Market Neutral or Relative Value	Seek out basic mispricings between securities	Aggressively use leverage to capitalize on otherwise small pricing differences.	Broad market risk is eliminated completely to capitalize on relative mispricings
Event Driven	Seek out valuation disparities produced by corporate events that are less dependent on overall stock market gains.	Use of leverage varies from situation to situation, but in general, leverage is used conservatively.	Portfolio is diversified among a number of positions to reduce the impact of any single position that does not work out as anticipated. Hedge against market risk by purchasing index put options and short selling.
Long/Short	Seek out mispriced securities based on the business prospects of the firms, using both long and short positions.	Historically, they maintain leverage positions ranging from slightly short to 100% long.	It is often accomplished through market neutral positions. Some accomplish this within industry groups and employ greater amount of leverage.
Tactical Trading: Systematic Managers	Seek out mispriced securities using statistical analysis, which is applied to historical data.	A high degree of leverage is used to capitalize on small, but statistically significant, return opportunities.	Risk control is vital. Managers eliminate all risk except the risk that their models indicate as profitable.
Tactical Trading: Discretionary Managers	Seek out mispricing in global currency, stock, and bonds market using derivatives.	Use of leverage is kept to a minimum due to lack of risk control.	Risk control is difficult to achieve because of low correlation between currencies and indices within a market.
Fund of Funds	Seek out diversification by investing in a variety of hedge funds.	Not applicable	Risk control is achieved through diversification of hedge funds.

V. Cluster Analysis

The literature on cluster analysis is quite large and diverse. Significant work on cluster analysis has been done in various fields. Cluster analysis has frequently been employed as a classification tool. Classification is concerned with the identification of discrete categories (taxonomies), whereas structural representation is concerned with the development of a faithful representation of relationships. Cluster analysis is a statistical method of classification, yet it is different from classification. Classification in its purest form pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups. In cluster analysis, no assumptions are made concerning the number of groups or the group structure. Grouping is based on similarities or distances (dissimilarities).

Cluster analysis is an exploratory technique in which the information provided by the analyst in the form of relevant attributes is used to come up with a natural grouping of data, if any. It is important to note that cluster analysis as a data-mining tool is a futile endeavor, because knowledge of subject matter is an important input for successful clustering. Cluster analysis is a tool of discovery that reveals structure and relations in data. The results of a cluster analysis can contribute directly to the development of classification schemes. Strictly speaking, a set of results applies only to the sample on which they are based; but through appropriate modification, technique employed can be extended to describe adequately the properties of other samples and ultimately the parent population.

Unlike discriminate analysis, cluster analysis makes no prior assumptions about important differences within a population. All data analytic methods except cluster analysis have developed within a particular discipline. Factor analysis and other scaling methods were developed within the discipline of psychology. Regression is used in a variety of disciplines, but econometricians have provided a large body of literature on the technique. Cluster analysis has been independently approached by numerous disciplines (engineering, econometrics, psychology, and biology). A detailed discussion of applications of cluster analysis in social science is found in Punj and Stewart (1993). The steps of cluster analysis are discussed below.

A. Choice of Attributes

This very important step depends on the researcher's knowledge of the subject matter. The data for clustering should be described in terms of their characteristics, attributes, class membership, and other such properties. These descriptors collectively are the attributes of the problem. Attributes that are highly correlated add little in terms of distinguishing the data units. At the same time, including attributes that have large variation among data units, but are not relevant to the problem at hand will provide misleading results. The choice as to the number of attributes is different for different fields of study. Statisticians and social scientists emphasize parsimony and thus seek to minimize the number of measured attributes. Proper selection of attributes is a difficult but important task.

B. Scaling and Standardization of Attributes

Once a decision is made as to the number of attributes to be included for clustering, the next step is to select the type(s) of attribute to be used. Attributes could be quantitative, qualitative or mixed type. The common problem in real data is the lack of homogeneity among attributes of interest. The philosophy of cluster analysis is based on measuring proximity between different data points in a multi-dimensional framework. The type of attribute and the scale of measurement influence the measure of similarity calculated for the data points. Most analysis techniques assume homogeneity of data types, whereas real data sets often have mixed types. There are various ways of handling these three variations in calculating the similarity matrix.

Measurement scales could be sequentially ordered as nominal, ordinal, interval, and ratio, with the progression reflecting increasing information demands for scale definition. Nominal and ordinal scales are referred to as qualitative attributes, and interval and ratio scales are referred to as quantitative attributes. If the problem at hand has mixed data type, one can reduce the quantitative attributes into qualitative attributes by dichotomizing the quantitative attributes. This strategy reduces the quantitative variable to the lowest common denominator. The process has the risk of losing information that the quantitative attributes may contain. This might be crucial in mathematical terms, but loss of information may not be crucial for clustering purpose.

Even after the decision has been made as whether to use mixed data type or to convert the attributes into a homogeneous type, there remains the issue of standardization of attributes. There are two main reasons for standardizing a data matrix. First, the units of measurement of the attributes can arbitrarily affect the similarities among data points. Standardization helps remove the arbitrary affects. Second, standardization makes attributes contribute more equally to similarities among data points. If in the original data matrix, the value of one particular attribute is much greater than the range of values of other attributes, the attributes with a larger value will carry more weight in determining the similarities among the data points. When this affects the clustering process adversely, the attributes should be standardized to remove the effect.

C. Measure of Similarity/Dissimilarity

Cluster analysis requires a measure of similarity to be defined for every pair-wise combination of entities to be clustered. The measure interacts with the cluster analysis criteria, so that some measures give identical results with one criterion and distinctly different results with another. The combined choice of attributes, data transformation, and similarity measures leads to successful natural grouping. A basic assumption of all clustering methods is that these numerical measures of distance are all comparable to each other. If the similarity measure for a pair is 100 and for another pair it is 70, then the second pair is more similar than the first pair. There are various ways of handling quantitative, qualitative or mixed type of data in calculating the similarity matrix. Romesburg (1988) discusses in detail the various resemblance coefficients that can be calculated for quantitative, qualitative, and mixed attributes.

D. Clustering Methods

Clustering methods are of two types: (1) hierarchical clustering methods and (2) nonhierarchical clustering methods.

1. Hierarchical Clustering Method

The hierarchical clustering method can be further divided into two types: (a) agglomerative hierarchical methods and (b) divisive hierarchical methods. The agglomerative methods start with a disjoint set of entities and merge them by certain rules into fewer and more inclusive clusters, until the formation of a conjoint set. The divisive techniques begin with the conjoint set and partition the sample into smaller and smaller subsets. There are several methods of hierarchical clustering, like the linkage methods and Ward's minimum variance method. These methods are discussed in Johnson and Wichern (1998).

In hierarchical clustering, there is no provision for reallocation of objects that may have been "incorrectly" grouped at an early stage. For a particular problem, it is important to try several clustering methods and, within a given method, a couple of different ways of assigning distances (similarities). One should conclude a natural grouping only if the outcomes of several methods are consistent with one another.

2. Nonhierarchical Clustering

For a data set of *m* entities, the hierarchical methods described above provide *m* nested classifications ranging from *m* clusters of one member each to one cluster of *m* members. Nonhierarchical clustering method is designed to cluster data units into a single classification of *k* clusters, where *k* is specified *apriori* or is determined as part of the clustering method. The main idea is to choose some initial partition of data units and then alter cluster memberships to obtain a better partition. The partitioning techniques differ from the hierarchical methods in several ways. First, partitioning leads to nonhierarchical single-rank solutions; second, it allows for correction of poor initial clustering by iteratively reallocating entities. The hierarchical method, by contrast, leads to multilevel structures and allows for only one assignment. The nonhierarchical methods can be applied to much larger data sets than the hierarchical methods, as there is no need to store the similarity matrix of distances. In nonhierarchical methods, a set of cases is iteratively partitioned to maximize some predefined criterion function. The various methods of nonhierarchical clustering are described in Anderberg (1982). K-means clustering method is one of the widely used nonhierarchical clustering methods.

K-means Clustering Method

The procedure, developed by MacQueen in 1967, partitions a sample of *n* entities into *k* sets based on a euclidean distance measure. The algorithm assigns each item to the cluster having the nearest centroid (mean). The method can be described in the following three steps:

1) Partition the items into *k* initial clusters.

