Clustering-based Sector Investing

MATTEO BAGNARA[§] AND MILAD GOODARZI[¶]

ABSTRACT

Industry classification groups firms into finer partitions to help investments and empirical analysis. No solution has been proposed yet for the well-documented problems of standard industrial definitions, like their stale nature and coarse categories for firms operating in multiple areas. We allocate firms to homogeneous groups maximizing the within-group explained variation through a clustering approach that employs 69 firm characteristics. The resulting novel economic sectors represent a better investment set compared to existent classification schemes for portfolio optimization and for trading strategies based on within-industry mean-reversion. We provide a new metric to quantify feature importance for clustering methods, finding that size drive differences across classical industries while book-to-market and financial liquidity variables matter for clustering-based sectors.

JEL classification: G12, C55, C58

Keywords: Empirical Asset Pricing, Risk Premium, Machine Learning, Industry Classification, Clustering

[§]Leibniz Institute for Financial Research SAFE, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60323, Frankfurt am Main, Germany, bagnara@safe-frankfurt.de.

[¶]Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60323, Frankfurt am Main, Germany, Milad.Goodarzi@hof.uni-frankfurt.de.

We thank Christian Schlag and Alexander Hillert for useful comments. We gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE.

1 Introduction

Classification, i.e. the grouping of objects into categories that share similarities, is one of main mechanisms of human thought (Rosch and Lloyd, 1978) and as such it is pervasive also in financial markets. In the context of portfolio allocation decisions, many investors adopt a "top-down" approach where they first identify broad asset classes and only afterwards decide how to allocate their funds across assets within a class (Barberis and Shleifer, 2003). One significant categorization concerns the formation of economic sectors or industries, that dates back to the 1930s with the introduction of the Standard Industrial Classification (SIC) codes in the U.S. for statistical purposes across governmental agencies.

Industries represents the focal point of several investors' trading strategies as they offer an off-the-shelf classification of firms into groups that share similar products. For each year from 1998 to 2010, industry knowledge was the most crucial research attribute of equity analysts according to *Institutional Investor Magazine*. Analysts often specialize in industries, issuing industry-level forecasts and recommendations (Kadan et al., 2012). Some institutions offer sector-oriented mutual funds like "Vanguard Information Technology" or "Vanguard Commodity Strategy Fund". Investment decision are influenced by industry categorization both at the institutional (Busse and Tong, 2012) and at the retail level (James et al., 2013). Furthermore, some financial phenomena often have a relevant industry-wide component, such as the *dot-com bubble* (James et al., 2013) and the momentum effect (Moskowitz and Grinblatt, 1999). Industries are critical also in research: between 1995 and 2003 they have been used for different purposes in 18 papers in American Economic Review, 70 in Journal of Finance and 467 in Journal of Financial Economics (Weiner, 2005).

What can market participants expect to earn by investing at the industry level? Figure 1 shows standard deviation and average excess return for 48 industries obtained using Fama and French (1997) categories for a large sample of firms between 1984 and 2019, where lighter colors denote higher Sharpe Ratios (SR).[1] To summarize their investment performance, we focus on the maximum SR portfolio built using industries as base assets. During the period considered, it earns an average excess return of 1.3% and an annualized SR of 1.26. As a benchmark, the market factor has a mean of 0.7% and an annualized SR of 0.58.[2] Economic sectors have therefore a strong potential to construct profitable investment strategies out of a contained number of assets.

In this paper, we revisit the long-standing problem of assigning firms to homogeneous groups with the help of clustering methods in order to provide economic sectors that represent a better investment set for mean-variance investors and that fully exploit the mean-reversion

[1] Further details about the data used can be found later in Section 4.

[2] See https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

of stock returns within the same group. We focus on sector investing because it makes it easier for market participants to optimize their portfolio choices by reducing the universe of individual stocks to a tractable number of assets. In research, economic sectors are useful for empirical analysis and modelling. For example, Fama and French (1997) industry portfolios represent a notoriously hard set of test assets that can inform about an Asset Pricing model's validity.[3]

Firms are usually classified according to four main criteria, i.e. SIC, North America Industrial Classification System (NAICS), the industries provided by Fama and French (1997) (henceforth FF) and the Global Industrial Classification Standard (GICS). The literature has highlighted several drawbacks affecting these schemes. For example, SIC codes often do not coincide across different data vendors (Guenther and Rosman, 1994; Kahle and Walkling, 1996) and struggle to identify firms with similar characteristics (Clarke, 1989). Despite being designed to research purposes, Fama and French (1997) document imprecise cost of equity estimates for their industries. More recent studies suggest the GICS classification outperforms the other systems. A prominent example is Bhojraj et al. (2003), who show that GICS codes are significantly better at explaining stock return comovements and cross-sectional variation in key valuation ratios using S&P 1500 firms. The reason for this improved performance is due to a more sophisticated categorization of firms into sectors. More dated classifications mainly focus on a company's largest product line. As such, they are inherently static (except for sporadic revisions carried out by the provider) and inevitably coarse for firms active in multiple areas. In contrast, GICS codes assign firms to economic sectors taking explicitly into account information from financial statements and investment research reports in order to satisfy the needs of investment professionals.

In our work, we follow and develop further this logic using firm characteristics to find economic sectors that deliver better investment perspectives compared to existing systems. More in detail, we use bisecting mean clustering on a large number of well-known return predictors from the literature (Gu et al., 2020) to find the groups that maximize the within-cluster explained variation. The average R^2 that a cluster portfolio achieves in explaining the returns of firms in that cluster is a natural metric to judge the validity of industry groups (Bhojraj et al., 2003). Using firm characteristics as starting point for the clustering exercise, we build a bridge between the anomaly literature and the industry classification issue. If characteristics predict returns, sectors constructed using this information have a tighter link with portfolio performance and improve investment profitability. To help interpreting the new clustering-based sectors we find, we introduce a novel approach to quantify the relative

[3]For instance, for the period July 1984 - June 2019, Fama and French (1993) model explains well above 80% of the variation of 25 portfolios sorted on size and book-to-market or on size and momentum, whereas it achieves only 59% for 30 industry portfolios. Data are from Prof. Kenneth's website.

importance of features in determining differences across clusters.

Our results are four-fold. First, clustering-based classification delivers sizeable improvements with respect to standard industries in the task of creating groups whose returns comove tightly. The average in-sample R^2 of 10 cluster sectors, obtained regressing each firm i's CAPM residuals in cluster k on the corresponding cluster portfolio k, is 9.31%. As a comparison, SIC codes explain only 5.98% and Fama-French 10 industries 8.51%. For any number of sectors K ranging from 5 to 48, cluster portfolios are better than any other standard classification scheme. In other words, firm characteristics contain information that leads to more homogeneous economic sectors.

Second, these novel cluster sectors deliver better investment opportunities for mean-variance investors. The maximum in-sample SR portfolio from $K = 10$ sectors earns an annualized SR of 1.98, which is higher than what one can obtain investing in any other set of industries (e.g. 1.67 with SIC codes, 1.81 with Fama-French industries and 1.84 with GICS codes). This holds for almost any K considered. More importantly, cluster portfolios outperform by far all other classifications Out-Of-Sample (OOS). By creating more uniform firm groups, our method finds portfolios whose returns spread much more widely, thereby improving investment opportunities.

Third, trading strategies based on the mean-reversion of stock returns belonging the same group are remarkably profitable for cluster sectors. Averaging mean-reversion portfolios across $K = 10$ cluster delivers a monthly mean excess return 0.40% and an alpha of 0.36% with respect to Fama and French (2015) plus momentum, both of which are highly statistically significant. Similar strategies based on any other industrial classification exhibit instead average returns close to zero and statistically insignificant. Clustering-based classification has therefore also a practical investment appeal.

Fourth, we provide a motivation for our findings that hinges on a new metric that quantifies the contribution of firm characteristics to distinguishing clusters from each other, the *Proportion of Across-Clusters Feature Spread (PAC-FS)*. The *PAC-FS* captures the percentage of variation across clusters and features that is due to a certain covariate. We find that while classical industries mostly differ in terms of size, the main drivers of differences across clusters are book-to-market and financial liquidity variables (quick and current ratio). These characteristics are likely responsible for the better investment performance of cluster sectors.

