Forming Stock Groups with a Cluster Analysis of Common Size Statements

Andy Fodor
Ohio University

Randy D. Jorgensen
Creighton University

John D. Stowe
Ohio University

Submitted to the 2015
Southwestern Finance Association
Annual Conference

JEL Classification: G11, G32, L00, M41

Keywords: Stock groups; Industries; Stock return comovement; Common size statements; Cluster analysis; Benchmarks

Forming Stock Groups with a Cluster Analysis of Common Size Statements

Abstract

Researchers often classify firms in various ways to test hypotheses about corporate finance or investments. Similarly, investors and analysts must classify firms (or portfolios) to conduct valuations or to evaluate performance. The primary method of forming industry groups is based on an economic analysis of the firm's production processes or of the market for the firm's final product or service. The comovement of stock returns is another method used to form industry groups. Our purpose is to form industry groups based on the structure of their financial statements. Using cluster analysis, a multivariate tool that can form groups where their characteristics are similar within groups and distinct across groups, we form clusters of stocks based on common size financial statements (percentage breakdowns of their balance sheets and income statements). We cluster the largest U.S. corporations and compare the financial properties of the resulting clusters and also compare the clusters with those of industrial groups formed by traditional methods. Because firms in the same industry (SIC) often have differing financial structures, and because firms with similar financial structures can come from very different industries, clusters based on financial statements can be a useful complement to, or substitute for, the usual industry classifications. We study the financial characteristics of the firms on each cluster, compare cluster and industry memberships, and predict stock return comovements (correlations) of all pairs of firms based on cluster and industry membership.

I. Introduction

This paper groups a large number of diverse U.S. companies according to the financial and operating characteristics expressed on their financial statements. We use cluster analysis, a multivariate statistical method, to form groups such that the members (firms) are relatively similar within groups and distinct between groups. The fundamental data that we use to form clusters are common size statement variables, percentage breakdowns of each firm's balance sheet and income statement.

Investors, analysts, managers, and economists routinely compare the performance of a firm to industry norms, peers, or benchmarks. Of course, how to form these groups is a fundamental economic problem. Economists form industries based on the use of common inputs used by firms, by the nature of the production process, or by the nature of the demand for the final product. For example, two firms can be classed in the same industry based on the price cross-elasticity of demand for their outputs. Federal agencies and financial services firms provide such industry groupings.[1]

We form clusters of 1,641 corporations that are relatively homogenous within clusters and distinct across clusters based on the structures of their financial statements. We provide considerable information about these clusters, including a comparison to industry classifications. Finally, we compare the stock return correlations within each cluster to correlations of stocks between clusters. Like prior research, the stock comovement within and between clusters shows important investment properties of the clusters. We examine the comovement of stock returns first using our financial statement clusters, second, using industry classifications, and finally using them simultaneously to establish any incremental value to firm clusters.

The following section discusses the intuition behind the study. Section III describes the sample and the data used in our analysis. It also outlines the statistical methods we employ. Section IV presents and interprets the empirical results. Section V summarizes and discusses.

II. Background

Stock classifications have been used in several ways by financial economists and investment managers.

- Investment performance evaluation of individual stocks is estimated relative to market, industry, or peer benchmarks.

[1] The commonly used classifications include the SIC, NAIC, GICS, and Fama-French systems.

- Management performance (and compensation) is often evaluated relative to a set of industry peers. If these peers have different asset, financing, and cost structures, the performance evaluations may be less relevant.
- Valuation of equities using financial multiples is based on comparisons to the multiples of otherwise similar stocks.
- Researchers often use matched pairs of firms as controls when they test hypotheses. The matched pair is often a firm in the same industrial classification and similar in size to the subject firm.
- Benchmarks or industries are used to analyze portfolio performance and to do an attribution analysis. They promise to be a useful device for the fundamental analysis used in holdings-based style portfolio analysis.

Stocks are characterized in various ways, sometimes by forming a natural grouping that may have some predictive power in explaining the dispersion of future returns. Mutual funds are also grouped by their investment objectives or the ‘style’ of their managers. Brown and Goetzmann (1997) use an empirical of manager ‘style’ that proves to be superior to industry classifications in predicting future performance, as well as past performance.

The kinds of information used to form groups can vary. For example, the two dominant ways to describe a portfolio are returns-based style analysis (RBSA) and holdings-based style analysis (HBSA). RBSA was first presented by Sharpe (1988) and it has been heavily used to characterize the portfolios of mutual funds and other managed funds. It is based on regressions of portfolio returns against a set of indexes. Holdings-based style analysis, in contrast to RBSA, is more of a “bottoms up,” fundamental approach to characterizing a portfolio. The data required for HBSA are descriptives of the stocks held in a portfolio. Often, instead of using several quantitative variables for each stock, the stocks classed into groups that have distinct characteristics. Industrial groups, such as SICs, are frequently used. Another popular such classification is Morningstar's style box, a two dimensional classification of size (small-, mid-, and large-cap) versus orientation (growth, core, and value). The financial clusters we form are an alternative way of trying to find stocks with common characteristics.

Industries are often defined economically, where firms have related final products/services and/or similar production processes. Examples of the economic classifications (used in the U.S.) include SIC codes, NAICS codes (which have largely replaced SICs), GICs, and the Fama-French industry classes. Comovements or correlations between stock returns have also been used to form industry groups (Elton and Gruber, 1970, Farrell, 1974). Our method is to form groups based on the multivariate makeup of firms' financial statements.

Bhojraj, Lee, and Oler (2003) compare the four industry classification systems, specifically SIC (Standard Industrial Classification), NAICS (North American Industry Classification System), GICS (Global Industry Classification Standard jointly developed by Morgan Stanley Capital International (MSCI) and S&P) and the Fama and French classifications. The latter is popular with academics and the others, especially GICS, are popular with practitioners. Bhojraj, et al. shows that the GICS classifications are better at explaining stock return comovements, valuation multiples, growth rates, R&D, and several financial ratios. The other three methods (SIC, NAICS, and Fama French) differ little from each other in most applications. Although our classification system, clustering firms based on common size financial statement data is very basic, it proves to have informational value on its own, and, importantly, to have incremental value if used as an adjunct to these other industry classification systems.

We use common size statement variables to characterize the firm's financial structure rather than financial ratios for practical reasons (Deakin, 1976; Pinches, Mingo, and Caruthers, 1973; Lev and Sunder, 1979). A basic issue with financial ratios is choosing the set of ratios for analysis as well as which definitions of a ratio to use, both of which can pre-ordain the results. Ratios are generally non-normally distributed and can be skewed and with extreme outliers. When the denominators of ratios include negative values, the distribution can be discontinuous and analysts often discard these observations as "not meaningful." Ratios often share similar variables, which induces spurious correlations across these ratios. While common size variables are not trouble free, their simplicity is an attractive property. The analyst has considerably less freedom to choose ratios, the denominators are (generally) positive, and their distributions are more regular than financial ratios. Our method is to form groups based on the multivariate makeup of firms' common size financial statements.

Users analyze common size statements, often called "vertical analysis," to summarize changes in the financial statement proportions over time for a given firm, or to characterize differences between firms or between a firm and its industry. Researchers have occasionally used common size variables. Stowe, Watson, and Robertson (1980) did a canonical correlation analysis between the two sides of the balance sheet. Frank and Goyal (2003) used common size statements to show changes over time, although they did not include the sets of common size variables in their statistical models.

Investors and analysts are very familiar with common size statements and have ready access to financial data (including common size statement data) from many sources such as Bloomberg, Morningstar, FactSet, Capital IQ, and the firms themselves. However, common size variables are generally used judgmentally and, excepting the examples above, have not been incorporated in a multivariate statistical analysis.