2) Proceed through the list of dataset items; assigning an item to the cluster whose centroid (mean) is nearest (Distance is usually computed using Euclidean measure, one of the popular distance measures, with either standardized or unstandardized observations). Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3) Repeat Step 2 until no more reassignments take place.

Rather than starting with a partition of all items into *k* preliminary groups in Step 1, one could specify *k* initial centroids (seed points) and then proceed to Step 2. K-means treats each observation in the data set as an object having a location in space. The objects are partitioned such that they are as close to each other as possible within each cluster, and far from objects in other clusters.

Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. The K-means method has a few limitations. It is necessary to take proper care in selecting initial number of clusters, avoiding local minima and misclassification.

VI. Data Organization and Methodology

The classification is carried out separately for hedge funds domiciled in the US and those domiciled outside the US. Domicile is an important characteristic that determines the organization structure of hedge funds and probably affects the return characteristics. At the same time, many hedge fund managers have two simultaneous operations, one domiciled in the US and another domiciled outside the US, but with same characteristics in terms of their investment strategy and investment style. Following attributes are considered for the classification.

1. **Asset Class**: The asset class is the broadest category, and defines the market in which the fund operates. For example, the asset classes could be stocks, bonds, currency (foreign exchange), options, futures, or warrants. Hedge funds refrain from giving out exact information as to their portfolio composition. They are not required to give this information and avoid giving too much disclosure to discourage herding. This attribute is subdivided into four different sub-attributes; stocks, bonds, currency and derivatives. Each of this subdivision is considered as a separate attribute for the clustering purpose. The attribute derivative is composed of options, futures, and warrants. No specific ordering is given to any of these three derivatives. In fact, they are taken as derivative 1, derivative 2, and derivative 3. If a hedge fund uses all three types of derivatives, proper weight is given to it to represent its use of all types of derivatives compared to a hedge fund that uses just one derivative type.
2. **Size**: The net asset value could be a measure of size. However, net asset value will change from year to year and will depend on the method of calculation. Mutual funds are valued daily with a published net asset value (NAV). There are no specific rules governing hedge fund pricing. U.S. hedge funds provide investors only a monthly estimate of percentage gain or loss. The minimum purchase is used as a proxy for size; it represents the size of a “unit share” in the particular hedge fund.
3. **Fee**: In general, hedge funds charge two types of fees: asset management fee and incentive fee. The asset management fee is based on percentage of assets in the fund, usually 1 or 2 percentage points per year. This includes legal, audit, administrative, and other expenses. It is paid monthly or quarterly and may be due at the beginning or end of each period. The fee is automatically deducted pro rata from each investor’s account. Asset management fee is almost same for all hedge funds. Therefore, the asset management fee is not considered as an attribute for the classification. The incentive fee or the ‘carried interest’ is the hedge fund manager’s share in the fund’s profit. This incentive fee is what differentiates hedge funds from mutual funds. Usually this is 20 percent, but it could vary from zero percent to 50 percent. Incentive fee is used as an attribute for classification.
4. **Leverage**: The Investment Company Act severely limits a mutual fund's ability to leverage or borrow against the value of securities in its portfolio. The SEC requires that funds practicing certain investment techniques, including the use of options, futures, forward contracts, and short selling, must "cover" their positions. The effect of these constraints has been to limit leveraging by mutual fund portfolio managers. Leveraging and other higher-risk investment strategies are a hallmark of hedge fund management. Leverage varies from zero to 70 times the asset value.
5. **Liquidity**: Some funds trade short term and invest in instruments that can be traded easily. Other funds are less liquid because of either their strategies, the types of instruments they hold, or the size of their holdings. Hedge funds refrain from disclosure of their specific trading strategies. There is no direct measure of liquidity that could be calculated for hedge funds. Redemption frequency is considered as a measure of liquidity. The redemption frequency varies from daily to annual.

A. Data Organization

The attributes considered for the classification have different scales of measurement (quantitative and qualitative). The attributes for asset class are qualitative type, giving the information as to whether the hedge fund invest in a particular asset class or not (no portfolio composition), whereas other attributes are quantitative attributes. It is decided to convert all the attributes to qualitative type, while care is taken to minimize the loss of information in the conversion process.

Binary logic is the popular choice for representation of qualitative attributes. However, binary logic cannot represent the multi-stage characteristics of many of the attributes. Multi-valued logic helps to represent the multi-stage character of a qualitative variable by assigning different logical states to different stages. For example, the attribute leverage if converted into a binary qualitative variable would just identify the presence or absence of leverage. Hedge funds are differentiated from other investment types not just by their mere use of leverage but by the varied degree of leverage. To accommodate this characteristic of the attribute, it is necessary to use multi-valued logic.

The evolutionary fuzzy logic, a form of multi-valued logic, is based on the notion of graded truth and falsity, similar to other multi-valued logical systems. However, it allows the researcher to change the number of states without changing the membership of already classified members. It also incorporates a state of unknown where truth and falsity merge. This feature helps to address missing information in the dataset for any attribute. In the context of this paper, this logical system allows to convert the quantitative variable into qualitative variable with ordinal measurement scale. For example, a hedge fund that identifies itself as using a leverage of 10%, has membership in *State 3 which* encompasses *State 1* (up to 1% leverage) and *State 2* (1%-2% of leverage).

The next step is to convert all the quantitative attributes into qualitative attributes using a two-stage process. In stage 1, the data is examined to get an idea about the ranges of each of the quantitative attributes. Table III lists the available ranges for each of the attributes along with the number of funds in each range. In stage 2, groups are formed using different ranges of the attributes. The new ranges are than assigned a state, using ordinal scale. Table III provides the assigned states and the number of funds in each state for the attributes of incentive fee, amount of leverage used, redemption frequency, and minimum purchase respectively.

The qualitative information of the asset class is used differently. As already mentioned above, the asset class is subdivided into four attributes; stocks, bonds, currency and derivatives. Hedge fund managers in the ZCM/Hedge database used three types of derivatives. No distinction is made as to the type of derivative used for assigning states to this attribute, but the number of derivatives that a hedge fund manager uses is given due consideration. For example if a hedge fund manager uses two of the derivative instruments (options and futures, future and warrants, etc) the fund is assigned *State 2* for the derivatives attribute.

The selection criteria provides eight attributes; four (stocks, bonds, currency and derivatives) for the asset class and, one each for incentive fee, leverage, redemption frequency and minimum purchase. These eight attributes are used to find out the similarities between the hedge funds in the database. Since the derivatives can have a maximum state of +3, the other attributes of the asset class are also given a state of +3 if they invest in the particular asset class and a state of -3 if they do not invest in that asset class. A *State* of zero (0) is assigned for hedge funds that do not have any information as to their use or lack of use, concerning the attribute in question. Finally, all attributes are converted to uniform logical states of +3 to -3 while maintaining the number of states related to each attribute. This is done to assign equal importance to each of the eight attribute used for the classification. All logical states are represented by a decimal number which is used for the computation of distance measure.

Table III. Logical States for Incentive Fee, Leverage, Redemption Frequency and Minimum Purchase Attributes.

Values	Number of Funds	Grouping	Logical State	Number of Funds
Panel A. Incentive Fee				
0%	140	Less than equal to 10%	1	351
0.2-8%	28			
9-10%	183			
12-15%	11	Greater than 10% and less than equal to 20%	2	2,253
15.01-20%	2,242			
20.02-60%	189	Greater than 20%	3	189
Panel B. Leverage				
No leverage	109	Less than equal to 1X	1	1,223
Less than equal to 0.8X	77			
Greater than 0.8X and less than equal to 1X	1,037			
Greater than 1X and less than equal to 1.25X	345	Greater than 1X and less than equal to 2X	2	1,013
Greater than 1.25X and less than equal to 1.5X	348			
Greater than 1.5X and less than equal to 2X	320			
Greater than 2X and less than equal to 9X	196	Greater than 2X and less than equal to 10X	3	225
Greater than 9X and less than equal to 10X	29			
Greater than 10X and less than equal to 25X	30	Greater than 10X and less than equal to 30X	4	44
Greater than 25X and less than equal to 30X	14			
Greater than 30X and less than equal to 35X	6	Greater than 30X and less than equal to 50X	5	14
Greater than 35X and less than equal to 50X	3			
Greater than 50X and less than equal to 70X	11	Greater than 50X and less than equal to 70X	6	3
Not known	273	Not declared	7	273
Panel C. Redemption Frequency				
Daily	61	Less than equal to monthly	1	1,166
Weekly	89			
Bimonthly	13			
Monthly	1,003			
Quarterly	1,084	Greater than monthly and less than equal to semiannually	2	1,259
Semiannually	175			
Annually	321	Greater than semiannually	3	372
More than Annual	51			
Panel D. Minimum Purchase				
<=$100	180	Less than equal to $25,000	1	386
$101-$5,000	17			
$5,001-$25,000	189			
$25,001-$50,000	87	Less than equal to $50,000	2	87
$50,001 - $100,000	383	Less than equal to $100,000	3	383
$100,001-$500,000	1,179	Less than equal to $500,000	4	1,179
$500,001-$1 million	649	Less than equal to $1 million	5	649
$1 million -$5 million	103	Less than equal to $25 million	6	111
$5 million -$25 million	8			
More than $25 million	2	Greater than $25 million	7	2