The rest of the paper is organized as follow. After providing a review of the literature in Section 2, we give an overview of the existent classification schemes in Section 3. Section 4 illustrates our data sample and Section 5 explains the method we use. Section 6 presents the empirical findings. Section 7 concludes.

2 Literature review

Our paper relates to various strands of literature. One concerns the validity of different industry classification schemes and the comparison among each other for a variety of purposes. Hrazdil et al. (2013) document the superiority of GICS codes for NYSE and NASDAQ firms following the approach in Bhojraj et al. (2003). Chan et al. (2007) find similar results concerning return covariation at increasingly finer levels of industry partitioning. Kile and Phillips (2009) argues GICS codes deliver improvements over SIC and NAICS in identifying technology firms. We treasure the result that a classification scheme that goes beyond mere product considerations like GICS offers better performance, and we extend this approach by considering 69 firm characteristics that have predictive power for expected returns from the literature to inform the classification algorithm.

The role of economic sectors for investment purposes has attracted the interest of several academics. Moskowitz and Grinblatt (1999) find that the bulk of momentum effect, one of the most famous investment anomalies, can be attributed to momentum at the industry level. Hameed and Mian (2015) document strong intra-industry reversal effects due to order imbalances and non-informational shocks. Busse and Tong (2012) shows that roughly one third of fund performance can be accounted for by industry selection while the rest is due to the performance of individual stocks relative to their own industry. James et al. (2013) suggest that industry-wide categorization influences the investment decisions of retail investors, with market participants chasing past winning industries.

Our work fits well the emerging literature that applies Machine Learning (ML) in Asset Pricing.[4] A benchmark in this context is given by Gu et al. (2020), who compare a large number of different ML techniques for predictions purposes. Freyberger et al. (2020) attempt at establishing which firm characteristics deliver independent information for the cross-section of expected returns using a method called adaptive group LASSO. Kozak et al. (2020) estimate the weights of the Stochastic Discount Factor (SDF) through a robust mean-variance optimization. Bryzgalova et al. (2020) suggest to use Asset Pricing restrictions to guide the pruning procedure while using random forest. Goodarzi et al. (2022) use fused LASSO to perform dynamic model selection. In our work we apply a classical *unsupervised* learning algorithm like bisecting K-means to find those groups of firms whose returns comove as tightly as possible, thereby engineering a pseudo-supervised classification technique.

Lastly, our research question is linked to cluster analysis, whose application has been scant so far in the field of Asset Pricing. Greengard et al. (2020) employ t-distributed stochastic neighborhood embedding (t-SNE) to cluster risk factors into 6 groups. In simi-

[4]For a comprehensive review of the methods employed in this area, see Giglio et al. (2022). Bagnara (2022) offers a thorough review of the empirical results.

lar spirit, Geertsema and Lu (2020) use agglomerative clustering to group anomalies based on correlation-based dissimilarity. In the context of industrial organization, Hoberg and Phillips (2016) group firms into industries using a clustering algorithm on the text of 10-K product descriptions, and Hoberg and Phillips (2018) document momentum effects using text-based industries. Differently from ours, their method does not account for firm characteristics. von den Hoff (2022) proposes a technique to quantify the economic value of clustering that helps uncovering patterns in the data that are due to investors' limited attention. Kakushadze et al. (2016) use information contained only in past returns to group stocks into clusters similar to industries. Weiner (2005) carries out an extensive comparison across different classification schemes and suggests that a cluster analysis may provide better results in terms of financial multiples. Evgeniou et al. (2021) assign firms to clusters to enhance the performance of a two-stage econometric model for individual firm predictions.

Our contribution differs from others as we provide a novel firm classification that represent an attractive set of portfolios for mean variance optimization accounting for the information contained in many firm characteristics. More specifically, we do not look for an optimal number of clusters based on prediction performance; rather, we fix K to match the number of economic sectors delivered by standard classification systems and we select as optimal cluster configuration the one that maximizes the within-cluster explained variation, which is a natural valuation metric for industrial categories. Furthermore, we benchmark our cluster sectors against every major classification scheme, and not only against SIC codes, which have been documented to exhibit several drawbacks. Finally, instead of individual stocks, we focus on economic sectors for investment purposes, as they represent the asset universe for many market participants and analysts (Kadan et al., 2012; Busse and Tong, 2012).

3 Standard Classification Schemes

3.1 An Overview

Firms are usually assigned to economic sectors according to four main classification schemes.

The oldest and probably the most notorious is the SIC, established in the 1930s by the Interdepartmental Committee on Industrial Classification under the Central Statistical Board. Its construction was aimed at providing the Federal Government with a standard classification to be adopted for statistical purposes. SIC codes are integers of 4 digits and follow a top-down approach, where the first 2,3 and 4 digits define major industry groups, industry groups and industries, respectively.The first digit is defined by the product line representing the largest percentage of sales in the 10-K filing. SIC classification was lastly

revised in 1987 as later on a new scheme would have replaced it.

The NAICS was introduced in 1999 under joint development by Canada, Mexico and United States to offer a system that would reorganize "industry groups to better reflect the dynamics of our economy, [...], allowing first-ever industry comparability across North America" (Saunders (1999), p.37). After their introduction, SIC codes were not discontinued and are still reported by several data vendors like CRSP and Compustat even nowadays. SIC and NAICS share many commonalities, including being issued by governmental agencies and following a hierarchical lineage. NAICS codes are in fact 6-digit long, where the first 2, 3, 4, 5 and 6 digits identify general categories of economic activity, subsectors, industry groups, NAICS industries and national industries, respectively.Another feature in common with SIC codes is that NAICS codes are product-oriented and far from concerns that can affect financial research and practice (Bhojraj et al., 2003).

The Fama-French industries (FF) were instead developed by academics "to have a manageable number of distinct industries that cover all NYSE, AMEX and NASDAQ stocks" (Fama and French (1997), p. 156), although they crucially hinge on SIC codes.[5] They were constructed, in fact, by reorganizing the existing SIC code-based industries into a total of 48 new groups that provide groups more likely to share common risk characteristics. As such, FF industries are also product-based although it is clear that research was trying to develop a classification system that could go beyond mere product considerations. The FF groups have been vastly used in the literature and have become the reference point for several works concerning economic sectors (e.g. Hameed and Mian (2015)).

The latest classification scheme is the GICS, born in 1999 from the collaboration between Morgan Stanley Capital International (MSCI) and Standard & Poor's (S&P). It significantly departs from the other supply-based approaches, as the industry assignments take into account a firm's principal business activity but are also informed by annual reports, financial statements and investment research reports which reflect market participants' perceptions. The goal of GICS code is to "enhance the investment research and asset management process for financial professionals worldwide" (S&P and MSCI, 2002). Furthermore, firms can be assigned to the Industrial Conglomerates subindustry (Industrial Sector) or to the Multi-sector Holdings subindustry (Financial Sector) if they do not fall neatly into a single category.[6] GICS code consists of up to 8 digits, where the first 2, 4, 6 and 8 digits identify sectors, industry groups, industries and sub-industries, respectively.[7]

[5]The classification was introduced in Appendix A of the article and it is available at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[6]If the company is engaged in at least two business categories, none of which constitutes 60% or more of the total revenues, a more extensive analysis is carried out to determine the appropriate classification.

[7]In Compustat, GICS codes are available even before 1980 for some companies. We do not find information about back-filling for data points, but it is likely that the data vendor has extended the first available code to some previous years in some cases.