III. Data and Methods

The purpose of this research is to classify firms into groups according to the financial and operating characteristics exhibited in their financial statements. The results allow for the formation of groups with similar financial characteristics, and these groups will differ from the widely used SIC, NAIC, GIC, or Fama-French industry codes. This section also includes a discussion of the cluster analysis methods applied to our firms.

A. The Data and Sample

The financial statement data for this study are from Standard and Poor's Compustat database. Financial firms (in the SIC 6000's) were not included in our sample. Our clustering of firms is based on twenty-one common size statement variables, including six asset variables, eight liability/shareholder equity variables, and seven income statement variables. The balance sheet variables are expressed as a proportion of total assets, and the income statement variables are expressed as a proportion of total sales. In addition to the common size statement variables used to form the clusters, we use several other firm-specific variables to analyze the results, including additional descriptors such as firm size, betas and return volatility, financial ratios, returns correlations, and industry classifications.

Table 1 Here

The study includes all U.S. firms with a fiscal year ending in 2012. To focus on the more important firms, we exclude all firms with total assets or sales less than $200 million. In addition, firms had to have firm data available in both Compustat and CRSP. Table 1 provides the descriptive statistics in two panels. Panel A includes information on the 21 common size variables. Panel B includes several market statistics and financial ratios.

In Panel A, the means of the six asset common size variables sum to 1.000, as do the means of the eight liability/shareholders' equity variables and the seven income statement variables. The standard deviations of the variables and their values at several percentile points are also shown in the table. The dispersions of the 21 variables, as seen in their standard deviations and the differences between their values for alternative percentile points, are large and provide a good opportunity to cluster firms based on the variation in common size variables. The other descriptive variables shown in Panel B are used to further describe the clusters of firms that are formed with the common size variables. The variables in Table 1 will be discussed when we compare their values for the clusters we derive.

Table 2 Here

The correlations between the 21 common size variables are given in Table 2.   Cluster analysis will seek to find patterns among the great variety of financial statement structures of the sample firms. Because the common size variables for three sets of variables—assets, liabilities and equity, and income statement—each sum to 1.0, there is a tendency for the correlations between pairs of variables within each set to be negative. Nonetheless, there should be an economic basis for the correlations among variables.

Focusing first on those variables with a high and positive correlation (greater than 0.25), there are several meaningful relationships in the table. Cash is positively related to common equity and to selling/general/administrative expenses. Riskier firms who select higher levels of equity also tend to hold higher cash balances. Accounts payable levels are positively related to receivables, inventories, and cost of sales. Other current liabilities are associated with other current debt. Short-term assets are often financed with short-term obligations. Current assets and liabilities are part of the working capital cycle and intimately related. Not surprisingly, net fixed assets is positively related to depreciation and to interest expense, and interest expense is positively related to long-term debt and to depreciation.

Commenting on those variables with low correlations (less than -0.25), net fixed assets is negatively related to short-term assets such as cash, receivables, and inventory as well as with the short-term debt accruals and selling/general/administrative expenses in the income statement. Depreciation expense is negatively correlated with receivables and inventory and with accounts payable. Interest expense is negatively related to receivables, common shareholders equity, and net income. Selling/general/administrative expenses are negatively related to net fixed assets. Net income is also negatively related to cost of sales and with other expenses.

The nature of the firm leads to its asset structure, liability structure, and income statement structure. Reciprocally, the multivariate makeup of these financial variables should help describe the nature of that firm. Across the entire sample, there are many associations between pairs of common size variables that are economically meaningful. In this paper, we do not look at these variables singly or pairwise, but use cluster analysis is to look at the overall multivariate relationships. This method helps to identify overall patterns that are common within the clusters formed and yet distinct across these clusters.

B. Methodology

An iterative-partitioning, non-hierarchical cluster technique is used to form the company groups. The specific program used is SAS FASTCLUS, which performs the cluster analysis based on Euclidian distances and is an efficient algorithm for clustering large data sets. The procedure iteratively assigns firms to clusters that minimize the sum of squared distances of the variable means of the sample firms from their respective assigned cluster means. This study employs the K-means method, which assigns each firm to one and only one cluster, the cluster with the nearest centroid.

Cluster analysis results depend on the sample of firms used, the variables employed, and the choice of cluster technique and clustering constraints. Changing these will alter the results. For example, clustering based on common size balance sheet and income statement variables is fairly straightforward due to the normalization of the variables achieved in the common sizing process. Clustering based on financial ratios instead of common size variables would alter the groupings.[2] In this study, clustering is based only on the 21 common size variables. Although financial ratios and other firm-specific variables are not used to form the clusters, these other variables (listed in Table 1, Panel B) are used to help describe the characteristics of the various clusters. Because there is no uniquely correct product of clustering analysis, the number of clusters and the minimum cluster size are subject to the analyst's judgment.

The clusters formed are described by comparing the mean values of the variables across the clusters as well with the entire sample. The median values for the financial ratios as well as SIC breakdowns and market data are presented to assist the interpretation. The empirical results and interpretations are presented next.

IV. Empirical Results

With the K-means cluster procedure, the sample of firms first is arbitrarily segregated into K different groups. Then the firms are iteratively reassigned to other clusters if the reassignments reduce the total Euclidian distance, which is the sum of the squared distances of the firms' variable means from the means of their assigned clusters. This process continues until

[2] Because of their distributions, for example, outliers have a large influence when financial ratios are used.

a convergence criterion is satisfied, that additional reassignments would have a small impact on the total Euclidian distance.

A. Formation and Interpretation of Clusters

Table 3 Here

In Table 3, we present the empirical results from forming 25 clusters based on the 21 common size variables. Table 3 shows the number of members in each of the 25 clusters.[3] As the table shows, the largest two clusters have 254 and 231 members. At the other extreme, five clusters had only one member.[4] The firms in the single-member clusters were so unique that combining them with any other firms would have increased the total Euclidian distance of the sample firms from their assigned cluster means. In this paper we concentrate our interpretation and discussion on the largest ten clusters, which held 1,576 (or 96.0%) of the total sample of 1,641 firms. The smallest fifteen clusters held 65 firms (4.0%) of the sample firms, and the smallest ten clusters held only 14 firms (0.9% of the sample firms).

Table 4 Here

Table 4 shows the amount of the variation in each of the 21 common size variables that is accounted for by cluster membership. For example, the standard deviation of L8 common shareholders equity around its overall mean is 0.252, but the standard deviation of common shareholders equity around the companies' respective cluster means is 0.137. The R-square for a variable is the between group variance (total group variance minus the within group variance) divided by the total variance, or, alternatively, the R-square is the proportion of the total variance of a variable that can be accounted for by cluster membership. Accounting for cluster

[3] The naming of clusters is arbitrary, so we named them on their descending number of members. "Cluster 1" is the cluster with the largest number of members and "Cluster 25" has the smallest.

[4] We will generally ignore the small clusters in our statistical analysis. Although they may be interesting, their number of members is too small for hypothesis testing.

membership reduces the variance of shareholders equity (square of the standard deviations) by 0.708, or by 70.8%. In the final column, RSQ / (1 – RSQ) is the ratio of between-cluster variance to within-cluster variance, which for shareholders equity is 2.423. The total standard deviations, within group standard deviations, and R-squares for all 21 common size variables are shown in the table.

The variance-weighted R-square is 0.659, which is the fraction of the total variance of the 21 common size statement variables that is explained by cluster membership. If the variables are equal-weighted, the average R-square for all variables is 0.484.[5] Eleven of the 21 variables have R-squares of 0.50 or greater, while three of the variables have R-squares of less than 0.20. The variables with the highest R-squares, in descending order, are A5 net fixed assets, A4 other current assets, L5 long-term debt, L8 common shareholders equity, IS1 cost of sales, A6 other long-term assets, IS7 net income, and IS2 selling, general, and administrative expenses. Cluster membership accounts for at least 60% of the variance for all of these variables.