B. Methodology

This paper uses the nonhierarchical K-means clustering method to classify hedge funds based on the attributes discussed above. The classification is done separately for hedge funds domiciled in US and those domiciled outside US. The attributes selected are such that the characteristics of a hedge fund represented with this set of attributes should not change with time unless the very structure of the hedge fund is changed. The market conditions, which will have an impact on the return of the hedge fund, will not change the cluster-membership of the hedge fund. Proper care is taken to ensure that the attributes selected represent the structure and not the return characteristics of the hedge fund. In general, all of these attributes may affect the return of the hedge fund like other attributes that affect the return but not considered for classification purpose. The K-means method identifies the closest cluster center (in terms of a distance measure) for each hedge fund and assigns the hedge fund to that cluster. The method then re-computes each cluster center based on the hedge funds assigned to that cluster. The assignment of hedge funds to clusters is done in a way that improves the within-cluster variance. Some issues, for the K-means clustering method, are discussed here.

1. Choice of Number of Clusters

The number of clusters must be specified in advance. It is true that the number of clusters is not known *apriori* and this is probably the most difficult problem in cluster analysis. When clustering is done by optimizing a criterion function like minimizing within-cluster variance, the usual approach is to repeat the clustering algorithm with different number of clusters. There is a reduction in error, each time number of clusters is increased by transfer of a single hedge fund to the new cluster. Therefore, the variance will decrease monotonically with increase in number of clusters. If *n* hedge funds are grouped into *k*-opt well-separated clusters, the variance should decrease rapidly until number of clusters *k*=*k*-opt, decreasing much more slowly after that until it reaches zero when *k*=*n*. This paper identifies the value of *k*-opt by plotting the within-cluster variance versus number of clusters and focusing on 'the knee of the curve'.

2. Avoiding Local Minima

Once the decision is made as to the range of cluster numbers that can be tried, the clustering algorithm is run for the range of values. The K-means method uses an iterative technique to minimize the within-cluster variance. It arbitrarily partitions the *n* hedge funds into *k* clusters and computes the coordinates of the cluster centroid (mean), $\left(\overline{\mathrm{X}}_1^k, \overline{\mathrm{X}}_2^k, \ldots \overline{\mathrm{X}}_j^k\right)$, where number of attributes *j* is equal to eight in this classification scheme. The centriod coordinate of cluster *k*, consisting of *m* hedge funds is $\overline{\mathrm{X}}_j^k$ and is given as follows.

$$\begin{aligned}\overline{\mathrm{X}}_j^k &= \frac{\left(\mathrm{X}_{1j}^k + \mathrm{X}_{2j}^k + \ldots + \mathrm{X}_{mj}^k\right)}{m} \\ &= \frac{1}{m}\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{X}_{ij}^k\end{aligned} \tag{1}$$

Distance of each hedge fund from the group centroids (*k* centroids) is then computed using one of the distance measures described below and each hedge fund is reassigned to the nearest cluster, in the sense that it minimizes the within-cluster variance. Since the initial partition is done arbitrarily, the K-means method like many other types of numerical minimizations, could reach local minima. The reaching of local-minima (and thus not optimizing the minimum variance criterion) is avoided by replicating the randomization process of *k* clusters many times. It is generally recommended that the replication be done for at least *k+1* number of times for *k* clusters. In this paper, the randomization process is replicated between 100 and 1000 times for each *k*.

3. Distance Measure

The K-means algorithm can use different distance measures. This study uses two popular distance measures: 'Squared Euclidean' and sum of absolute distances ('City-block' distance measure). The 'Squared Euclidean' distance, d_{xy}^{2}, between two p-dimentional observations (items) $\mathrm{X}=\left[x_1, x_2, \ldots, x_p\right]'$ and $\mathrm{Y}=\left[y_1, y_2, \ldots, y_p\right]'$ is

$$\begin{aligned} d_{xy}^{2} &= \left(x_1-y_1\right)^2+\left(x_2-y_2\right)^2+\ldots\left(x_p-y_p\right)^2 \\ &= \left(\mathrm{X}-\mathrm{Y}\right)'\left(\mathrm{X}-\mathrm{Y}\right) \end{aligned} \tag{2}$$

The 'City-block'distance, d_{xy}, between two p-dimentional observations (items) $\mathrm{X}=\left[x_1, x_2, \ldots, x_p\right]'$ and $\mathrm{Y}=\left[y_1, y_2, \ldots, y_p\right]'$ is

$$\begin{aligned} d_{xy} &= \left(x_1-y_1\right)+\left(x_2-y_2\right)+\ldots\left(x_p-y_p\right) \\ &= \sum_{j=1}^{p}\left(x_j-y_j\right) \end{aligned} \tag{3}$$

The 'Squared Euclidean' distance measure is selected for the classification as discussed in the results section.

4. Silhouette Value

Silhouette value provides a measure of separation of clusters. The silhouette value for each hedge fund is a measure of how similar that hedge fund is to the hedge funds in its own cluster compared to the hedge funds in other clusters. The value ranges from +1 to -1. The silhouette value, $S(i)$, for the i - th hedge is defined as

$$S(i)=((\min(b(i,k))-a\,(i))\,/\,(\max(a\,(i),\min(b(i,k))) \tag{4}$$

where:

$a\,(i)$ is the average distance of the i - th hedge from other hedge funds in the same cluster,

$b\,(i,k)$ is the average distance of the i - th hedge fund from hedge funds in the neighboring cluster k,

$\min(\,b\,(i,k))$ is the minimum of all the average distances of the i - th hedge fund from hedge funds in other clusters.

A silhouette value of +1, for a(i)=0, for a particular hedge fund indicates that the hedge fund is very distant from the neighboring clusters. A silhouette value of 0, for $\min(b(i,k)=a\,(i)$, indicates that the hedge fund is not distinctly in one cluster or another. A silhouette value of -1, for $\min(b(i,k)=0$, indicates that the hedge fund is assigned to the wrong cluster. The same information can be displayed using the silhouette plot. Instead of visually comparing silhouette plots for different number of clusters, *k*, the decision as to the optimal number of clusters (with minimum misclassification) is done by comparing the mean silhouette values for different cluster sizes.

5. Tools

A stand-alone Workstation equipped with dual 2.2 GHz Intel Xeon Processors, 1.0 GB of RAM, dual 18 GB SCSI hard drives, Microsoft Windows XP Professional OS, and, MATLAB software package with Statistics Toolbox and Microsoft Office XP has been used for all data analysis.

VII. Results

The first step in using the K-means clustering method is to check the validity of the implemented algorithm. K-means method minimizes the sum of distances of all hedge funds from the cluster centroid. This distance should decrease as the number of clusters (*k*) is increased. Figure V and Figure VI shows the sum of the distances (Y-axis) for different number of clusters (X-axis) using two distance measures, 'City-block' and 'Squared Euclidean', respectively. The distance decreases as the number of clusters is increased. These plots also provide the range of *k*-clusters that can be used for the classification purpose.

The rate of decrease in the distance measure should be rapid initially; gradually the rate of decrease should slow down, and ultimately the distance measure should be zero when the value of *k* equals the number of funds in the class. The figures show that this is true for both the distance measures and for both class of hedge funds. In fact, the distance measure reaches zero even before the number of clusters reach the total number of funds; indicating that there are at least some funds in the class of hedge funds that have the same characteristics as measured by the attributes chosen for the classification.