3.2 Drawbacks of Standard Classification Schemes

The classification schemes discussed above present several drawbacks. First of all, there might some discrepancies across different databases. Guenther and Rosman (1994) find in fact that the primary two-digit SIC codes from CRSP and Compustat do not coincide 38% of the time. Weiner (2005) argues that the concordance of SIC codes across different data vendors decreases over time. Second, SIC codes have hard times at identifying firms with similar characteristics like sales changes, profit rates or stock price changes (Clarke, 1989). The introduction of newer industry codes over time has aimed at sidestepping some of these issues. For example, Krishnan and Press (2003) find that NAICS deliver some improvement for some industries compared to SIC in terms of intra-industry variation in financial ratios. However, the degree of success strongly depends on the framework considered. Although designed for academics, Fama and French (1997) industries are still dependent on SIC codes and thus on their shortcomings. More recently, evidence suggests that the GICS system stands out among the standard classification schemes (Bhojraj et al., 2003; Hrazdil et al., 2013), and should be used as benchmark by both academics, regulators and practitioners. We propose a new classification algorithm that uses information contained in a large number of return predictor from the "factor zoo" literature (Cochrane, 2011), *de facto* treasuring the results that classification schemes based merely on the firm primary line of business are inherently static and coarse, and inevitably perform worse than other systems which update more frequently and that pay attention to information coming from financial markets.[8]

4 Data

Our dataset coincides with the updated version of Gu et al. (2020).[9] The original data includes 94 firm characteristics for CRSP stocks in the NYSE, AMEX, and NASDAQ, that we merge with CRSP return data. To avoid artificial influence to the time-series fluctuation (Chen et al., 2020), we do not impute the cross-sectional mean to the missing values. Instead, we include covariates with not more than 37.5% of missing values in the full sample, which leaves us with 69 characteristics.[10] Then, we require to have at least 1000 stocks per month, retaining only those that have data for all characteristics and at least 60 months of available return data.[11] The sample spans July 1984 to June 2019 for a total of 7052 firms. On

This, however, seems not worrisome as only a small percentage of firms had a change in their GICS code assignment in recent years (Bhojraj et al., 2003).

[8] MSCI and S&P claim that GICS codes are revised annually.

[9] We thank the authors to make the data available at https://dachxiu.chicagobooth.edu/#research.

[10] See Appendix A for a detailed description of the variables included.

[11] As we estimate our model every 12 months, in this way we ensure that we can track the behavior of each firm for a non-negligible amount of the time across different model estimations.

average, there are 2822 firms per month and 3016 per year. SIC codes are obtained from CRSP because changes over time are not covered by other data vendors. NAICS and GICS codes are acquired from Compustat.

Similarly to Kozak et al. (2020), we cross-sectionally rank-transform the firm characteristics and map them into the unit interval to make the results insensitive to outliers. Finally, characteristics are cross-sectionally standardized so that they are all on the same scale.

To keep the comparison as precise as possible, industry portfolios are replicated applying each classification scheme to the firms contained in our data sample. Using the first digit of the SIC codes delivers 9 industries; the first digit of NAICS 18; the first and second digits of GICS result in 11 and 24 sectors, respectively.[12][13] Finally, Kenneth French's website offers 5, 10, 12, 17, 30 and 48 industries, which are arbitrary numbers without a clear economic motivation. We keep 48 as the maximum number of potential industries and thus rule out using further digits of other codes both to ensure comparability across all the methods considered and to avoid having groups with very few firms.[14] Below, we use abbreviations of the type "SIC9" to denote industries formed following the classification given by the letters (here: SIC) that results in a number of groups indicated by the digits (here: 9).

5 Methodology

We group firms into clusters using bisecting K-means, choosing the optimal clusters based on a measure of within-cluster commonality. Further, we develop a new metric to quantify feature importance for clustering methods. We illustrate our approach in what follows.

5.1 Clustering Algorithms

Our technique is based on bisecting K-means, an improvement over the basic K-means. To help the reader who is not familiar with clustering analysis, we provide here a quick overview of the standard algorithm.

Cluster analysis aims at grouping a sample of data points into a user-specified number of subsets or "clusters" K such that the dissimilarity of observations within a cluster is minimized. Let x^i be a vector containing P different features (characteristics) for observation i. Assuming the data points have already been assigned to a certain cluster, one straight-

[12] Missing data can cause the number of industries we obtain to be smaller than what the classification system should deliver. While there are 10 1-digit-SIC industries and 20 1-digit-NAICS industries, we have 9 and 18, respectively.

[13] Notice that some studies report an outdated number of industries for 1-digit GICS codes. For example, in Bhojraj et al. (2003) there are only 10 industries while there are now 11 (see https://www.msci.com/our-solutions/indexes/gics). We also take care of the revisions to the GICS structure that took place in 2016 and 2018.

[14] For example, 2-digit SIC codes result in 67 industries; 2-digit NAICS 97 and 3-digit GICS 76.

forward way to formalize the notion of similarity between two observations is to use the Euclidean distance over all the P features (Hastie et al., 2009):

$$d(x^i, x^j) = ||x^i - x^j||^2 = \sum_{p=1}^{P} (x_p^i - x_p^j)^2 \tag{1}$$

where x_p^i denotes the p-th characteristic for the i-th observation. In cluster analysis, it is customary to compute the dissimilarity of a cluster as the average of the Euclidean distance between every point belonging to a cluster and the so called *centroid* of the cluster, i.e. its centre. This such measure is often called *Within-Cluster Sum of Squares (WCSS)* or *inertia*, and the objective is to minimize it. Said it differently, clusters are formed in order to contain data points that are very close to each other in the P-dimensional feature space. Hastie et al. (2009) show that minimizing the WCSS is equivalent to maximize the between-cluster dissimilarity. But how does one assign observations to clusters in the first place? The most accurate method would be using combinatorial optimization, which evaluates every possible arrangement of the data into K clusters. Operationally, this is computationally infeasible unless the dataset is very small. Therefore, feasible strategies based on "iterative greedy descent" are needed. These algorithms specify a (random) initial configuration of data points into clusters, and at each iteration the cluster assignments are changed in order to reduce the WCSS. When no further improvement is possible, the algorithm stops. In this way, only limited fraction of all the possible assignments is examined, which makes the algorithm feasible. To avoid getting stuck on local optima and ensure robustness, changing the initial cluster configuration becomes pivotal.

Among these greedy strategies, the K-means is one of the most popular. Figure 2 illustrates a simple example in a two-dimensional space. Assume the data points belong to $K = 3$ unknown groups, identified by different colors, that we want to uncover with K-means. To initialize the algorithm, K initial guesses for the centroids are needed. These can simply be random numbers or, instead, be determined by the user's expert judgment. In the first panel, centroids are depicted as fully colored circles. The second step consists in computing the Euclidean distance for all observations in the sample with respect to all centroids, and clusters are determined assigning data points to the closest centroid such that the pre-specified number K of groups is formed. These clusters or partitions are delimited with black lines in the second panel. Then, the WCSS is computed and new centroids for each cluster are obtained as the average of all observations belonging to the same clusters. In the third panel, one can in fact see that the centroids corresponding to each region have changed. The last two steps are repeated until the WCSS does not change anymore. At that

point, K-means provides the clusters that minimize the WCSS.

5.2 Bisecting K-means

Bisecting K-means is a hybrid approach between divisive clustering (i.e. top-down recursive clustering, Hastie et al. (2009)) and K-means clustering. While standard K-means partition the dataset into K clusters at each iteration, bisecting K-means recursively splits one cluster into 2 sub-clusters at each step of the algorithm using K-means, until K clusters are obtained. From this perspective, it is a simple development of the basic K-means whose approach might remind the reader of conditional portfolio sorts or regression trees.

More specifically, the user specifies a desired number of clusters K. In the first step, K-means is used to partition the data into 2 clusters, such that the resulting intra-cluster similarity is maximized. This corresponds to the first panel in Figure 3. Then, one of the two clusters is split again into 2 sub-clusters using K-means. To choose which cluster to split further, one can select either the cluster with the highest WCSS or the one with the most data points. We follow the first strategy which resembles the classical minimization of a loss function.[15] The result is depicted in the second panel. Regardless the guiding criterion, the splitting procedure always provides 2 sub-groups each time, and this is where the term "bisecting" comes from. After the second split, the algorithm continues splitting the cluster with the highest WCSS at each step (like in the bottom panel), until K clusters are found.