Table 5 Here

The following three tables, Tables 5, 6, and 7, provide information on the characteristics of the stocks assigned to each cluster. We will briefly comment on each table and then make summary comments about the clusters following the tables. Panel A of Table 5 provides the variable means (the cluster centroids) of the common size variables for the firms in the ten largest clusters. The firms within each cluster are closer to their own cluster centroid than that of any other cluster. The means for specific variables differ across clusters. For example, cash is highest for Clusters 6 and 7 and lowest for Cluster 1. Common equity is highest for Clusters 6

[5] The variance-weighted R-square is higher because the variables with higher variances tended to have higher R-squares.

and 5 and lowest for Clusters 9 and 10. Cost of sales is highest for Clusters 5, 4, and 3 and lowest for Cluster 7.

Panel B of Table 5 provides mean data for market equity, recent equity returns, the return standard deviation, beta, ROA and ROE for each cluster. The largest firms, based on market value of equity, tend to be in Clusters 1, 2, and 7, while the smallest firms are in Clusters 3 and 10. Clusters 1, 2, and 7 have low betas and low return standard deviations, while Cluster 10 (especially) and Clusters 4 and 9 have high values.[6] Cluster 1 firms had a high dividend yield while several clusters had low dividend yields (six clusters had yields below 0.02). Clusters 9 and 10 may exhibit a "dividend trap" where their high dividend yields are not because of profitability but due to poor fundamentals that depress their stock prices. With high book-to-market ratios, firms in Clusters 1 and 3 might be value companies while low ratios might indicate that the firms in Clusters 4, 6, and 7 are growth companies.

Table 6 Here

The 1,641 sample firms were classed according to their Fama-French 30 (FF) industry codes. The twenty FF codes with the largest numbers of sample firms are shown in rows of Table 6, with the remaining ten codes (116 firms) in the "other" row.[7] The Fama-French industries are listed in descending order based on their number of members in our sample. The columns show the largest ten clusters, with the "other" column including the remaining fifteen clusters that had 65 firms. There are interesting patterns. For example, for the 98 firms in the petroleum and natural gas industry (FF19), the majority (70) are in Clusters 1 and 8. Of the 90 public utilities (FF 20), 77 of them are in Cluster 1. Wholesale trade firms (FF 26) tended to fall in Clusters 4 and 5 (50 out of 70 firms). Without mentioning all examples, most of the time, the

[6] The betas average more than 1.0 because the small firms tend to have higher betas and our means are equal-weighted across firms, not market-weighted.

[7] A listing of the 30 Fama French industries is given in the appendix.

majority of the members of an industry located in one or two clusters. While there are obvious associations between financial structures (clusters) and industry type (FF classes), the relationships are not simple. This is not surprising given the complexities within an industry and the complexities of financial statements. Financial structures (represented by the clusters formed on financial statement variables) can differ a lot within an industry group. There can be diversity of asset, liability, and income structures for many industries. There are instances where at least 10% of the members of an industry code are spread out across four, five, or more clusters. Reciprocally, when looking at the columns, for a given cluster (financial structure), there can be a variety of industries represented.

We test the hypothesis of no relationship between the distribution of firms across clusters and industry classifications using a χ^2goodness of fit test. The test was performed for the data in Table 6 as a whole and for a subset of the table (top 6 clusters and top 12 industries). The p-values for the estimated χ^2's were extremely low and we can reject the null hypothesis and accept the alternate that the financial clusters and industry classes are significantly related.

Table 7 Here

Table 7 lists the names of the ten largest companies in each cluster based on total equity market capitalization.[8] Although the ten largest firms do not perfectly represent all members of a cluster, the largest members are the most interesting. Cluster 1 includes three integrated oil companies (Exxon, Chevron, and ConocoPhillips), six utilities (Duke, Exelon, Nextera, Southern Co, AEP, and PG&E), and a large retailer (Wal-Mart). The largest ten firms in Cluster 8 are mostly oil and gas companies or mining companies (Occidental, Apache, Anadarko, Marathon, Goldcorp, Newmont Mining, EOG Resources, Noble Resources, Plains Exploration, and Pioneer

[8] The ten largest clusters had substantially more than ten members, ranging from a high of 254 down to 63. The smaller clusters had fewer members—the smallest twelve clusters had less than ten members each.

Natural Resources). Cluster 7 includes tech and biotech firms among its top ten, including Pfizer, Johnson & Johnson, Microsoft, Merck, Cisco, Intel, Oracle, Abbott Labs, and Amgen).

Based on the differences across clusters of the means of the common size variables, financial ratios and other firm characteristics, industry classifications, and largest members, we can briefly characterize each cluster. Obviously, the 1,641 firms cannot be placed into a set of homogeneous clusters. However, each cluster should include a set of firms that share the traits of their assigned cluster mean more than they share the traits of the other cluster means.

For parsimony, we will not describe each cluster, but briefly will mention two of them as an example, specifically Cluster 1 and Cluster 7.

Cluster 1: The 254 firms in Cluster 1 had a high level net fixed assets (A5) and low current assets (A1 through A4). They had a high cost of sales (IS1) and levels of depreciation and interest expense (IS2 and IS3). These firms had a high dividend yield, low P/E ratio, and fairly low beta and return volatility. This cluster is heavily represented with firms that can be considered large-cap value stocks. As mentioned, this cluster includes several oil and gas companies, utilities, and transportation companies.

Cluster 7: The 138 firms in Cluster 7, compared to Cluster 1, had much more cash (A1) and relatively little net fixed assets (A5). This cluster used more equity financing (L8) and less long-term debt financing (L5). Cluster 7 had a very high level of selling, general, and administrative expenses (IS2) while Cluster 1 had a very low level. Cluster 7 included several tech and biotech firms. The industry memberships are very different for these two clusters. This cluster included firms in personal and business services, business equipment, and healthcare.

The fit between the clusters and industry classifications, though highly significant, is far from perfect. In many cases, the firms in an industry will fall on a single cluster, because firms in the industry have a similar financial structure. However, in many industries, the members will fall into different clusters. This means that firms in the same industry may have very different financial structures. Assuming that firms in the same industry are financially homogeneous is a clear mistake in such cases. A related point is that firms from different industries may fall in the same cluster. Even though these firms are in different industries, their financial structures are similar. Thus, we often assume firms are the same when they are not, and, conversely assume that firms are different when they are (financially, at least) very similar.

Table 8 Here

Table 2 included the distance between centroid for each cluster and its nearest neighbor. Table 8 gives the distances between all pairs of clusters. This is a crude measure of financial dissimilarity across clusters. For example, for Cluster 1, the nearest other centroid is that of Cluster 4 (distance = 0.485), and it is farthest away from Cluster 7 (distance = 1.015). For all possible pairs of clusters, the most distant pairs are Clusters 6 and 10 (distance = 1.110) and Cluster 7 and 10 (distance = 1.090). The closest pairs are Clusters 4 and 5 (distance = 0.365) and Clusters 3 and 5 (distance = 0.366).

B. Return Comovements of Clusters (Correlations)

Chan, Lakonishok, and Swaminathan (2007) compare the economic relatedness of firms within an industry and without an industry. One way to do this is to calculate the average of the pairwise correlations of the firms within an industry to the average correlations of the firms in an industry to firms outside the industry.   In their study, the correlations for raw returns are higher than for excess returns. Also, the correlations for large stocks are higher than for small stocks. In general, the inside-group correlations are higher than the outside-group correlations. We do a similar comparison of the return correlations for our clusters formed on common size statements and compare the results to similar information using Fama-French industry classifications.