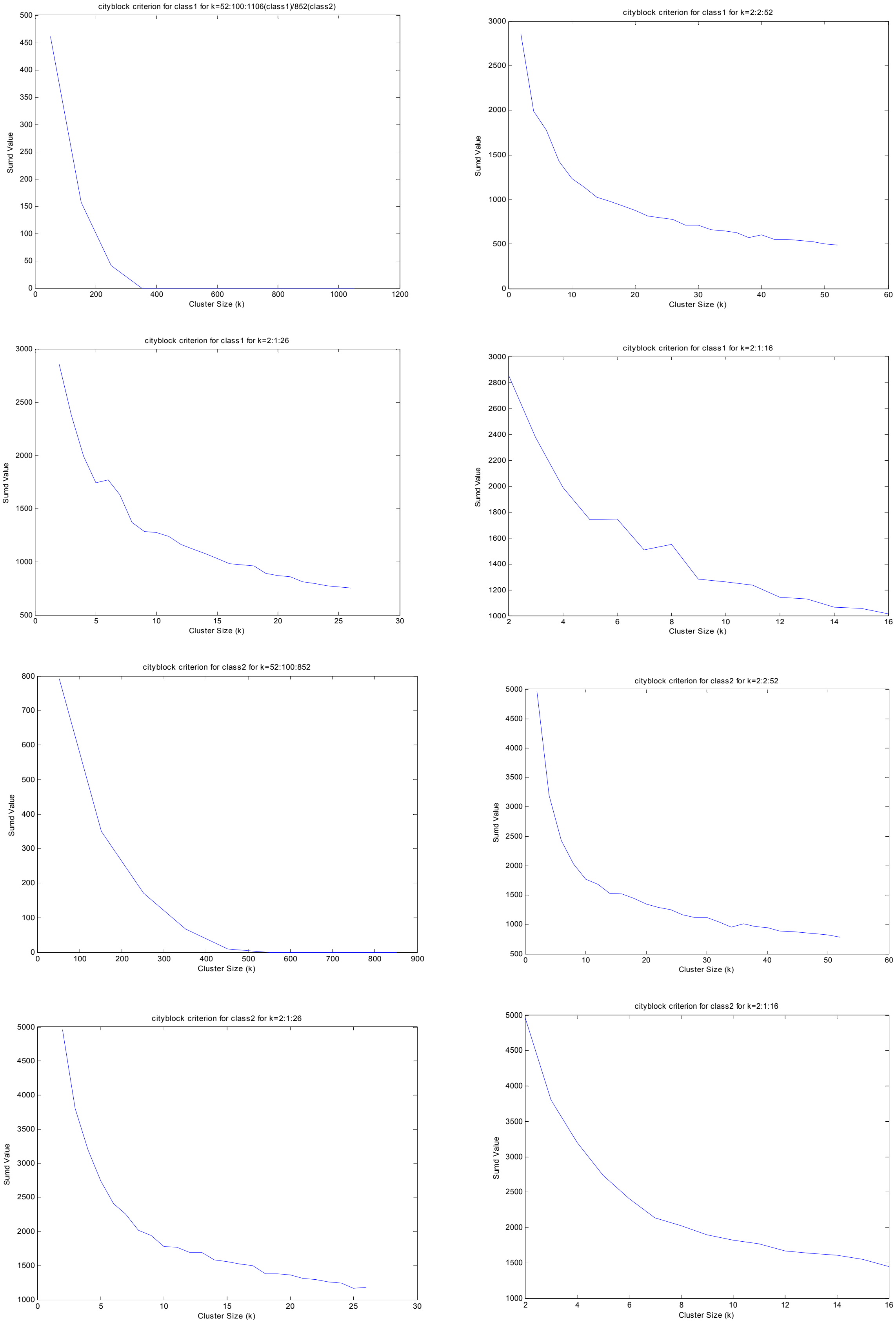
There are 1,106 hedge funds in Class 1 and 852 hedge funds in Class 2. The number of clusters is varied in steps of 100 in the first plot for the range of 52 to 1106 ; it is varied in steps of 1 for the range of 2 to 52 for the second plot. The next two plots reduce the range to 2 to 26 and 2 to 16 respectively to help focus on the decrease in distance measure over the relevant range. The relevant range for number of clusters is between five and nine for Class 1 using the 'City-block' distance measure; and it is between five and eight using the 'Squared Euclidean' measure. For Class 2, the relevant ranges for number of clusters are four to ten and seven to ten respectively using two different distance measures. Note that the distance measure using 'City-block' criterion does not decrease smoothly for Class 1 hedge funds.

It is important to check the separation of clusters for classifications using different number of clusters. This is checked by plotting silhouette values of hedge funds for different number of clusters. The silhouette value varies from +1 to -1. The greater the similarity of a hedge fund within the cluster, the closer the silhouette value is to +1. Figure VII and Figure VIII shows the silhouette values for Class 1 and Class 2 hedge funds respectively, using two different distance measures. Silhouette plots are shown for the two sets of classification where transition in silhouette value occurs. For example, the silhouette value is +1 for all funds for the classification of Class 1 hedge funds using 352 or more number of clusters (using both distance criteria). The silhouette value decreases for some funds in the 252-cluster classification. There is a trade-off between the silhouette value and the number of clusters. Obviously, a classification scheme which classifies 1106 hedge funds into 352 clusters is not of much practical use. Probably, each hedge fund is different. The idea of classification is to come up with a manageable number of groups of hedge funds, without having hedge funds with very different characteristic being grouped together.

In order to find the optimum number of clusters, mean silhouette values are computed for different number of clusters, ranging from 2 to 16. The range for the number of clusters is decided from the plots of distance measure shown in Figure V and VI. Table VII compares the distance measures and mean silhouette values for different number of clusters. As expected, the distance measure decreases as the number of cluster increases. The distance measure gives an idea of the proximity of hedge funds that are members of the same cluster. Lower the distance measure, better the optimization is in terms of closeness of hedge funds within a cluster.

As mentioned above, the distance measure decreases as the number of cluster increases, but the rate of decrease slows down after reaching a certain number of clusters. The mean silhouette values, shown in Table IV, indicate the misclassification of hedge funds for a particular cluster-size. The closer the value is to 1, the closer the hedge funds in that cluster. There are seven different number of clusters for Class 1 hedge funds, with the mean silhouette value greater than 0.5 and, four different number of clusters for Class 2 hedge funds, using city-block distance measure. There are thirteen different number of clusters for Class 1 of hedge funds and fifteen number of clusters (all) for Class 2 of hedge funds with mean silhouette values greater than 0.5, using the 'Squared Euclidean' distance measure. The highlighted values in Table IV are the number of clusters that need further investigation. The criteria considered for selecting the number of clusters, applicable to both class of hedge funds, as candidates for further investigation are the rate of decrease in distance measure and a simultaneous increase in the mean silhouette value.

Figure V. Distance Measures for Class 1 and 2 Funds using ‘City-block’ Criterion.



cityblock criterion for class1 for k=52:100:1106(class1)/852(class2)