Bisecting K-means presents several advantages over simple K-means. First, it is more efficient when K is large, because at each step it utilizes the data points and the centroids within only one cluster instead of the whole sample. This reduces the computation time. Second, it is prone to deliver clusters of more similar sizes, while K-means is known to produce groups with wildly different number of observations. This is likely to be the reason why overall bisecting K-means outperforms K-means under several metrics (Steinbach et al., 2000). Third, it can identify clusters of any shape, while K-means can uncover only spherical ones. Fourth, the intuition behind the method is very similar to conditional portfolio sorts, which facilitates its interpretation.

5.3 Clustering Firms Using Firm Characteristics

We cluster firms into K clusters at the end of June of each year t, for $t = 1984, ..., 2019$. Depending on their nature (e.g. accounting versus return-based), characteristics change over time but the bisecting K-means requires a one-dimensional vector of inputs for each firm (column). To comply with this restriction, we aggregate the information using their time-

[15] Results obtained with both approaches are usually very similar, especially when there are many data points available.

series average over the past year at the firm level. By repeating the clustering exercise every 12 months, we are able to track potential changes over time to a reasonable degree, whereas traditional classification schemes are updated only sporadically and are essentially static. As an alternative to average characteristics over the period considered, one could use the entire matrix of characteristics pooled together as input, but this would result in repeated observations of the same firm belonging to either the same or potentially different clusters, which is difficult to interpret.[16] To initialize the calculation of the cluster centroids, we use the "k-means++" method, which uses the information contained in the empirical probability distribution to sample and initialize the cluster centroids, thereby speeding up the process. We run the algorithm for 1000 different initial random states ("seeds"). To make sure results are robust, for each seed the algorithm is run starting from 5 different random initial choices for the centroids, and the one producing the best outcome in terms of inertia is chosen for the model, as is customary with clustering methods. Next, we describe the procedure we follow to choose the optimal clusters among the 1000 different seeds.

We start by computing the CAPM residuals of each stock i, i.e. $\hat{\varepsilon}^i$. For a given random seed and for cluster k, $k = 1, ..., K$, we regress the CAPM residuals of each stock belonging to that cluster, $\hat{\varepsilon}^{i,k}$, on the CAPM residuals of k-th "cluster portfolio", wihch is the value-weighted average of all stocks in cluster k: $\hat{\varepsilon}^k = \sum_{i \in C_k} w_i \hat{\varepsilon}^{i,k}$, where C_k denotes the set of firms in cluster k and w_i are value weights.[17] The average within-cluster R^2 provides a measure of commonality in the cluster, which is the ultimate goal of industry classification in terms of research-related purposes (Bhojraj et al., 2003).[18] We use the average fit across all the K clusters to pick the best cluster configuration from the 1000 random seeds. By adding this Asset Pricing criterion to inform cluster selection, we build a bridge between Machine Learning and Finance with a simple and clear economic interpretation.

The classification algorithm is repeated for $K = 5, 9, 10, 11, 17, 18, 24, 30, 48$. We use these numbers to closely match the number of industries provided by other schemes that we list in Section 4. When calculating the within-industry R^2, we follow Bhojraj et al. (2003) and consider only "functional" groups, i.e. those with at least 5 firms. The same requirement is employed also elsewhere, such as in Berger and Ofek (1995) and Villalonga (2004).

Before moving on, a remark is necessary. Clustering algorithms like K-means and bi-

[16] While the frequency with which the procedure is repeated could be matter for discussion as it might disregard information contained in relatively fast-changing variables like short-term reversal, we believe that 12 months is a reasonable time window considering that most of the variables used have a low frequency, such as operating profitability and book-to-market and in general those derived from balance sheet information.

[17] While seeking the best cluster configuration, we look at CAPM residuals but elsewhere in the text "cluster portfolios" denotes the returns to portfolios formed aggregating the excess returns of stocks belonging to the same cluster.

[18] Differently from Bhojraj et al. (2003), we maximize over the intra-cluster R^2 of the residuals rather than of the overall returns. In this way, we remove the component of returns that is due to the exposure to the market factor, which would interfere with the construction of groups of firms similar to each other.

secting K-means assign observations to clusters based on a measure of inertia, as mentioned above. The label that is given to clusters has no meaning, i.e. the clusters are not ordered according to some measure. Therefore, if one repeats the clustering procedure more than once on the same sample, the clusters to which firms are assigned to will remain the same, but not necessarily the labels assigned to the clusters. For instance, consider splitting a sample of 4 firms into 2 clusters, C_1 and C_2, such that firms 1 and 2 belong to C_1 and firms 3 and 4 belong to C_2. If the clustering is repeated, we will obtain again two clusters with same shape and size as C_1 and C_2, but it could happen that now firms 1 and 2 belong to C_2 whereas firms 3 and 4 belong to C_1. To sidestep this issue, we assess our clustering method using several tests which refer to the year in which the algorithm is run, as better explained below. In this way, not only we make sure that the issue of changing labels does not impact the results, but also provide a more transparent set of evaluation metrics that model users can look at each time they utilize our algorithm.

5.4 Feature Importance in Clustering: A novel Metric

Finally, we introduce a novel evaluation metric to gauge feature importance in the context of clustering algorithms. Clustering finds homogeneous groups in order to minimize the WCSS, as explained above. All features contribute equally to the WCSS, and that is why normally there is no such thing as feature importance when talking about cluster techniques, in contrast to other unsupervised ML paradigms like dimension reduction (e.g. PCA). Nonetheless, we can still measure the contribution of each feature to determining differences *across* clusters. After all, if the WCSS is minimized in the optimal cluster configuration, we can expect meaningful differences related to features only across clusters and not within one. We measure this dimension with what we call *Proportion of Across-Cluster Feature Spread (PAC-FS).*

The computation of this metric goes as follows. First, we calculate x_p^k, the value-weighted mean of the firms in cluster k, $k = 1, ...K$, for each feature p, $p = 1, ..., P$: $x_p^k = \sum_{i \in C_k} w_i x_p^{i,k}$. This resembles value-weighted portfolios returns at the characteristic level. Second, we compute the range of variation for each characteristic *across* custers, i.e. the *spread* between the cluster with the highest and the lowest feature value:

$$S_p = \max_{k=1,...,K} \left\{ x_p^k \right\} - \min_{k=1,...,K} \left\{ x_p^k \right\}$$

PAC-FS$_p$ is the ratio between the spread of a characteristic and the sum of spreads over all

characteristics P:

$$PAC\text{-}FS_p = \frac{S_p}{\sum_{p=1}^{P} S_p} \tag{2}$$

$PAC\text{-}FS_p$ captures the proportion of variation across clusters and features that is due to feature p. Exactly like K-means considers the Euclidean distance across all feature when minimizing the WCSS, *PAC-FS* accounts for the spread over all characteristics, too. It quantifies how much of the differences across clusters, characteristic-wise, are driven by each feature. This is a novel metric to gauge feature importance for clustering methods and thus belongs to the contributions of our paper.

6 Empirical Results

We now illustrate the results of the empirical analysis we carry out applying the methods discussed above.

6.1 Descriptive Statistics

We begin by presenting descriptive statistics concerning the number of firms in each economic sector. Table 1 follows Bhojraj et al. (2003) and reports information regarding the distribution of firms divided into three groups of classification schemes. The left panel reports the results for SIC9, FF10, GICS11 and $K = 10$ clusters. The middle panel shows figures for FF17, NAICS18, and $K = 18$ clusters, and the right panel refers to GICS24 and 24 clusters. We group the various methods in this way such that the number of industries is comparable, because different standard classification schemes provide unequal numbers of industries. We therefore focus mostly on K between 9 and 11, between 17 and 18, and 24. Whenever possible, we compute the results for each other intermediate K for clusters to make the comparisons more precise.[19] Notice that the statistics refer to the "functional" sectors ($N \geq 5$ firms), like when computing the within-cluster explained variation. Hence, we can expect to observe differences in the average number of firms across different methods. In all three panels, the distribution appears very similar across all approaches. The average number of firms varies between 271 for GICS and 331 for SIC for $K = 5$ and decreases with more industries, reaching 126 for GICS24 and 124 for Cluster24. The standard deviation is high in all cases, being close to the mean. Distributions are all right-skewed, and the kurtosis is relatively high for FF17 and NAICS18. Overall, clustering-based sectors share common patterns with other classification schemes in terms of number of firms per industry.