The correlation matrix is based on monthly returns for 36 months, 2010-2012, which avoids returns from the recent financial crisis. To the extent that within-cluster correlations are higher than the outside cluster correlations, the clusters can be economically meaningful to investors. We provide this information in Tables 9 through 12 for the return comovement for pairs of stocks within and without the financial statement clusters and within and without the 30-industry Fama-French classifications. There are 1,641 stocks, so the number of unique correlations is $(1{,}641^2 - 1{,}641) / 2 = 1{,}365{,}620$, which is the number of observations in Tables 9 through 12.

Table 9 Here

Panel A of Table 9 shows the mean correlation between stocks within a cluster and between stocks in a cluster and stocks in other clusters (in mean and out mean). The difference between these is also given. Several of the clusters have a within-cluster correlation that is significantly greater than the correlation between the stocks in the cluster and other clusters. This is consistent with what researchers have found for industry groups. The overall mean difference between the within cluster average correlation and between cluster average correlation is 0.022.

Panel B of Table 9 gives similar information for the Fama French industries. The within-industry correlations are higher for stocks within an industry than between industries. For our sample firms, the within industry average correlation tends to be greater than the between industry average correlation. Overall, the in mean exceeds the out mean by 0.045, which is a significant and greater than the 0.022 difference for clusters.

Table 10 Here

Table 10 presents the average correlations for all combinations of firms cross the ten clusters and largest ten industry groups. This is a more detailed analysis of the average correlations given in Table 9. Panel A presents the average correlations between stocks in pairs of clusters. The average correlations on the diagonal are for correlations of stocks within a given cluster, and the other average correlations in the panel are between clusters. Panel C shows the difference between the average correlation between two clusters and the geometric average of the within-cluster average correlations of stocks in the two clusters. As Panel C shows, the between cluster correlations are consistently below those of the within-cluster correlations.[9]

The same analysis is done for the largest ten industry groups in Panels B and D. The same general pattern occurs, and the between cluster correlations are consistently lower than the within cluster correlations. The largest reductions in correlations tend to be concentrated in two industries, Industry 19 and Industry 20.

Based on Table 10, it is clear that the average correlation between the returns on two stocks from different clusters is less than the average of the correlations of the stocks in those two clusters. The same statement also applies to the average correlation for stocks from different industries.

Table 11 here

Table 11 provides the summary results of a regression analysis using dummy variables for the clusters or industries from which a pair of stocks belong. We use dummies for all pairings of the largest ten clusters (55 dummy variables) and for all possible pairs of the largest

[9] This comparison eliminates the positive differences observed in Table 9. In Table 9, a cluster with high (low) within cluster correlations will tend to have high (low) between cluster correlations. Table 10 compares the between cluster correlations to the correlations of the two different clusters from which the firms come. This eliminates a potential bias in our results (and from Chen et al (2007)).

ten FF industries (also 55 dummy variables). The sample size is 1,365,510 correlations, all possible unique correlations between the 1,621 companies in the sample. The table includes three models, model 1 using the cluster dummies as independent variables, model 2 using industry dummies as independent variables, and model 3 using both cluster and industry dummy R-square for model 1 (cluster variables only) is 0.0201 and the R-square for model 2 (industry variables only) is 0.0473. The F-values for both models show that the models are highly significant. Obviously, the industry dummies explain a higher proportion of the variance of the dependent variable (correlations) than the cluster variables.

Model 3, with both cluster and industry dummies, has the highest R-square (0.0617). Because our focus is on the value of the cluster variables, we test whether the increase in R-square from model 2 (industry variables only) to model 3 (both cluster and industry variables) is significant. An F-test for the significance of the additional variables in model 3 shows that the cluster variables have significant additional explanatory power over industry variables alone.[10] This suggests that the clusters formed using common size statements are valuable in explaining the correlations between stock returns.

V. Conclusion

In this study, the financial statements of 1,641 firms are used to cluster the firms into strategic groups according to their financial and operating characteristics. Based on the financial data analyzed, the 1,641 firms in the sample formed distinct clusters that are not segregated by the usual industrial classifications. For many firms, the structure of their financial statements

[10] The F-test for this question (ability of the additional variables to explain variation in the dependent variable) is $F_{q,n-k} = \frac{(R^2_{new} - R^2_{old})/q}{(1 - R^2_{new})/(n-k)}$, where n is the sample size, k the number of parameters in the new regression, and q is the number of additional parameters in the new regression. For our regression, n = 1,365,620, k = 110, and q = 55, and the computed F-value is 381.02. The critical F-value of a probability = 0.001 is 0.512, so we reject the hypothesis that the cluster variables do not help to explain the variation in stock correlations.

more closely resembles the other firms in their clusters than that of other firms in their own industry that fell into other clusters.

The basic comparison of the firms within the various clusters is undertaken via comparison of the mean values of the 21 variables used for cluster formation. Further interpretation of the financial characteristics of each cluster is enhanced by the use of selected additional financial ratios (not used in the cluster formation process), industry data, and other market data. Finally, the financial characteristics and industry membership are significantly different across clusters.

The second method of evaluating the stock clusters is by their ability to explain the correlations between firms in different clusters. We find that cluster membership explains a significant amount of the variation in the correlations between the returns on different stocks. We compared the ability of clusters and industries to explain these returns correlations, and find that both are fairly strong. Importantly, we find that while clusters and industries may have much redundant information, they both have significant incremental ability (over the other) to account for the correlations between stock returns. Hence, clusters formed based on financial statement information may have potential in many applications.

The financial cluster analysis that we study is both fairly easy to do and, as we have shown, can have incremental information compared to industries formed by other methods. Likewise, researchers or analysts seeking to find peer firms can include the structure of financial statements to be a criterion in selecting peer or benchmark firms. If a firm’s structure and that of its industry have a great diversity of financial and operating characteristics, our approach may have value. Cluster analysis allows for comparisons of firms with similar financial characteristics, thereby deriving more appropriate comparisons. External analysts and researchers as well as internal company managers may find this cluster analysis beneficial.

Financially, the firms within a cluster more closely resemble their own cluster means than those of any other cluster. In many cases, their financial makeup looks more like that of firms from other industries than many firms in their own industry that were assigned to other clusters. Managers make many strategic decisions in managing their firms, involving real investing decisions as well as financial ones. Many of these financial decisions such as those concerning capital structure, dividend policy, corporate liquidity, and working capital policies are not the fundamental basis for assigning a firm to an industry group. Thus, firms with various financial policies may be lumped together. An advantage of the methods employed here is that to the extent that operating and financial policies are communicated on financial statements, these policies are reflected in the financial clusters that we form. In such cases, the financial clusters we form can enhance any analysis where a firm is compared to an industry or a peer.

References

Barberis, Nicholas, and Andrei Shleifer (2003) Style investing. *Journal of Financial Economics* 68(2), 161–199.

Bhojraj, Sanjeev, and Charles M. C. Lee (2002) Who is my peer? A valuation based approach to the selection of comparable firms. *Journal of Accounting Research* 40(2), 407–439.

Bhojraj, Sanjeev, Charles M. C. Lee, and Derek K Oler (2003) What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41(5), 745–774.

Brown, Stephen J., and William N. Goetzmann (1997) Mutual fund styles. *Journal of Financial Economics* 43(3) 373–399.

Chan, Louis K.C., Josef Lakonishok, and Bhaskaran Swaminathan, (2007) Industry classifications and return comovement. *Financial Analysts Journal* 62(6), 56–70.

Deakin, Edward B. (1976) Distributions of financial accounting ratios: Some empirical evidence. *The Accounting Review* 51(1), 90–96.