cityblock criterion for class1 for k=2:2:52

cityblock criterion for class1 for k=2:1:26

cityblock criterion for class1 for k=2:1:16

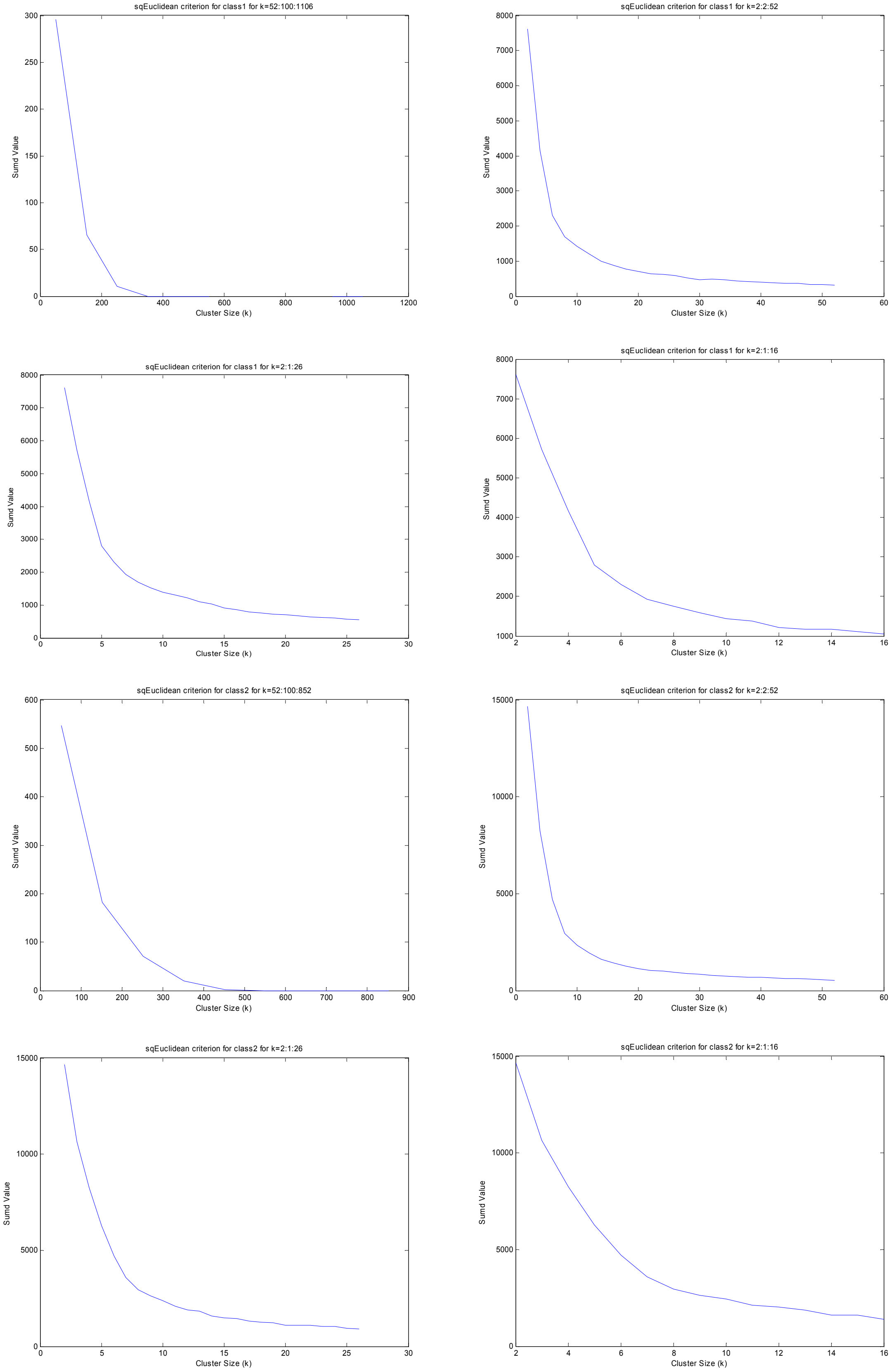
cityblock criterion for class2 for k=52:100:852

cityblock criterion for class2 for k=2:2:52

cityblock criterion for class2 for k=2:1:26

cityblock criterion for class2 for k=2:1:16

Figure VI. Distance Measures for Class 1 and 2 Funds using 'Squared Euclidean' Criterion.



sqEuclidean criterion for class1 for k=52:100:1106

sqEuclidean criterion for class1 for k=2:2:52

sqEuclidean criterion for class1 for k=2:1:26

sqEuclidean criterion for class1 for k=2:1:16

sqEuclidean criterion for class2 for k=52:100:852

sqEuclidean criterion for class2 for k=2:2:52

sqEuclidean criterion for class2 for k=2:1:26

sqEuclidean criterion for class2 for k=2:1:16

Figure VII. Silhouette Plots for Different Cluster Sizes for Class 1 Hedge Funds.

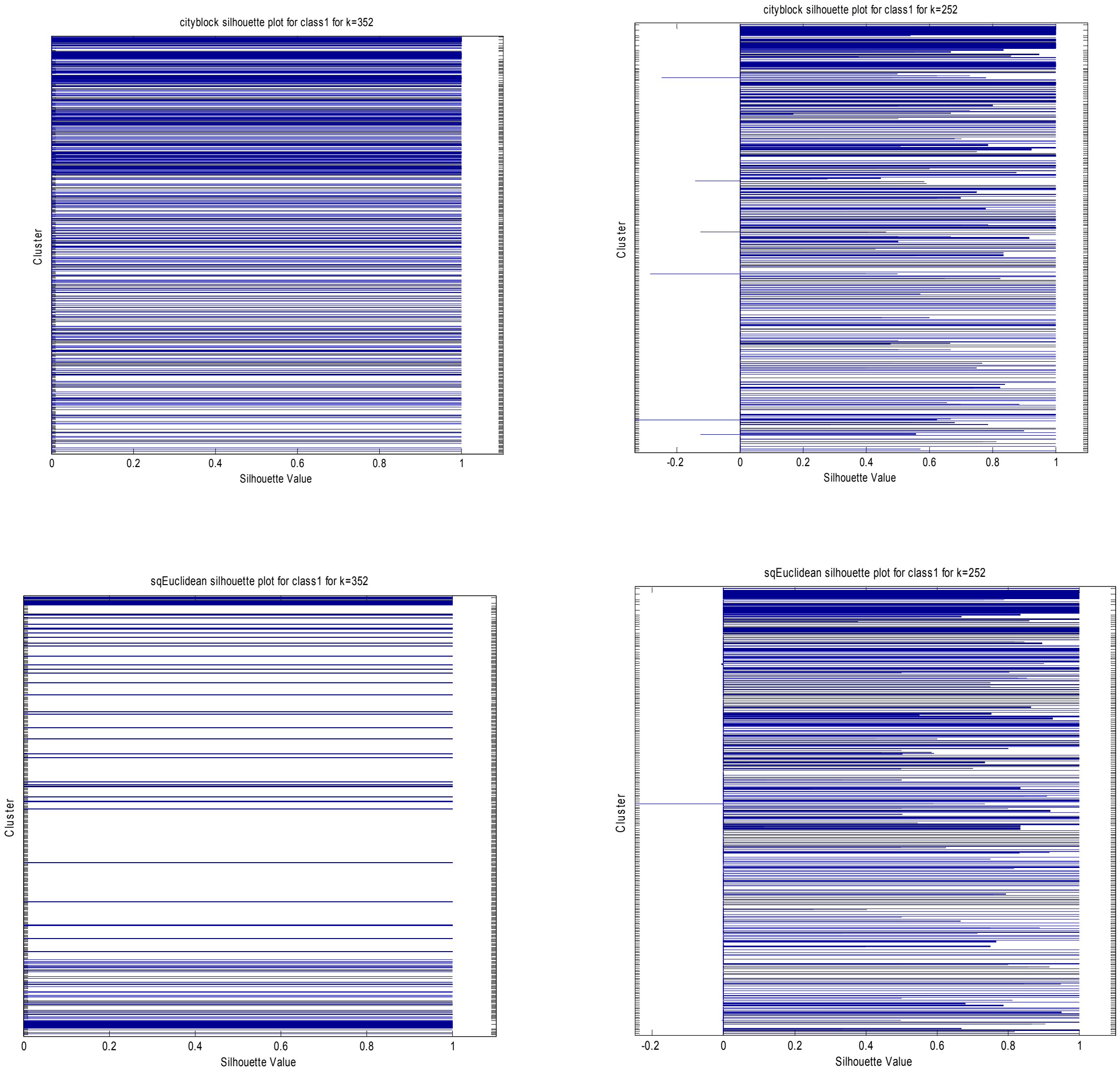


Figure VIII. Silhouette Plots for Different Cluster Sizes for Class 2 Hedge Funds.