[19] We keep the same approach also later on.

6.2 Within-cluster Explained Variation

We compute the average within-group R^2 for economic sectors formed with SIC and NAICS codes, Fama and French (1997) industries, GICS codes and those obtained from our clustering algorithm at the end of June of each year (i.e. when clusters are formed) and report time-series averages for different K in Table 2. Clustering is performed using the 69 firm variables detailed in the Appendix. When calculating the within-sector R^2, only firms with at least 10 available observations over the past 12 months are considered to ensure meaningful statistics. Furthermore, only "functional" industries are considered (Bhojraj et al., 2003).

The table shows that clustering outperforms all other classification methods for every K. With 5 sectors, the average cluster portfolio alone is able to explain 7.92% of the variation in CAPM-adjusted firm residuals in the same group, while FF capture slightly above 4%. With $K = 10$, clustering achieves an R^2 of 9.31% against 5.98% for SIC9, 8.51% for FF10 and 8.80% for GICS11.[20] When K is higher, all classification schemes are better at summarizing the within-industry variation, as one might expect. With $K = 18$, the R^2 related to clustering-based sectors is 10.10%, for NAICS18 it is 8.90% and for FF17 it is 9.98%. With $K = 24$, the R^2 for clustering is 10.40% whereas for GICS24 it is 9%. Among standard industrial classifications, the best results are achieved by GICS, both with K close to 10 and with K close to 24, a finding in line with previous literature (e.g. Hrazdil et al. (2013)). When increasing the number of industries even further, we observe that the differences become smaller: for K sufficiently high, the groups of firms become so small and so homogeneous that differences among algorithms are less crucial than for fewer clusters. Nonetheless, the main takeaway is that informing the clustering procedure based on information from firm characteristics deliver tremendous improvements over static product-based classifications that do not account for it.

Before concluding this section, we briefly address a further aspect of interest besides within-sector commonalities, i.e. the degree of covariation across different clusters. Economic intuition would suggest that, the tighter firms comove in the same cluster, the less correlated different clusters should be among each other. We measure this behaviour by computing the average pairwise correlation of each sector with the others, for each classification method.[21] Results are presented in Table 3. We notice two patterns. First, as expected, across-cluster correlations generally decrease for higher K. Second, there are no great differences between classification schemes for comparable K. For example, the correlation across $K = 5$ clusters is 0.74 and for FF5 it is 0.75. The difference in terms of within-cluster explained variation is

[20] Since clustering is more flexible, we report here the results also for intermediate K which show that varying K by a few units does not change much the results.

[21] As explained above, from here on we use excess returns of cluster portfolios, not CAPM residuals.

instead sizeable as discussed above. This finding is not at odd with the nature of clustering methods: while minimizing the WCSS is equivalent to maximizing the dissimilarity *between* the clusters (i.e. the distance between observations in one cluster and data points in different clusters), comovements *across clusters* represent a different dimension. Here, we document that differences across industry definitions do not impact it significantly.

6.3 Investment Perspective

A distinctive trait of our clustering approach is that clusters are determined based on a clear objective, i.e. the maximization of the explained variation within each sector. We have shown that firms comove more tightly within clustering-based sectors than within classical industries. Furthermore, we base the cluster formation on a large number of return predictors from the literature. If characteristics predict returns, sectors built with such information have a closer link to portfolio performance and we have reason to believe that this can lead to improvements in investment profitability relative to standard classification schemes that are mainly product-oriented and are not designed to meet financial professionals' needs. As mentioned earlier, sector investing is a key activity for both retail and institutional investors, and we aim at enhancing it through our method. We now illustrate two investment applications in which clustering-based sectors are particularly appealing, namely the construction of the tangency portfolio and a trading strategy that exploits within-cluster mean-reversion.

6.3.1 Tangency portfolio

In the spirit of using economic sectors as a reduced asset space from which market participants can pick from, as it often happens in practice (Kadan et al., 2012), we perform a classical mean-variance optimization to find the maximum SR portfolio (or tangency portfolio) using industries and clusters as base assets. More formally, we solve the problem:

$$\begin{aligned} \max_{w} \left\{ \frac{\delta' \mu}{\sqrt{\delta' \Sigma \delta}} \right\} \\ s.t.\ \delta' \mathbf{1} = 1 \\ 0 \leq \delta_i \leq 1 \ \forall \ i = 1, ..., K \end{aligned} \tag{3}$$

where μ represents an $K \times 1$ vector of expected excess returns, Σ is the corresponding variance-covariance matrix, $\delta = (\delta_1, ..., \delta_N)$ is a vector of portfolios weights for the K available assets and $\mathbf{1}$ is a $K \times 1$ vector of ones. The second constraint imposes short-sale restrictions. We solve the optimization problem at the end of June of each year after performing firm clustering, and compute the SR of the resulting tangency portfolio. In Table 4 we report

the results for both standard industrial classifications and clustering-based sectors. The left panel shows the in-sample results that refer to the entire period 1984-2019. Clustering surpasses the other standard industries for almost any number of sectors K. For example, Cluster10 deliver an annualized SR of 1.98 against 1.67, 1.81 and 1.80 for SIC9, FF10 and GICS11, respectively. Increasing K increases the tangency portfolio SR thanks to a larger investment universe, in line with economic intuition. For example, with $K = 18$, clustering yields a 2.28 SR versus 1.69 and 2.24 for FF17 and NAICS18, respectively. For $K = 24$, GICS and clusters essentially give the same results.

More importantly, clustering is the best-performing method also out-of-sample. The tangency portfolio is computed over the next 12 months keeping the classification unchanged at the end of June of year t.[22] This is an important exercise as it represents the results of feasible investment strategies beyond simple backtesting. With $K = 10$, clustering-based sectors can be combined into a tangency portfolio that earns an annualized SR of 2.05, against 1.65, 1.79 and 1.77 for SIC9, FF10 and GICS11, respectively. For $K = 18$, clusters deliver a SR of 2.47, also higher than that earned using FF17 and NAICS18 (1.67 and 2.19, respectively). With $K = 24$, the SR is 2.61 for clusters and only 2.18 for GICS.

Thanks to the tighter intra-cluster return commonality, clustering provides economic sectors that represent a better investment set for mean-variance investors relative to the existent industries, thereby revealing potential for the financial industry.

6.3.2 Within-cluster Mean-Reversion Strategy

A second way in which an investor can take advantage of industry classification is by using a mean-reversion argument (Kakushadze, 2015). If we believe that returns of firms within an economic sector k are linked together and comove, we expect that stocks that temporarily underperform the mean-sector return will outperform it in the future, and vice-versa for stocks that are currently outperforming. This idea is similar to that of statistical arbitrage for two identical assets with temporary different prices. Our conjecture is that such strategy is particularly profitable for clustering-based sectors as they capture higher within-cluster commonality as shown above. Hence, we design a trading strategy that, for each cluster k, goes long stocks whose returns at time t, r_t^i, are below the corresponding value-weighted cluster return r_t^k and that shorts stocks with returns above it. In other words, we form a

[22] Of course, there is no actual OOS for standard industries. We simply calculate the tangency portfolio for the period corresponding to the one that we use for the OOS test for cluster portfolios for a more precise comparison and to avoid that results are driven by year-fixed effects.

value-weighted portfolio of the type

$$p_t^{k,MR} = \sum_{i \in C_k} w_i \mathrm{D}_t(i,k) r_t^{i,k} \tag{4}$$

where

$$\mathrm{D}_t(i,k) = \begin{cases} +1, & \text{if } r_t^i < r_t^k \\ -1, & \text{otherwise} \end{cases} \tag{5}$$

and MR stands for Mean-Reversion. We refer to returns in excess of the risk-free rate.