Elton, Edward J., and Martin J. Gruber. (1970) Homogeneous groups and the testing of economic hypotheses. *Journal of Financial and Quantitative Analysis* 4(5), 581–602.

Fama, Eugene F., and Kenneth R. French. (1997) Industry costs of equity. *Journal of Financial Economics* 43(2), 153–193.

Farrell, Jr., J. L. (1974) Analyzing covariation of returns to determine homogeneous stock groupings. *Journal of Business* 47(2), 186–207.

Frank, Murray Z, and Vidhan K. Goyal. (2003) Testing the pecking order theory of capital Structure. *Journal of Financial Economics* 67(2), 217–258.

Gupta, Manak C., and Ronald J. Huefner (1972) A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research* 10(1), pp. 77–95.

Kahle, Kathleen M., and Ralph A. Walkling. (1996) The impact of industry classifications on financial research. *Journal of Financial and Quantitative Analysis* 31(3), 309–335.

Lev, Baruch, Shyam Sunder. (1979) Methodological issues in the use of financial ratios. *Journal of Accounting and Economics* 1(3), 187–210.

Liu, Jing, Doron Nissim, and Jacob Thomas (2002) Equity valuation using multiples. *Journal of Accounting Research* 40(1), 135–172.

Pinches, George E., Kent A. Mingo, and J. K. Caruthers. (1973) The stability of financial patterns in industrial organizations. *Journal of Finance* 28(2), 389–396.

Sharpe, William F. (1988) Determining a fund's effective asset mix. *Investment Management Review* 2(6), 59–69.

Sharpe, William F. (1992) Asset allocation: Management style and performance measurement. *The Journal of Portfolio Management* 18(2), 7–19

Stowe, John D., Collin J. Watson, Terry D. Robertson. (1980) Relationships between the two sides of the balance sheet: A canonical correlation analysis. *Journal of Finance* 35(4), 973–980.

Table 1: Descriptive Statistics

Table 1 presents descriptive statistics for the full sample. Means, medians, standard deviations, and percentile 1, 5, 95, and 99 values are presented. In Panel A, descriptive statistics are presented for common size variables. Values of balance sheet variables are expressed as a percentage of total assets. Values of incomes statement variables are expresses as a percentage of total sales. Variables are classified as 'Type' either A, L, or IS (asset, liability, or income statement). In Panel B, descriptive statistics are presented for market equity, cumulative returns (2011), standard deviation of returns (2010-2012), beta (2010-2012), return on assets and return on equity.

Panel A: Descriptive Statistics for Common Size Variables

Type	Variable	Mean	Median	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl
A1	Cash	0.137	0.091	0.141	0.000	0.004	0.437	0.615
A2	Accounts Receivable	0.128	0.107	0.103	0.000	0.016	0.317	0.509
A3	Inventories	0.109	0.069	0.129	0.000	0.000	0.370	0.622
A4	Other Current Assets	0.026	0.029	0.105	-0.668	0.005	0.087	0.161
A5	Net Fixed Assets	0.301	0.212	0.251	0.010	0.030	0.801	0.910
A6	Other Assets	0.299	0.253	0.218	0.007	0.024	0.708	0.916
L1	Accounts Payable	0.073	0.051	0.073	0.003	0.008	0.217	0.375
L2	Accruals	0.054	0.048	0.047	0.000	0.000	0.137	0.205
L3	Notes Payable	0.010	0.000	0.037	0.000	0.000	0.055	0.178
L4	Other Current Debt	0.067	0.045	0.090	-0.094	0.001	0.222	0.370
L5	Long-Term Debt	0.236	0.209	0.222	0.000	0.000	0.619	0.907
L6	Other Debt	0.118	0.093	0.105	0.003	0.014	0.312	0.472
L7	Preferred Stock	0.003	0.000	0.025	0.000	0.000	0.002	0.080
L8	Common Stock	0.438	0.451	0.252	-0.327	0.063	0.791	0.887
IS1	Cost of Sales	0.617	0.645	0.243	0.098	0.200	0.897	0.964
IS2	Sell/Gen/Admin Exp	0.209	0.175	0.170	0.000	0.000	0.519	0.721
IS3	Depreciation	0.066	0.041	0.082	0.003	0.009	0.210	0.415
IS4	Interest Expense	0.028	0.012	0.053	0.000	0.000	0.113	0.198
IS5	All Other Income	0.012	0.002	0.088	-0.165	-0.043	0.101	0.376
IS6	Income Taxes	0.023	0.022	0.058	-0.179	-0.021	0.086	0.146
IS7	Net Income	0.045	0.053	0.186	-0.519	-0.133	0.222	0.388

Panel B: Descriptive Statistics for Performance Variables

Variable	Mean	Median	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl
Market Equity (MM)	7.02	1.46	23.82	0.05	0.14	27.17	114.91
Return	-0.023	-0.026	0.313	-0.719	-0.531	0.473	0.851
Return STD	0.107	0.100	0.048	0.033	0.044	0.190	0.265
Beta	1.315	1.258	0.664	0.107	0.340	2.448	3.244
ROA	0.044	0.046	0.114	-0.296	-0.094	0.165	0.274
ROE	0.100	0.107	2.421	-2.231	-0.354	0.413	1.609
Dividend Yield	0.020	0.006	0.085	0.000	0.000	0.071	0.141
Earnings Yield	-0.050	0.051	2.443	-1.134	-0.261	0.129	0.271
Book-to-Market	0.477	0.493	3.318	-1.423	0.072	1.355	2.295

Table 2: Correlations

Table 2 presents correlations of Table 1 Panel A(B) variables in Panel A(B). Correlations significant at the 10% level or better are bolded.

Panel A: Common Size Variable Correlations

	Cash	AR	Inv	OCA	NFA	OA	AP	Acc	Notes	OCD	LTD	OD	PS	CS	COS	SGE	Dep	Int	Oth	IT	NI
Cash	1.00																				
Accounts receivable	-0.03	1.00																			
Inventories	**-0.04**	**0.05**	1.00																		
Other current assets	0.02	-0.01	**-0.25**	1.00																	
Net fixed assets	**-0.40**	**-0.41**	**-0.25**	-0.01	1.00																
Other assets	**-0.16**	-0.01	**-0.18**	**-0.33**	**-0.54**	1.00															
Accounts payable	**-0.09**	**0.44**	**0.43**	0.03	**-0.15**	**-0.25**	1.00														
Accruals	**0.16**	**0.21**	**0.10**	**0.12**	**-0.26**	-0.02	**0.08**	1.00													
Notes payable	**-0.11**	**0.05**	**0.20**	**-0.05**	-0.01	-0.04	0.03	**-0.06**	1.00												
Other current debt	**0.18**	**0.06**	**-0.16**	**0.39**	**-0.18**	-0.03	**-0.05**	-0.04	**-0.07**	1.00											
Long-term debt	**-0.33**	**-0.22**	**-0.18**	**-0.14**	**0.29**	**0.16**	**-0.17**	**-0.12**	**-0.04**	**-0.13**	1.00										
Other debt	**-0.24**	**-0.10**	**-0.10**	**-0.20**	**0.23**	**0.10**	**-0.06**	**-0.06**	0.01	**-0.13**	**0.09**	1.00									
Preferred stock	-0.01	-0.01	0.01	-0.04	0.03	-0.01	0.01	-0.04	0.00	0.00	0.04	0.03	1.00								
Common stock	**0.33**	**0.05**	**0.08**	**0.05**	**-0.19**	**-0.09**	**-0.12**	**-0.05**	**-0.09**	**-0.15**	**-0.79**	**-0.42**	**-0.15**	1.00							
Cost of sales	**-0.31**	**0.20**	**0.21**	**-0.09**	**0.18**	**-0.18**	**0.32**	0.01	**0.10**	**-0.13**	**0.11**	**0.13**	0.03	**-0.22**	1.00						
Sell/gen/admin exp	0.48	-0.03	0.00	0.11	-0.47	0.19	-0.14	0.18	-0.08	0.23	-0.22	-0.25	-0.03	**0.23**	**-0.65**	1.00					
Depreciation	**-0.20**	**-0.33**	**-0.35**	-0.01	**0.54**	**-0.12**	**-0.25**	**-0.23**	**-0.08**	**-0.10**	**0.28**	**0.08**	0.03	**-0.11**	**-0.18**	**-0.07**	1.00				
Interest expense	**-0.20**	**-0.29**	**-0.23**	**-0.10**	**0.36**	**0.04**	**-0.23**	**-0.21**	-0.03	**-0.11**	**0.53**	**0.07**	**0.07**	**-0.35**	**-0.05**	**-0.10**	**0.56**	1.00			
All other income	**-0.06**	0.01	**-0.05**	**0.06**	**-0.04**	**0.09**	-0.02	0.03	-0.03	**0.07**	**0.07**	**0.04**	-0.03	**-0.09**	-0.02	0.02	**-0.07**	0.01	1.00		
Income taxes	**0.10**	0.00	**-0.09**	**0.10**	-0.01	**-0.04**	**-0.07**	-0.01	0.00	0.03	**-0.13**	0.02	-0.03	**0.12**	**-0.32**	0.03	**-0.06**	**-0.13**	**-0.19**	1.00	
Net income	**0.11**	-0.01	-0.01	-0.01	**-0.12**	**0.08**	**-0.08**	-0.02	0.01	-0.01	**-0.21**	-0.03	-0.03	**0.23**	**-0.52**	-0.02	**-0.25**	**-0.33**	**-0.37**	**0.23**	1.00