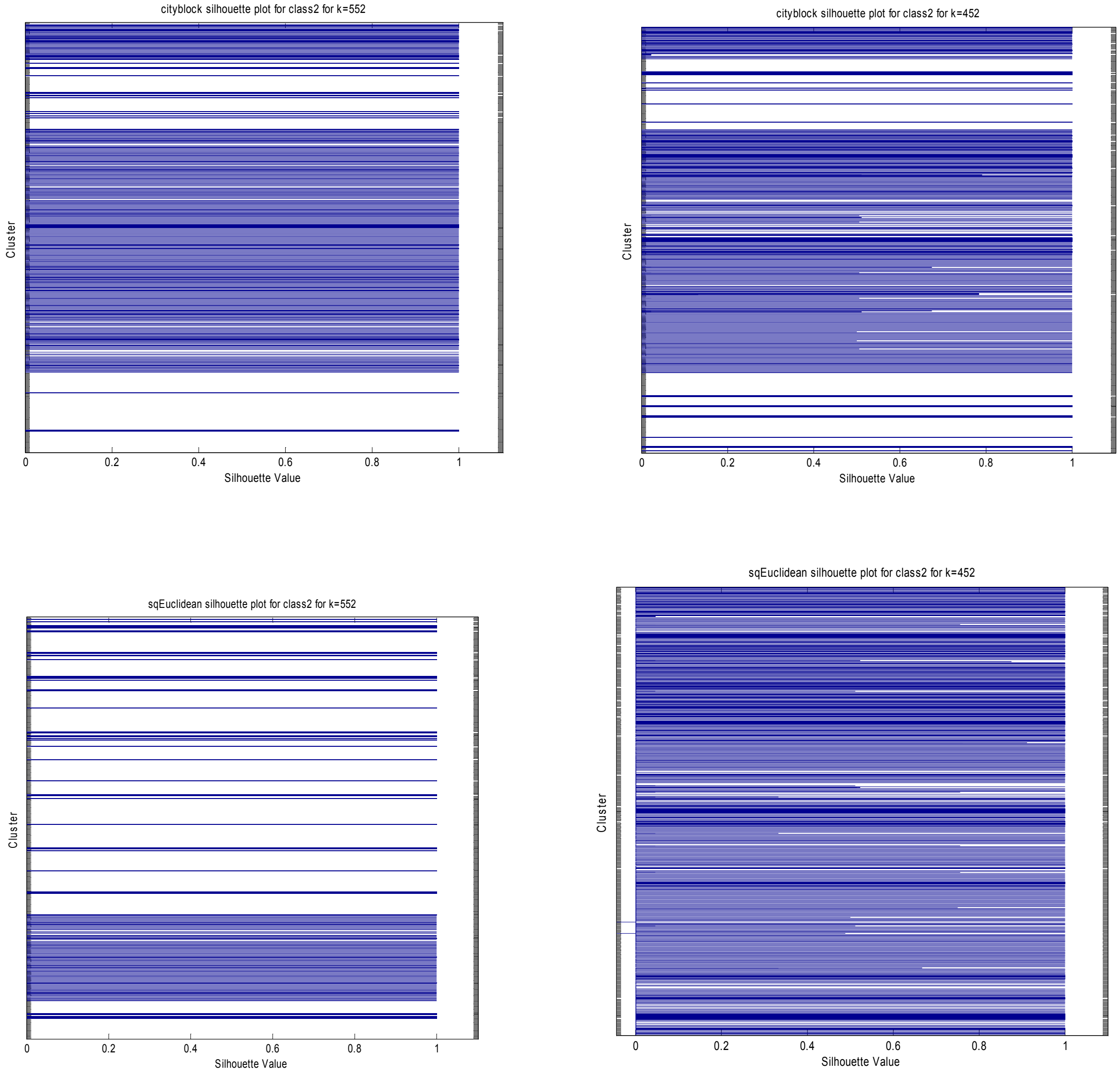


Table IV. Comparison of Distance Measures and Silhouette Values for different Number of Clusters.

Number of Clusters (*k*)	'City-block' Criterion				'Squared Euclidean' Criterion			
	Class 1		Class 2		Class 1		Class 2	
	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value
2	2852	0.69	4965	0.46	7619	0.80	14665	0.55
3	2377	0.69	3801	0.47	5716	0.82	10634	0.59
4	1991	0.65	3197	0.44	4169	0.84	8251	0.62
5	1742	0.66	2738	0.46	2799	0.84	6282	0.65
6	**1747**	**0.50**	**2409**	**0.52**	**2308**	**0.86**	**4721**	**0.73**
7	1509	0.42	2133	0.56	**1933**	**0.87**	**3598**	**0.78**
8	1550	0.35	2029	0.48	1752	0.53	2960	0.79
9	**1285**	**0.52**	**1895**	**0.51**	**1595**	**0.57**	**2619**	**0.81**
10	1261	0.38	1825	0.52	1442	0.68	2426	0.82
11	1236	0.40	1766	0.41	1373	0.63	2134	0.83
12	1142	0.54	1666	0.42	1211	0.70	2022	0.73
13	1130	0.42	1632	0.48	1167	0.50	1866	0.74
14	1067	0.47	1606	0.43	1175	0.57	1614	0.72
15	1059	0.35	1549	0.41	1108	0.42	1620	0.61
16	1016	0.46	1450	0.43	1054	0.46	1380	0.69

It is important to check the robustness of the number of clusters. This is achieved in two ways. In the clustering algorithm, the randomization of the seed points is done 1000 times for each *k*-cluster classification to avoid the local minima. Even after replicating the randomization process for 1000 times, the clustering algorithm is run four more times to check if there is any significant change in the distance measure and the mean silhouette values. Table V provides the distance measure and the mean silhouette value for different runs for the six-cluster and the nine-cluster classification. There is no change in distance measures and mean silhouette values for the six-cluster classification, using 'Squared Euclidean' distance measure. The nine-cluster classification shows a slight variation in mean silhouette values for Class 1 hedge fund but the variation is not significant. The values which are different for different runs using city-block distance measure are highlighted in Table V. It appears that the city-block distance measure lacks robustness compared to the 'Squared Euclidean' measure. The mean silhouette values are also much lower than the corresponding values using Euclidean distance measure. However, seven-cluster classification using 'Squared Euclidean' distance measure also show variations in distance measure and mean silhouette values for Class 2 funds.

Table VI provides the same information for the seven-cluster classification using 'Squared Euclidean' distance measure only. The seven-cluster classification is not a candidate for investigation using city-block distance measure. The fourth and the last column in the Table VI provides the order of cluster number for each run taking run 1 as the base run. It may be noted that the order of cluster number changes because of the randomization process, but there is no change in distance measure or the mean silhouette values.

Table VII compares the distance measures and mean silhouette values for the six-cluster, seven-cluster, and nine-cluster classifications using 'Squared Euclidean' distance measure. The seven-cluster classification appears to perform better in terms of optimizing the minimum distance criterion and reducing misclassification for both the classes, whereas the nine-cluster classification performs better for Class 2 only.

Table V. Robustness of Six-Cluster and Nine-Cluster Classifications.

Number of Clusters (*k*)	Run No	‘City-block’ Criterion				‘Squared Euclidean’ Criterion			
		Class 1		Class 2		Class 1		Class 2	
		Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value
6	1	1578	0.51	2410	0.52	2308	0.86	4720	0.73
	2	**1610**	**0.42**	2410	0.52	2310	0.86	4720	0.73
	3	1580	0.51	2410	0.52	2310	0.86	4720	0.73
	4	1580	0.51	2410	0.52	2310	0.86	4720	0.73
	5	1580	0.51	2410	0.52	2310	0.86	4720	0.73
9	1	**1297**	**0.50**	**1870**	**0.53**	1554	0.68	2620	0.81
	2	**1260**	**0.42**	**1870**	**0.53**	1550	0.68	2620	0.81
	3	**1310**	**0.41**	**1830**	**0.51**	1550	0.68	2620	0.81
	4	**1270**	**0.41**	**1830**	**0.52**	1530	**0.66**	2620	0.81
	5	**1290**	**0.40**	**1840**	**0.59**	1540	**0.64**	2620	0.81

Table VI. Robustness of ‘Squared Euclidean’ Distance Measure For Seven-Cluster Classification.

Run No.	Class 1			Class 2		
	Distance Measure	Mean Silhouette Value	Order of Cluster Number	Distance Measure	Mean Silhouette Value	Order of Cluster Number
1	1933.35	0.87	1234567	3597.514	0.78	1234567
2	1933.35	0.87	1267543	3597.514	0.78	7523461
3	1933.35	0.87	1376245	3597.514	0.78	6523471
4	1933.35	0.87	1364527	3597.514	0.78	2674153
5	1933.35	0.87	5136427	3597.514	0.78	2541763
6	1933.35	0.87	6315742	3597.514	0.78	5721634
7	1933.35	0.87	4351627	3597.514	0.78	2657413
8	1933.35	0.87	7435162	3597.514	0.78	5467213
9	1933.35	0.87	7541326	3597.514	0.78	1732645
10	1933.35	0.87	6127543	3597.514	0.78	2374165

Table VII. Comparison of Six-, Seven- and Nine-Cluster Classifications using ‘Squared Euclidean’ Distance Measure.