Operationally, this strategy is formed and the end of June of each year for all industry classifications and all K reported previously, for each cluster k. To summarize its performance, we take an equal-weighted average of the K mean-reversion portfolios in each case. Figure 4 illustrates three different valuation metrics, where standard industry systems are represented by red bars and clustering-based ones by blue bars. Table 5 reports the actual figures with corresponding statistical tests. The first panel shows mean excess returns. While no mean-reversion strategy based on existing industries provide more than 0.1% per month on average, using clustering-based sectors earns between 0.4% and 0.57%. From Table 5, average excess returns for mean-reversion strategies are significant only for clusters (all well below the 1% significance level) and never for other classification schemes. Noteworthy, the magnitude tends to increase with K, confirming that a higher within-cluster commonality (which rises in K as per Table 2) is beneficial for trading strategies that exploit within-group mean-reversion.

The second panel shows annualized Sharpe Ratios. Clustering-based strategies are much more attractive than industry-based ones in terms of remuneration per unit of total risk. The fourth column of Table 5 reports the t-statistic for Sharpe Ratios based on Bailey and Lopez de Prado (2012), who show that this is standard-normally distributed. Hence, the values can be compared to classical critical values. Sharpe ratios are statistically different from zero at any conventional significance level for clustering-based sectors.

Finally, the third panel shows the alpha of each strategy with respect to the Fama and French (2015) model plus momentum ('FF6'). Even controlling for several important risk factors, mean-reversion strategies that use clusters instead of standard industry classification remain highly profitable with alphas that increase in K and are very similar to the full average returns. Said differently, such mean-reversion portfolios are largely unspanned by traditional risk factors, as captured by the adjusted R^2 in the last column of Table 5. Alphas are not significant for mean-reversion strategies based on standard industries.

Mean-reversion strategies exploit the comovements among stocks belonging to a certain group. As clustering is particularly powerful in providing sectors where the constituents

are tightly linked, it represents a much more valuable investment tool compared to existing industrial classifications.

6.4 Characteristics Importance

We use the new metric we develop, the *PAC-FS*, to identify which firm characteristics help to distinguish one economic sector from another. In Figure 5 we report the time-series average of the *PAC-FS*$_p$ for the twenty variables with the largest values, in descending order. Following the same approach as above, we compare the results across different industry classification for comparable K. Size (`mvel1`) is the first most important characteristic for both SIC9 and GICS11 and the second one for FF10. Its industry-adjusted version (`mve_ia`) belong to the top-6 features in all three cases. The amount of overall across-cluster across-feature spread these two characteristics determine together varies from around 8% (FF10) to 11% (SIC9). Market capitalization is therefore a distinctive trait for industries that are formed mainly looking at the firm's product lines. `sin` is the most important feature for FF10, which is not ideal as a binary categorization into sin and non-sin industries is a very coarse metric to distinguish many sectors from each other. Other variables that are important across the three systems are industry-adjusted change in profit margin (`chpmia`), industry momentum (`indmom`) and industry sales concentration (`herf`): although their relative rank varies, they all belong to the top-8 features.

A different pattern emerges when considering Cluster10. The book-to-market ratio (with the industry-adjusted variant) explain most of the across-cluster difference, i.e. almost 9%. The two next most important characteristics are financial liquidity ratios, i.e. the quick and the current ratio (`quick` and `currat`). The importance of size is considerably downsized compare to standard classification systems. Furthermore, it is noteworthy that after the 8th characteristic, the *PAC-FS*$_p$ flattens out almost completely, which means that even if all 69 covariates play the same role in forming the clusters as they enter with the same weight in the WCSS, only 8 of them can meaningfully be used to distinguish one cluster from another. This demonstrates that $PAC - FS$ is useful in uncovering interesting patterns related to feature importance.

Do similar findings hold also for higher K? In Figure 6 we show results for FF17, NAICS18, GICS24 and Cluster24.[23] Once again, size (and its industry-adjusted version) plays one of the biggest roles in the across-cluster across-feature spread for existing industrial classifications, taking the first, third and second place for FF17, NAICS18 and GICS24, respectively. The first characteristic for NAICS18 is `sin`, similar to FF10 but, interestingly,

[23] Results for Cluster18 are very similar and thus we omit them for clarity of exposition.

this is much less important for FF17. In fact, the first 5 positions for the latter change evidently with a larger number of industries. A similar pattern can be observed for GICS24 compared to GICS11. Now leverage (`lev`) becomes the most important feature, followed by size, `sin` and industry-adjusted size. A remarkable change happens for industry-adjusted book-to-market, which now is the third most important feature whereas it did not even appear in Figure 5. `chpmia`, `indmom` and `herf` remain relevant, in some cases more than in the case with $K = 10$ (e.g. `herf` for FF17). The situation, instead, remains substantially unchanged for clustering-based sectors: once again, book-to-market is the most crucial feature, followed by quick-and current ratio. Focusing on these 4 features, after which the *PAC-FS*$_p$ drops considerably and flattens out, one can conclude that clustering-based sectors are more stable to increases in K than other classification schemes where, besides size, differences are mainly determined by a rotating groups of features that varies with the number of industries considered. It is interesting that this happens in spite of the fact that classical industry codes are more static than cluster-based one: as we argued above, firm characteristics are a major source of comovement across individual firms and should be considered when grouping firms into economic sectors.

Another noteworthy phenomenon that emerges from Figure 6 is that the most salient features do not necessarily have lower *PAC-FS*$_p$ for higher K, i.e. the role they play in the overall characteristic-spread does not change substantially. Said differently, with higher K it is not more difficult to disentangle clusters from each other even if they become more "similar". Notice that this result does not go against the idea that smaller clusters are more homogeneous: *PAC-FS* measures the weight that each feature has in the differences *across* clusters *and* across features. Hence, it can well be that more homogeneous clusters differ more in terms of the same characteristic among each other, but the relative importance that each feature has with respect to other variables remains unaltered. *PAC-FS* is thus a suitable measure to capture differences across clusters that is not sensitive to the number of groups used.

Overall, differences across sectors identified by standard classification schemes tend to be driven first by variables related to the equity portion of the balance sheet (size), and second by elements connected to profitability (changes in profit margin) or to the recent market performance and the level of competition in an industry (industry momentum and sales concentration). In contrast, the main determinants of differences across clusters refer to a firm's "value" (book-to-market) or to its ability to meet its short-term obligations with its most liquid assets. These marked discrepancies are likely to explain the superior performance of cluster sectors relative to standard industries, which means our clustering algorithm can be useful to identify candidate variables that enhance the performance of trading strategies

at the economic-sector level.

7 Conclusion

Standard product-oriented industry classification presents several drawbacks that have led the profession to look for new schemes over time. Using the information contained in a large number of stock return predictors, we propose a new classification method that links the power of bisecting K-means with an easy-to-interpret Asset Pricing criterion, namely the maximum within-cluster explained variation, a natural metric to asses the goodness of industry assignments. Results reveal strong potential both for research and investment purposes. Clustering surpasses all existent classification schemes for every number of industries considered. Clustering-based sectors offer better investment opportunities for mean-variance investors willing to simplify the decision process from the entire universe of individual stocks to a tractable number of groups, both in-sample and out-of-sample. Significant gains from clustering classification arise also when exploiting mean-reversion trading strategies. Equipped with a new metric developed to quantify feature importance for clustering methods, we find that classical industries mainly differ in terms of size while book-to-market and financial liquidity variables are useful to distinguish clusters from each other.

References

Bagnara, M. 2022. Asset Pricing and Machine Learning: A critical review. *Journal of Economic Surveys* .

Bailey, D. H., and M. Lopez de Prado. 2012. The Sharpe ratio efficient frontier. *Journal of Risk* 15:13.

Barberis, N., and A. Shleifer. 2003. Style investing. *Journal of financial Economics* 68:161–199.

Berger, P. G., and E. Ofek. 1995. Diversification's effect on firm value. *Journal of financial economics* 37:39–65.

Bhojraj, S., C. M. Lee, and D. K. Oler. 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41:745–774.