Panel B: Performance Variable Correlations

	ME	RET	STD	Beta	ROA	ROE	DY	EY	B/M
Market Equity (MM)	1.00								
Return	**0.11**	1.00							
Return STD	**-0.23**	**-0.33**	1.00						
Beta	**-0.16**	**-0.36**	**0.72**	1.00					
ROA	**0.12**	**0.38**	**-0.21**	**-0.16**	1.00				
ROE	0.01	**0.05**	-0.01	0.00	**0.05**	1.00			
Dividend Yield	**0.01**	**-0.01**	**0.01**	**-0.01**	**-0.07**	**0.00**	1.00		
Earnings Yield	**0.01**	**0.06**	**-0.12**	**-0.07**	**0.18**	**0.00**	**-0.87**	1.00	
Book-to-Market	-0.01	**0.00**	-0.09	-0.06	**0.09**	-0.01	-0.86	**0.97**	1.00

Table 3: Cluster Summary Statistics

Table 3 presents descriptive statistics for the forming of the 25 clusters. The number of firms in each cluster, root mean square standard deviation, maximum distance from seed to observation for each cluster, the cluster nearest the cluster for which statistics are presented and the average distance between cluster centroids.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	254	0.072	0.825	8	0.483
2	231	0.078	0.989	5	0.440
3	218	0.080	0.779	5	0.366
4	192	0.085	0.745	5	0.366
5	178	0.087	1.082	4	0.366
6	166	0.080	0.852	7	0.426
7	138	0.097	1.035	6	0.426
8	71	0.089	0.870	1	0.483
9	65	0.112	1.175	2	0.546
10	63	0.108	1.120	4	0.548
11	15	0.142	1.162	8	0.682
12	13	0.090	0.855	13	0.874
13	11	0.126	0.838	2	0.842
14	5	0.119	0.576	4	0.977
15	4	0.137	0.624	2	0.853
16	3	0.135	0.609	6	0.796
17	3	0.139	0.613	16	0.971
18	2	0.150	0.486	10	1.524
19	2	0.171	0.555	24	1.191
20	2	0.189	0.612	11	1.325
21	1	.	0	18	2.379
22	1	.	0	24	2.093
23	1	.	0	14	1.658
24	1		0	19	1.191
25	1	.	0	11	1.741

Table 4: Cluster Variable Descriptive Statistics

In Table 4, for each common size variable in Table 1, Panel 1 presents the total standard deviation, standard deviation when the variable in pooled within clusters, R-squared for predicting the variable from the cluster, and the ratio of between-cluster variance to within cluster variance RSQ/(1 – RSQ). Also presented are the overall statistics Pseudo F-Statistic, Approximate Expected Overall R-Squared, and Cubic Clustering Criterion.

Type	Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
A1	Cash	0.141	0.100	0.503	1.013
A2	Accounts Receivable	0.103	0.083	0.364	0.571
A3	Inventories	0.129	0.094	0.484	0.939
A4	Other Current Assets	0.105	0.050	0.780	3.554
A5	Net Fixed Assets	0.251	0.113	0.800	4.011
A6	Other Assets	0.218	0.122	0.693	2.255
L1	Accounts Payable	0.073	0.063	0.266	0.363
L2	Accruals	0.047	0.045	0.109	0.122
L3	Notes Payable	0.037	0.037	0.032	0.033
L4	Other Current Debt	0.090	0.077	0.276	0.382
L5	Long-Term Debt	0.222	0.114	0.741	2.860
L6	Other Debt	0.105	0.085	0.365	0.576
L7	Preferred Stock	0.025	0.025	0.027	0.028
L8	Common Stock	0.252	0.137	0.708	2.423
IS1	Cost of Sales	0.243	0.126	0.733	2.739
IS2	Sell/Gen/Admin Exp	0.170	0.105	0.620	1.629
IS3	Depreciation	0.082	0.052	0.596	1.478
IS4	Interest Expense	0.053	0.034	0.596	1.473
IS5	All Other Income	0.088	0.071	0.359	0.559
IS6	Income Taxes	0.058	0.044	0.425	0.740
IS7	Net Income	0.186	0.101	0.710	2.451
	Overall	0.147	0.086	0.659	1.935
	Pseudo F Statistic			130.28	
	Approximate Expected Overall R-Squared			0.484	
	Cubic Clustering Criteria			79.65	

Table 5: Cluster Variable Means

In Panel A, Table 5 presents mean values of the common size variables for the firms in the ten largest clusters. Panel B presents the mean values for several other financial variables for these same clusters of firms.