Number of clusters (*k*)	Class 1		Class 2	
	Distance Measure	Mean Silhouette Value	Distance Measure	Mean Silhouette Value
6	2308	0.86	4720	0.73
7	**1933**	**0.87**	**3597**	**0.78**
9	1530-1595	0.57-0.68	2620	0.81

It is important to identify the distribution of funds from the original ZCM/Hedge classification to the new classification scheme for different number of clusters before any conclusion as to the appropriate number of clusters is reached. Table VIII shows the cross-tabulation of ZCM/Hedge classification versus the six-cluster classification for Class 1 hedge funds using the algorithm described in this paper. It may be noted that the cluster 3, which is the largest in terms of total number of funds, consists of hedge funds from all the categories of ZCM/Hedge database. It is interesting to see that the new classification scheme has not kept intact any of the original classification of ZCM/Hedge database. It can be inferred that the hedge fund categories of the ZCM/Hedge database consists of heterogeneous hedge funds with regards to the attributes that are considered in the present study.

Table VIII. Six-Cluster Classification of Class 1 Funds using 'Squared Euclidean' Distance Criterion.

	Number of Funds in Cluster						**Total**
ZCM/Hedge Category	**1**	**2**	**3**	**4**	**5**	**6**	**Funds**
Event Driven	79	**0**	8	13	1	17	**118**
Global International	17	3	1	**0**	**0**	4	**25**
Global Regional Established	217	15	35	8	4	5	**284**
Global Regional Emerging	31	7	5	0	0	10	**53**
Global US	63	8	4	9	5	5	**94**
Global Macro	36	4	8	3	6	8	**65**
US Opportunistic	20	**0**	2	7	2	1	**32**
Long Only/Leveraged	23	1	1	**0**	0	3	**28**
Market Neutral	137	3	15	16	5	81	**257**
Sector	115	1	5	4	**0**	**0**	**125**
Short Sellers	21	**0**	2	2	**0**	**0**	**25**
Total Funds	**759**	**42**	**86**	**62**	**23**	**134**	**1106**

Table IX shows the cross-tabulation for the ZCM/Hedge classification versus the seven-cluster classification for Class 1 hedge funds using the algorithm described in this paper. It may be noted that the only difference between the six-cluster and the seven-cluster classification is in the break-up of cluster number 6 of the six-cluster classification into two clusters, 6 & 7, for the seven-cluster classification. All other distributions remain unchanged. The specific cluster number is arbitrary, since the number changes every time the algorithm is run. However, the composition of the clusters is very robust. The comparison is done for different runs and results are the same.

Table IX. Seven-Cluster Classification of Class 1 Funds using 'Squared Euclidean' Distance Criterion'.

	Number of Funds in Cluster							**Total**
ZCM/Hedge Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Funds**
Event Driven	79	0	8	13	1	2	15	**118**
Global International	17	3	1	0	0	0	4	**25**
Global Regional Established	217	15	35	8	4	1	4	**284**
Global Regional Emerging	31	7	5	0	0	2	8	**53**
Global US	63	8	4	9	5	1	4	**94**
Global Macro	36	4	8	3	6	2	6	**65**
US Opportunistic	20	0	2	7	2	0	1	**32**
Long Only/Leveraged	23	1	1	0	0	1	2	**28**
Market Neutral	137	3	15	16	5	16	65	**257**
Sector	115	1	5	4	0	0	0	**125**
Short Sellers	21	0	2	2	0	0	0	**25**
Total Funds	**759**	**42**	**86**	**62**	**23**	**25**	**109**	**1106**

Table X and Table XI show the cross-tabulation of ZCM/Hedge classification versus the six-cluster and seven-cluster classification respectively for Class 2 hedge funds. It is clear from Table VII, that for Class 2 hedge funds the rate of decrease in distance measure and the increase in mean silhouette value is much higher for the seven-cluster classification compared to the six-cluster classification. It may be noted that the main difference between the six-cluster and the seven-cluster classification is in the break-up of cluster number 6 of the six-cluster classification into two clusters, 6 & 7. However, four funds of the 'Global International' category that were in cluster number 6 of the six-cluster classification got split into two different clusters 5 and 7 of the seven-cluster classification; one fund moved to cluster number 5 and three funds moved to cluster number 7. All the eight funds of the same category in six-cluster classification moved to cluster number six of the six-cluster classification. All other distributions remain unchanged. It is also observed that the category 'Short Sellers' seem to be well classified, since almost all hedge funds in this category have remained in the same cluster for different number of clusters.

Table X. Six-Cluster Classification of Class 2 Funds using 'Squared Euclidean' Distance Criterion.

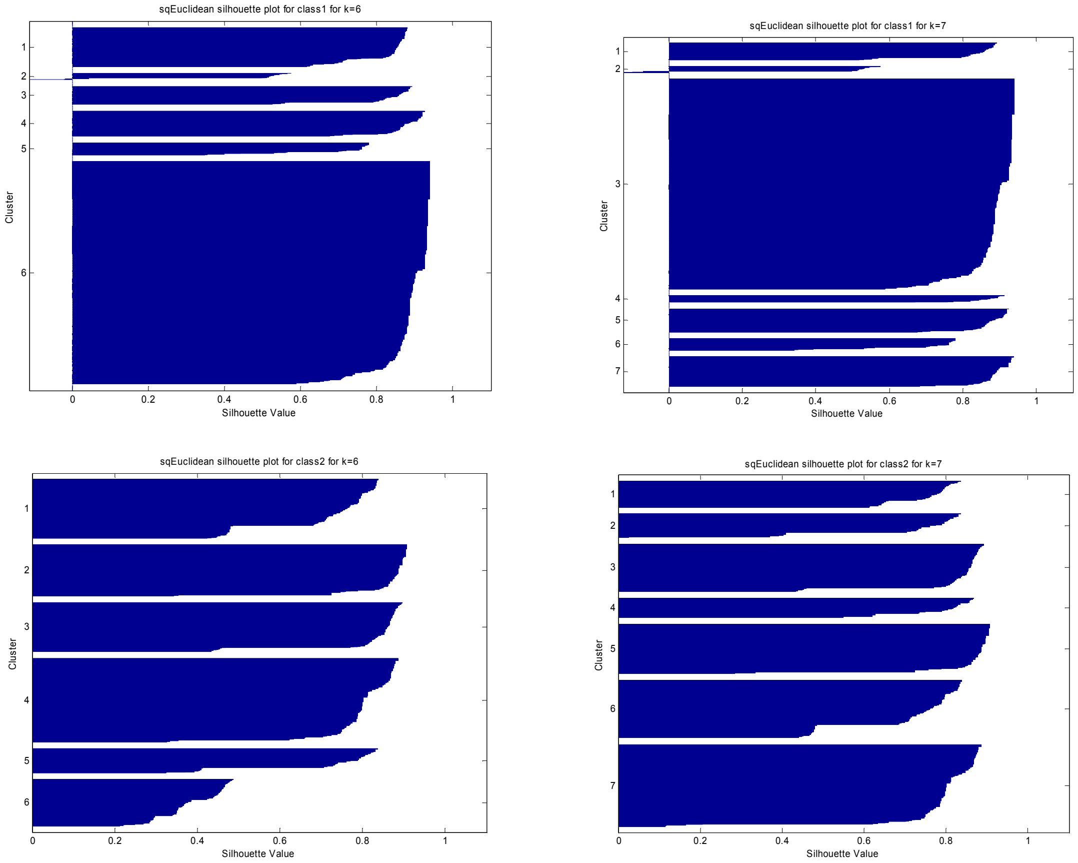
ZCM/Hedge Category	**Number of Funds in Cluster**						**Total Funds**
	1	**2**	**3**	**4**	**5**	**6**	
Event Driven	8	14	10	8	33	**0**	**73**
Global International	15	10	5	3	**8**	**4**	**45**
Global Regional Established	25	78	42	15	12	1	**173**
Global Regional Emerging	2	10	7	20	24	19	**82**
Global US	6	15	16	20	13	10	**80**
Global Macro	3	13	4	35	13	19	**87**
US Opportunistic	**0**	3	3	**0**	2	**0**	**8**
Long Only/Leveraged	**0**	1	1	1	3	1	**7**
Market Neutral	8	45	31	30	46	71	**231**
Sector	**0**	29	18	1	5	1	**54**
Short Sellers	**0**	9	2	**0**	1	**0**	**12**
Total Funds	**67**	**227**	**139**	**133**	**160**	**126**	**852**

Table XI. Seven-Cluster Classification of Class 2 Funds using 'Squared Euclidean' Criterion'.