Bryzgalova, S., M. Pelger, and J. Zhu. 2020. Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458* .

Busse, J. A., and Q. Tong. 2012. Mutual fund industry selection and persistence. *The Review of Asset Pricing Studies* 2:245–274.

Chan, L. K., J. Lakonishok, and B. Swaminathan. 2007. Industry classifications and return comovement. *Financial Analysts Journal* 63:56–70.

Chen, L., M. Pelger, and J. Zhu. 2020. Deep learning in asset pricing. *Available at SSRN 3350138* .

Clarke, R. N. 1989. SICs as delineators of economic markets. *Journal of Business* pp. 17–31.

Cochrane, J. H. 2011. Presidential address: Discount rates. *The Journal of finance* 66:1047–1108.

Evgeniou, T., A. Gueciouer, and R. Prieto. 2021. Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning. *Journal of Financial and Quantitative Analysis* pp. 1–36.

Fama, E., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33:3–56.

Fama, E. F., and K. R. French. 1997. Industry costs of equity. *Journal of financial economics* 43:153–193.

Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model. *Journal of financial economics* 116:1–22.

Freyberger, J., A. Neuhierl, and M. Weber. 2020. Dissecting characteristics nonparametrically. *The Review of Financial Studies* 33:2326–2377.

Geertsema, P., and H. Lu. 2020. The correlation structure of anomaly strategies. *Journal of Banking & Finance* 119:105934.

Giglio, S., B. Kelly, and D. Xiu. 2022. Factor Models, Machine Learning, and Asset Pricing. *Annual Review of Financial Economics* 14:null.

Goodarzi, M., C. Schlag, and S. von den Hoff. 2022. A New Model Every Month? — Dynamic Model Selection for Stock Return Prediction. *Available at SSRN 4028673* .

Greengard, P., Y. Liu, S. Steinerberger, and A. Tsyvinski. 2020. Factor Clustering with t-SNE. *Available at SSRN 3696027* .

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33:2223–2273.

Guenther, D. A., and A. J. Rosman. 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics* 18:115–128.

Hameed, A., and G. M. Mian. 2015. Industries and stock return reversals. *Journal of Financial and Quantitative Analysis* 50:89–117.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hoberg, G., and G. Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124:1423–1465.

Hoberg, G., and G. M. Phillips. 2018. Text-based industry momentum. *Journal of Financial and Quantitative Analysis* 53:2355–2388.

Hrazdil, K., K. Trottier, and R. Zhang. 2013. A comparison of industry classification schemes: A large sample study. *Economics Letters* 118:77–80.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*, vol. 112. Springer.

Kadan, O., L. Madureira, R. Wang, and T. Zach. 2012. Analysts' industry expertise. *Journal of accounting and economics* 54:95–120.

Kahle, K. M., and R. A. Walkling. 1996. The impact of industry classifications on financial research. *Journal of financial and quantitative analysis* 31:309–335.

Kakushadze, Z. 2015. Mean-reversion and optimization. *Journal of Asset Management* 16:14–40.

Kakushadze, Z., W. Yu, et al. 2016. Statistical Industry Classification. *Journal of Risk & Control* 3:17–65.

Kile, C. O., and M. E. Phillips. 2009. Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance* 24:35–58.

Kozak, S., S. Nagel, and S. Santosh. 2020. Shrinking the cross-section. *Journal of Financial Economics* 135:271–292.

Krishnan, J., and E. Press. 2003. The north american industry classification system and its implications for accounting research. *Contemporary Accounting Research* 20:685–717.

Moskowitz, T. J., and M. Grinblatt. 1999. Do industries explain momentum? *The Journal of finance* 54:1249–1290.

Rosch, E., and B. B. Lloyd. 1978. *Cognition and categorization*. L. Erlbaum Associates Hillsdale, NJ.

Saunders, N. C. 1999. The North American Industry Classification System: Change on the horizon. *Occupational Outlook Quarterly* 43:34–37.

S&P, and MSCI. 2002. Global Industry Classification Standard - A guide to the GICS Methodology .

Steinbach, M., G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. Tech. rep.

Villalonga, B. 2004. Does diversification cause the" diversification discount"? *Financial Management* pp. 5–27.

von den Hoff, S. 2022. Partitioning the Cross-Section of Stocks: The Economic Value of Statistical Clusters. *Available at SSRN 4214621* .

Weiner, C. 2005. The impact of industry classification schemes on financial research. *Available at SSRN 871173* .

Tables

TABLE 1: **Number of Firms per Sector: Descriptive Statistics**

This table reports the distribution of the number of firms within each economic sector for different classification methods. Only "functional" sectors with $N \geq 5$ firms are considered. The first panel groups schemes that yield between 9 and 11 sectors; the second one between 17 and 18; the third refers to 24 sectors. Ordinal numbers denote distribution percentiles. Data refer to the period July 1984 - June 2019.

	SIC9	FF10	GICS11	Cluster10	FF17	NAICS18	Cluster18	GICS24	Cluster24
Mean	331	298	271	300	186	164	171	126	128
Std. dev.	275	209	195	277	274	331	154	90	117
Skewness	1.48	0.73	0.81	1.14	3.1	3.76	1.14	1.01	1.2
Kurtosis	1.81	-0.74	-0.43	0.6	9.18	13.4	0.84	0.39	1.04
Min	7	28	21	5	22	5	5	5	5
1st	9	41	22	5	24	6	5	10	5
50th	248	237	196	211	103	76	131	104	97
99th	1180	784	740	1073	1383	1686	640	384	488
Max	1209	830	805	1193	1438	1832	719	404	559

TABLE 2: **Within-sector Explained Variation**

This table reports the average within-sector R^2 obtained by regressing the CAPM residuals of each firm i in cluster k on the cluster portfolio k, for different K corresponding to each industry classification. Data refer to the period July 1984 - June 2019.

K	SIC	NAICS	FF	GICS	Clustering
5	-	-	4.17%	-	7.92%
9	5.98%	-	-	-	9.04%
10	-	-	8.51%	-	9.31%
11	-	-	-	8.80%	9.40%
12	-	-	8.00%	-	9.58%
17	-	-	9.98%	-	10.02%
18	-	8.90%	-	-	10.10%
24	-	-	-	9.00%	10.40%
30	-	-	9.88%	-	10.77%
48	-	-	11.20%	-	11.75%

TABLE 3: **Across-cluster Correlation**

This table shows the average across-cluster correlation obtained computing the average of the pairwise correlation of each sector k with all other others, for different K corresponding to each industry classification. Data refer to the period July 1984 - June 2019.

K	SIC	NAICS	FF	GICS	Clustering
5	-	-	0.75	-	0.74
9	0.69	-	-	-	0.69
10	-	-	0.60	-	0.65
11	-	-	-	0.61	0.67
12	-	-	0.63	-	0.63
17	-	-	0.57	-	0.60
18	-	0.60	-	-	0.62
24	-	-	-	0.59	0.57
30	-	-	0.55	-	0.55
48	-	-	0.51	-	0.50

TABLE 4: **Sector Investing: In-sample and Out-of-sample SR**

This table shows the Sharpe Ratio of the tangency portfolio obtained using economic sectors as base assets for different K corresponding to each industry classification. Portfolio weights are computed at the end of June of each year. The left panel refers to the Sharpe Ratio over the entire sample period from July 1984 to June 2019. The right panel reports results for out-of-sample periods, where we keep the classification into clusters fixed over the next 12 months and compute the corresponding tangency portfolio.