Panel A: Common Size Variable Cluster Means

Type	Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
-	Number of Members	254	231	218	192	178	166	138	71	65	63
A1	Cash	0.047	0.068	0.151	0.089	0.137	0.352	0.285	0.056	0.085	0.103
A2	Accounts Receivable	0.071	0.122	0.131	0.152	0.269	0.131	0.105	0.043	0.097	0.093
A3	Inventories	0.041	0.068	0.253	0.191	0.120	0.105	0.035	0.012	0.026	0.064
A4	Other Current Assets	0.022	0.033	0.040	0.040	0.048	0.047	0.056	0.008	0.024	0.032
A5	Net Fixed Assets	0.665	0.128	0.299	0.266	0.123	0.151	0.101	0.789	0.141	0.474
A6	Other Assets	0.154	0.581	0.126	0.263	0.303	0.214	0.418	0.092	0.627	0.234
L1	Accounts Payable	0.050	0.047	0.107	0.113	0.141	0.051	0.031	0.041	0.032	0.069
L2	Accruals	0.033	0.049	0.058	0.066	0.076	0.064	0.061	0.022	0.056	0.059
L3	Notes Payable	0.010	0.007	0.012	0.024	0.010	0.005	0.005	0.002	0.005	0.006
L4	Other Current Debt	0.052	0.063	0.053	0.068	0.095	0.067	0.137	0.022	0.076	0.064
L5	Long-Term Debt	0.297	0.251	0.070	0.309	0.110	0.033	0.119	0.354	0.598	0.685
L6	Other Debt	0.178	0.117	0.082	0.146	0.091	0.049	0.084	0.131	0.134	0.113
L7	Preferred Stock	0.002	0.000	0.002	0.004	0.001	0.000	0.000	0.000	0.007	0.015
L8	Common Stock	0.378	0.467	0.615	0.270	0.477	0.728	0.563	0.427	0.092	-0.010
IS1	Cost of Sales	0.746	0.579	0.700	0.738	0.784	0.458	0.234	0.352	0.445	0.651
IS2	Sell/Gen/Admin Exp	0.054	0.237	0.182	0.159	0.132	0.349	0.537	0.117	0.258	0.168
IS3	Depreciation	0.085	0.050	0.036	0.034	0.023	0.044	0.061	0.263	0.096	0.082
IS4	Interest Expense	0.038	0.021	0.004	0.019	0.006	0.003	0.011	0.083	0.098	0.064
IS5	All Other Income	0.005	0.026	0.002	0.014	0.015	0.006	0.020	-0.012	0.074	0.025
IS6	Income Taxes	0.025	0.027	0.025	0.016	0.016	0.038	0.036	0.056	0.024	0.008
IS7	Net Income	0.047	0.060	0.050	0.020	0.025	0.101	0.100	0.141	0.005	0.002

Panel B: Performance Variable Cluster Means

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Market Equity	10.14	11.24	2.86	4.94	3.07	4.12	16.19	8.59	6.13	2.32
Return	0.027	0.024	-0.016	-0.093	-0.041	0.010	0.011	0.014	-0.087	-0.154
Return STD	0.088	0.089	0.112	0.118	0.106	0.112	0.099	0.103	0.128	0.148
Beta	1.110	1.151	1.334	1.533	1.411	1.356	1.114	1.304	1.431	1.713
ROA	0.030	0.043	0.069	0.024	0.041	0.102	0.061	0.039	0.005	0.012
ROE	0.079	0.097	0.113	0.114	0.091	0.142	0.121	0.098	0.743	-0.866
Dividend Yield	0.029	0.017	0.012	0.017	0.013	0.016	0.009	0.024	0.036	0.035
Earnings Yield	0.017	0.029	0.042	-0.021	0.030	0.043	0.031	0.046	-0.014	-0.140
Book-to-Market	0.784	0.587	0.712	0.534	0.673	0.473	0.422	0.592	0.182	-0.187

Table 6: Cluster Members by Industry

Table 6 shows the number of members in each cluster that belong to Fama-French 30 industries. Only the top 10 clusters and top 20 industries are presented specifically.

Industry	Cluster											Total
	1	2	3	4	5	6	7	8	9	10	Other	
22	6	41	4	4	43	26	55	7	15	4	5	210
23	4	33	27	13	16	66	32	0	3	2	2	198
27	11	11	59	25	9	7	1	0	1	6	2	132
8	5	33	3	7	2	21	36	0	6	7	2	122
19	33	2	7	1	2	0	0	37	0	1	15	98
20	77	1	0	1	0	0	0	8	0	0	3	90
25	44	1	4	3	8	0	1	5	2	3	1	72
26	2	3	6	21	29	2	2	0	1	2	2	70
13	2	11	23	13	9	8	0	0	0	1	1	68
21	1	17	0	1	3	2	3	5	18	12	6	68
11	4	6	8	8	19	1	0	0	1	2	13	62
9	6	4	11	21	4	5	0	0	0	0	1	52
1	0	16	14	13	2	2	0	0	2	2	0	51
24	10	2	4	15	5	0	0	0	1	1	1	39
15	1	1	10	11	8	1	0	0	0	1	4	37
4	8	3	4	1	1	2	1	2	2	11	0	35
28	16	1	4	5	0	0	0	2	1	3	1	33
12	4	3	9	10	5	0	0	0	1	0	0	32
6	0	8	4	3	2	5	2	0	3	1	0	28
7	0	7	5	2	1	12	0	0	0	1	0	28
Other	20	27	12	14	10	6	5	5	8	3	6	116
Total	254	231	218	192	178	166	138	71	65	63	65	1641

Table 7: Largest Firms in Cluster

Table 7 presents the ten largest companies in each of the 10 largest clusters based on market capitalization. Within clusters companies are listed in descending order from largest to smallest.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Exxon Mobil Corp	AT&T Inc	Archer-Daniels-Midland	Hewlett Packard Co
Chevron Corp	Verizon	Walgreen Co	Caterpillar Inc
Wal-Mart Stores Inc	Apple Inc	Halliburton Co	Boeing Co
ConocoPhillips	Comcast Corp	Costco Wholesale	Dow Chemical
Duke Energy Corp	Procter & Gamble Co	Baker Hughes Inc	Du Pont
Exelon Corp	Intl Business Machines	Mosaic Co	Delta Air Lines Inc
Nextera Energy Corp	United Technologies	Micron Technology	United Contl Holdings Inc
Southern Co	Mondelez Int	Western Digital	Amazon.com Inc
AEP	Disney (Walt) Co	Cummins Inc	Intl Paper Co
PG&E Corp	Pepsico Co	Tyson Foods Inc -Cl A	Macy's Inc
Cluster 5	Cluster 6	Cluster 7	Cluster 8
Schlumberger Ltd	Google Inc	Pfizer Inc	Occidental Petroleum
Dell Inc	Corning Inc	Johnson & Johnson	Apache Corp
Honeywell Intl	Monsanto Co	Microsoft Corp	Anadarko Petroleum
General Dynamics	Nike Inc -Cl B	Merck & Co	Marathon Oil Corp
McKesson Corp	Sandisk Corp	Cisco Systems Inc	Newmont Mining Corp
National Oil Well Varco	CF Industries Holdings	Coca-Cola Co	EOG Resources Inc
Johnson Controls Inc	Lam Research Corp	Intel Corp	Las Vegas Sands Corp
Automatic Data Proc	Priceline.com Inc	Oracle Corp	Noble Energy Inc
Bunge Ltd	Cognizant Tech Solns	Abbott Laboratories	Plains Exploration
Cardinal Health Inc	Marvell Technology Gp	Amgen Inc	Pioneer Nat Resources
Cluster 9	Cluster 10		
Sprint Nextel Corp	MGM Resorts Intl		
Time Warner Cable	DIRECTV		
Liberty Global Inc	Frontier Comm		
Philip Morris Intl	Dish Network Corp		
Altria Group Inc	Community Health Sys		
Crown Castle Int	Avis Budget Group		
Davita Healthcare Part	Level 3 Communications		
Kellogg Co	Supervalu Inc		
American Tower Corp	United Rentals Inc		

Table 8: Cluster Centroids

Table 8 presents distances between cluster centroids for the ten largest clusters (in terms of number of member companies).

Cluster	1	2	3	4	5	6	7	8	9	10
1	0.000									
2	0.743	0.000								
3	0.582	0.597	0.000							
4	0.485	0.467	0.461	0.000						
5	0.669	0.440	0.366	0.365	0.000					
6	0.869	0.607	0.447	0.718	0.560	0.000				
7	1.015	0.566	0.745	0.812	0.752	0.426	0.000			
8	0.483	0.908	0.797	0.781	0.952	0.924	0.984	0.000		
9	0.905	0.546	0.993	0.641	0.835	1.008	0.829	0.984	0.000	
10	0.616	0.822	0.931	0.548	0.875	1.110	1.090	0.762	0.585	0.000

Table 9: Correlation Differences

Table 9 presents average correlations within clusters (both firms in the same cluster) and outside clusters (one firm in a cluster and another outside the cluster) in Panel A. Average correlations within industries (both firms in the same industry) and outside industries (one firm in an industry and another outside the industry) are presented in Panel B. Differences are also presented.