ZCM/Hedge Category	**Number of Funds in Cluster**							**Total Funds**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
Event Driven	8	14	10	8	33	**0**	**0**	**73**
Global International	15	10	5	3	**1**	**8**	**3**	**45**
Global Regional Established	25	78	42	15	12	1	**0**	**173**
Global Regional Emerging	2	10	7	20	24	5	14	**82**
Global US	6	15	16	20	13	5	5	**80**
Global Macro	3	13	4	35	13	7	12	**87**
US Opportunistic	**0**	3	3	**0**	2	**0**	**0**	**8**
Long Only/Leveraged	**0**	1	1	1	3	1	0	**7**
Market Neutral	8	45	31	30	46	51	20	**231**
Sector	**0**	29	18	1	5	1	**0**	**54**
Short Sellers	**0**	9	2	0	1	**0**	**0**	**12**
Total Funds	**67**	**227**	**139**	**133**	**153**	**79**	**54**	**852**

Figure IX shows the silhouette plots for the six-cluster and the seven-cluster classifications using 'Squared Euclidean' distance measure. There is no misclassification of hedge funds for Class 2. The misclassification (slight negative value) of Class 1 hedge fund is almost negligible. The seven-cluster classification appears to be a better choice, since the mean silhouette value for both the Class 1 and Class 2 hedge funds is closer to 0.8.

Figure IX. Silhouette Plots for Six- and Seven-cluster Classifications for Class 1 and 2 Funds.



VIII. Summary and Conclusion

Investment strategy and/or investment style are the basis of the classification scheme of hedge funds employed by different database providers. This classification varies from database to database. There is considerable variation in the definitions, return calculation methodologies, and assumptions. There exists a myriad of classifications, some overlapping and some mutually exclusive. Moreover, the source of information for the classification scheme is the questionnaire filled by the hedge fund manager. It is important to note that the strategy definitions themselves are sometimes not clear. At a certain time, a hedge fund manager may think that the fund's investment strategy matches with a particular category and, the same manager may think otherwise at a different time-period, although there may not have been any fundamental change in the strategy followed by the manager.

In this paper, hedge funds in the ZCM/Hedge database are classified using cluster analysis. The attributes used for classification in this paper are those that influence the return characteristics of the hedge fund. These attributes will affect the hedge fund return, but the return will not affect the classification scheme. The source of information for the attributes used in this classification scheme is also the hedge fund manager.

The cluster analysis is based on asset class, size of the hedge fund, incentive fee, risk-level, and liquidity of hedge funds. All attributes are converted into a uniform scale of measurement.

Similarities in attributes of interest are the basis of the formation of clusters for the hedge funds. The classification is done separately for hedge funds domiciled in US and those domiciled outside US. The study uses nonhierarchical K-means clustering method for classifying hedge funds. The range of values for the optimal number of clusters is decided by focusing on 'the knee of the curve' in the plot of distance measure versus the number of clusters. The randomization process for cluster seeds is repeated 100 to 1000 times for each number of clusters to avoid local minima. The robustness of the algorithm is examined by comparing the distance measure and mean silhouette value for different runs of the algorithm. The classification is based on optimization of the within cluster variance and minimization of misclassification of hedge funds.

The classification uses two distance measures, 'City-block' and 'Squared Euclidean'. The results show that the 'Squared Euclidean' measure is robust as compared to the 'City-block' measure. Distance measures and mean silhouette values for the six-cluster, seven-cluster, and nine-cluster classifications are compared. The seven-cluster classification performs better both in terms of optimizing the distance criterion and reducing misclassification. The new classification obtained using K-means clustering method is compared with the existing classification of the ZCM/Hedge database. The new classification has not kept intact any category of the existing classification. This suggests that the existing classification does not consider the attributes that this paper uses for classification. The attributes used in this paper do not have any subjective criteria that would change from manager to manager. The attributes will change only if the basic characteristics of the hedge fund change.

The new classification can be used to develop benchmarks for evaluating the performance of hedge funds. Hedge funds from other databases can be classified using this classification scheme. Further work needs to be done to determine the representative characteristics of a typical hedge fund for membership to a particular cluster, which will help in the identification of hedge funds to this new classification method.

References

Anderberg Michael R. (1982): "Cluster Analysis for Applications," *Academic Press.*

Brown Stephen J., Goetzmann William N. and Park James (2000): "Hedge funds and the Asian Currency Crisis of 1997," The Journal of Portfolio Management, Summer, p 95-101.

Babinec Tony (2002): "Editorial: Some musings on Cluster Analysis," Journal of Targeting, Measurement and Analysis for Marketing, Vol. 10, No. 4, p 301-303.

Caldwell Ted: "Classifying Hedge Funds: What's in a Name"? Investment Policy, Vol. 1, No. 3, p 30-41.

Das, Nandita, Kish, Richard J., Muething, David L. and Taylor Larry, W. (2002a): "Literature on Hedge Funds", Lehigh University Martindale Center Working Paper, 2002 Series # 2.

Das, Nandita, Kish, Richard J., Muething, David L. and Taylor Larry, W. (2002b): "An Overview of the Hedge Fund Industry", Lehigh University Martindale Center Working Paper, 2002 Series #1.

Das, Nandita (2000), "Hedge Funds: An Alternative Investment Option," A Report Prepared for SEI Investments.

Franklin R. Edwards (1999): "Hedge Funds and the Collapse of Long Term Capital Management", Journal of Economic Perspectives, Vol. 13, No. 2, Spring, p189-210.

Fung, William and Hsieh David A. (2001): "Benchmarks of hedge-fund performance: Information content and measurement bias", Financial Analyst's Journal, May.

Fung, William and Hsieh David A. (1997): "Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds", The Review of Financial Studies, Vol. 10, No. 2, Summer, p 275-302.

Goldman Sachs & Co. & FRM (2000): "Hedge Funds Revisited", Pensions and Endowment Forum, Jan.

Goldman Sachs & Co. & FRM (1998): "Hedge Funds Demystified: Their Potential Role in Institutional Portfolios", Pensions and Endowment Forum, July.

Hoppner Frank, Klawonn Frank, Kruese Rudolf, and Runkler Thomas (1999): "Fuzzy Cluster Analysis," *John Wiley and Sons Ltd.*

Johnson Richard A. and Wichern Dean W. (1998): "Applied Multivariate Statistical Analysis," *Prentice Hall.*

Kuo R. J., Ho L. M. and Hu C. M. (2002): "Cluster Analysis in Industrial Market Segmentation through Artificial Neural Network," Computers and Industrial Engineering, Vol. 42, p 391-399.

Lim G. C.: "Hedge Funds and Currency Crisis", Australian Economic Review, Vol. 32, No. 2, p191-196.

Lorr Maurice (1983): "Cluster Analysis for Social Scientists," *Jossey-Bass Publications.*

President's Working Group on Financial Markets (2000): "Sound Practices for Hedge Fund Managers".

Punj Girish and Stewart David W. (1983): "Cluster Analysis in Marketing Research: Review and Suggestions for Application," Journal of Marketing Research, March, Vol. 20, May, p 134-48.

Purcell David, and Paul Crowley (1999): "The Reality of Hedge Funds", The Journal of Investing, Fall, p 26-44.

Romesburg Charles H. (1988): "Cluster Analysis for Researchers," *Lifetime Learning Publications.*

Seber G. A. F. (1984): "Multivariate Observations," *John Wiley and Sons.*

Spath Helmuth (1985): "Cluster Dissection and Analysis," *Ellis Horwood Limited.*

Schneeweis, Thomas (1998): "Dealing with the myths of Hedge Fund Investment", The Journal of Alternative Investments, Winter, p 11-14.

Tarassenko Lionel (1998): "A Guide to Neural Computing Applications," *John Wiley and Sons Inc.*

Yago Glenn, Ramesh Lalita and Hochman Noah E. (1999) "Hedge Funds: Structure and Performance", The Journal of Alternative Investments, Summer.