Panel A: Cluster Correlation Differences

Cluster	In Mean	Out Mean	Difference	
1	0.334	0.309	0.026	***
2	0.361	0.337	0.024	***
3	0.327	0.320	0.007	***
4	0.363	0.339	0.024	***
5	0.400	0.351	0.049	***
6	0.334	0.322	0.012	***
7	0.291	0.298	-0.008	***
8	0.380	0.323	0.057	***
9	0.297	0.303	-0.007	*
10	0.295	0.301	-0.006	
All	0.344	0.322	0.022	***

Panel B: Industry Correlation Differences

Ind	In Mean	Out Mean	Difference	
8	0.290	0.274	0.016	***
13	0.509	0.390	0.118	***
19	0.464	0.340	0.125	***
20	0.420	0.235	0.185	***
21	0.318	0.308	0.010	**
22	0.334	0.324	0.010	***
23	0.408	0.344	0.064	***
25	0.360	0.323	0.037	***
26	0.368	0.341	0.027	***
27	0.283	0.283	0.000	
All	0.367	0.322	0.045	***

Table 10: Average Correlations within and between Clusters and Industries

Table 10 presents average correlations for pairs firms grouped by clusters (Panel A) and industries (Panel B). The average correlations on the diagonal are the average of the correlations when both firms are within the same cluster (industry). The other average correlations are those between firms in cluster i (industry i) and firms in another cluster j (industry j), where i ≠j. Panels C and D present the differences between the mean correlations between companies in two different clusters (industries) and the average correlation of firms within those two clusters (industries). For example, in Panels C and D, for clusters 1 and 2, the difference is the average correlation between firms in cluster 1 and cluster 2 minus the geometric average of the correlations of firms within cluster 1 and within cluster 2.

Panel A	Cluster j										
		1	2	3	4	5	6	7	8	9	10
Cluster i	1	0.334									
	2	0.322	0.360								
	3	0.303	0.335	0.328							
	4	0.325	0.358	0.339	0.363						
	5	0.336	0.373	0.356	0.376	0.400					
	6	0.299	0.339	0.324	0.339	0.358	0.333				
	7	0.273	0.316	0.288	0.308	0.321	0.302	0.290			
	8	0.330	0.330	0.319	0.341	0.356	0.316	0.281	0.382		
	9	0.293	0.319	0.293	0.319	0.330	0.295	0.272	0.304	0.294	
	10	0.279	0.316	0.296	0.321	0.327	0.299	0.270	0.286	0.287	0.294

Panel B	Industry j										
		8	13	19	20	21	22	23	25	26	27
Industry i	8	0.290									
	13	0.307	0.509								
	19	0.278	0.443	0.464							
	20	0.190	0.279	0.278	0.420						
	21	0.273	0.364	0.321	0.254	0.318					
	22	0.285	0.391	0.340	0.224	0.308	0.334				
	23	0.299	0.435	0.361	0.223	0.329	0.359	0.408			
	25	0.246	0.409	0.325	0.260	0.300	0.320	0.353	0.360		
	26	0.286	0.423	0.367	0.241	0.321	0.343	0.377	0.349	0.368	
	27	0.239	0.334	0.287	0.198	0.271	0.283	0.302	0.282	0.297	0.283

Table 10, continued

Panel C	Cluster j										
		1	2	3	4	5	6	7	8	9	10
Cluster i	1	0.334									
	2	-0.025	0.360								
	3	-0.028	-0.008	0.328							
	4	-0.024	-0.003	-0.006	0.363						
	5	-0.030	-0.007	-0.006	-0.005	0.400					
	6	-0.035	-0.007	-0.007	-0.009	-0.008	0.333				
	7	-0.039	-0.007	-0.020	-0.016	-0.020	-0.008	0.290			
	8	-0.027	-0.041	-0.035	-0.031	-0.035	-0.041	-0.052	0.382		
	9	-0.021	-0.006	-0.018	-0.008	-0.013	-0.018	-0.020	-0.031	0.294	
	10	-0.034	-0.010	-0.014	-0.006	-0.016	-0.014	-0.022	-0.049	-0.007	0.294

Panel D	Industry j										
		8	13	19	20	21	22	23	25	26	27
Industry i	8	0.290									
	13	-0.077	0.509								
	19	-0.089	-0.043	0.464							
	20	-0.159	-0.183	-0.164	0.420						
	21	-0.031	-0.038	-0.063	-0.111	0.318					
	22	-0.026	-0.021	-0.053	-0.151	-0.018	0.334				
	23	-0.045	-0.021	-0.075	-0.191	-0.032	-0.010	0.408			
	25	-0.077	-0.019	-0.084	-0.129	-0.039	-0.027	-0.030	0.360		
	26	-0.041	-0.010	-0.046	-0.153	-0.021	-0.007	-0.010	-0.015	0.368	
	27	-0.047	-0.045	-0.075	-0.147	-0.028	-0.024	-0.038	-0.037	-0.025	0.283

Table 11. Regressions Predicting Stock Correlations with Cluster and Industry Membership

Table 11 presents regression results where the dependent variable is the correlation between two firms' returns. The first set of independent variables includes dummy variables for each possible pair of the 10 largest clusters. There are 55 unique cluster pair dummies. The second set of independent variables includes dummy variables for each possible pair of the 10 industries with the most members. There are 55 unique industry pair dummies. Model 1 includes the cluster dummies only, model 2 includes the FF industry dummies only, and model 3 includes both sets of dummy variables. The coefficients of the dummy variables are not shown for parsimony. For each model, and F-test rejects the hypothesis that the regressors equal zero at the 0.001 significance level. For model 3, an F-test of the hypothesis that the additional dummies in model 3 compared to model 1 (and to model 2) shows that the additional variables are significant.

	Model 1	Model 2	Model 3
55 cluster dummies for all pairs of the largest 10 clusters	Yes	No	Yes
55 industry dummies for all pairs of the largest 10 industries	No	Yes	Yes
R-Square	0.0201	0.0473	0.0617
N	1,365,620	1,365,620	1,365,620

Appendix
Fama French 30 Industry Codes

The names of the 30 FF industry codes are listed below. For the complete listing of the SICs assigned to each of these codes, download the industry definitions from Kenneth French's website:
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/changes_ind.html

1	Food	Food Products
2	Beer	Beer & Liquor
3	Smoke	Tobacco Products
4	Games	Recreation
5	Books	Printing and Publishing
6	Hshld	Consumer Goods
7	Clths	Apparel
8	Hlth	Healthcare, Medical Equipment, Pharmaceutical Products
9	Chems	Chemicals
10	Txtls	Textiles
11	Cnstr	Construction and Construction Materials
12	Steel	Steel Works Etc
13	FabPr	Fabricated Products and Machinery
14	ElcEq	Electrical Equipment
15	Autos	Automobiles and Trucks
16	Carry	Aircraft, ships, and railroad equipment
17	Mines	Precious Metals, Non-Metallic, and Industrial Metal Mining
18	Coal	Coal
19	Oil	Petroleum and Natural Gas
20	Util	Utilities
21	Telcm	Communication
22	Servs	Personal and Business Services
23	BusEq	Business Equipment
24	Paper	Business Supplies and Shipping Containers
25	Trans	Transportation
26	Whlsl	Wholesale
27	Rtail	Retail
28	Meals	Restaurants, Hotels, Motels
29	Fin	Banking, Insurance, Real Estate, Trading
30	Other	Everything Else