	In-sample					Out-of-sample				
K	SIC	NAICS	FF	GICS	Clustering	SIC	NAICS	FF	GICS	Clustering
5	-	-	1.64	-	1.55	-	-	1.63	-	1.70
9	1.67	-	-	-	1.75	1.65	-	-	-	2.01
10	-	-	1.81	-	1.98	-	-	1.79	-	2.05
11	-	-	-	1.80	2.01	-	-	-	1.77	2.10
12	-	-	1.84	-	2.04	-	-	1.82	-	2.21
17	-	-	1.69	-	2.19	-	-	1.67	-	2.49
18	-	2.24	-	-	2.28	-	2.19	-	-	2.47
24	-	-	-	2.19	2.18	-	-	-	2.18	2.61
30	-	-	2.12	-	2.26	-	-	2.10	-	2.68
48	-	-	2.54	-	2.47	-	-	2.52	-	3.04

TABLE 5: **Sector Investing: Mean-Reversion strategies**

This table shows average excess returns (in percent), annualized Sharpe Ratios and alphas (in percent) with respect to the Fama and French (2015) plus momentum ('FF6') for equal-weighted mean-reversion strategies for different industrial classification schemes and different number of industries K. The corresponding t-statistic is reported on the right next to each metric, with bold numbers for values above conventional significance levels. The t-statistic for the Sharpe Ratio is computed following Bailey and Lopez de Prado (2012). The last column refers to the FF6 model. Strategies are rebalanced at the end of June of each year between 1984 and 2019.

	Avg. Excess Ret. (%)	t-stat	Ann. SR	t-stat	Alpha (%)	t-stat	Adj. R^2
FF5	0.03	0.58	0.1	0.59	-0.01	-0.4	0.5
SIC9	0	-0.08	-0.01	-0.09	-0.03	-0.78	0.46
FF10	0.02	0.39	0.07	0.39	-0.04	-1.28	0.55
GICS11	0.05	1.17	0.2	1.2	0.03	0.86	0.38
FF12	0.01	0.22	0.04	0.23	-0.04	-1.18	0.55
FF17	0.03	0.63	0.11	0.64	-0.01	-0.29	0.43
NAICS18	-0.06	-1.17	-0.2	-1.19	-0.05	-1.19	0.34
GICS24	0.03	0.82	0.14	0.84	0.01	0.45	0.4
FF30	0.01	0.24	0.04	0.25	-0.03	-0.72	0.38
FF48	0.03	0.54	0.09	0.56	-0.03	-0.63	0.42
Cluster5	0.45	4.17	0.71	4.6	0.37	4.46	0.47
Cluster9	0.41	4.15	0.71	4.43	0.35	4.42	0.44
Cluster10	0.4	4.17	0.71	4.45	0.36	4.68	0.45
Cluster11	0.43	4.57	0.78	4.91	0.39	5.25	0.45
Cluster12	0.47	4.97	0.85	5.42	0.42	5.73	0.46
Cluster17	0.5	5.39	0.92	5.81	0.43	6.46	0.54
Cluster18	0.53	5.68	0.97	6.1	0.48	6.91	0.52
Cluster24	0.56	5.08	0.87	5.84	0.52	6.33	0.5
Cluster30	0.56	5.31	0.91	6.07	0.52	6.62	0.5
Cluster48	0.57	5.86	1.01	6.61	0.54	7.64	0.53

Figures

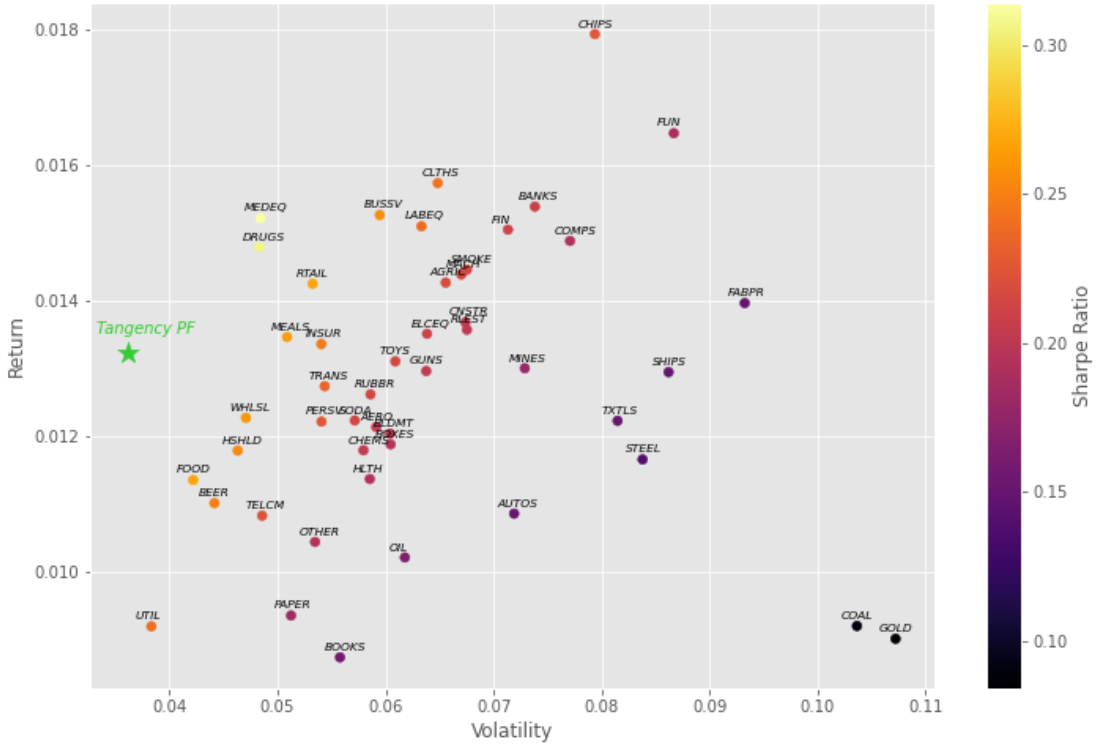

FIGURE 1: **Average Excess Returns and Standard Deviation, replicated Fama and French (1997) Industries**

This figure shows the average excess returns and the standard deviation for 48 industries built following Fama and French (1997). Lighter colors indicate higher Sharpe Ratios, as illustrated from the colored bar on the right. The star denotes the tangency portfolio (maximum Sharpe Ratio portfolio) that results from using the industries as base assets. Data refer to the period July 1984 - June 2019.

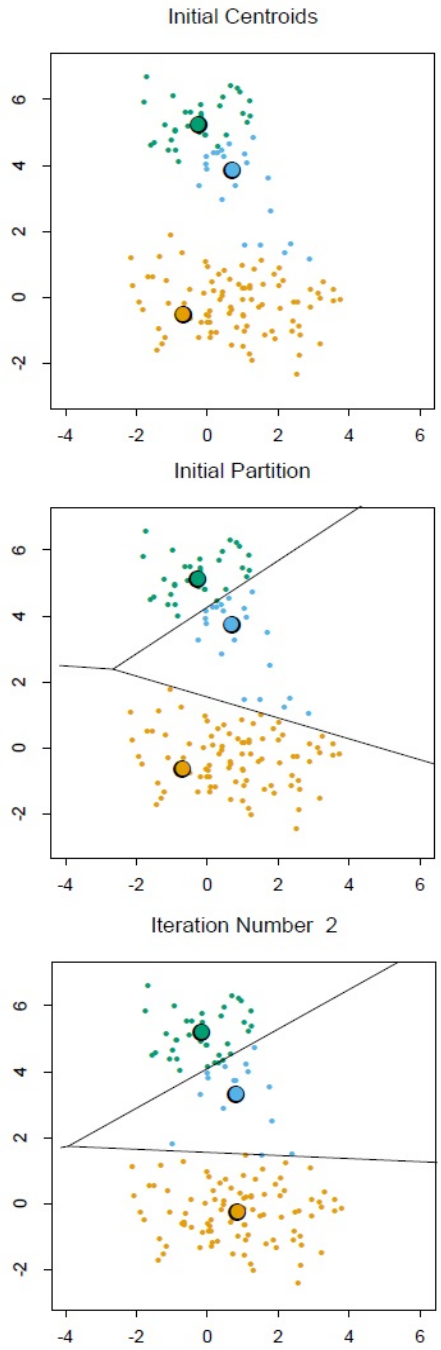

FIGURE 2: **Example of K-means Clustering**
This figure is adapted from Hastie et al. (2009) and shows an example of clustering through the K-means algorithm in a two-dimensional space, broken down into different steps.

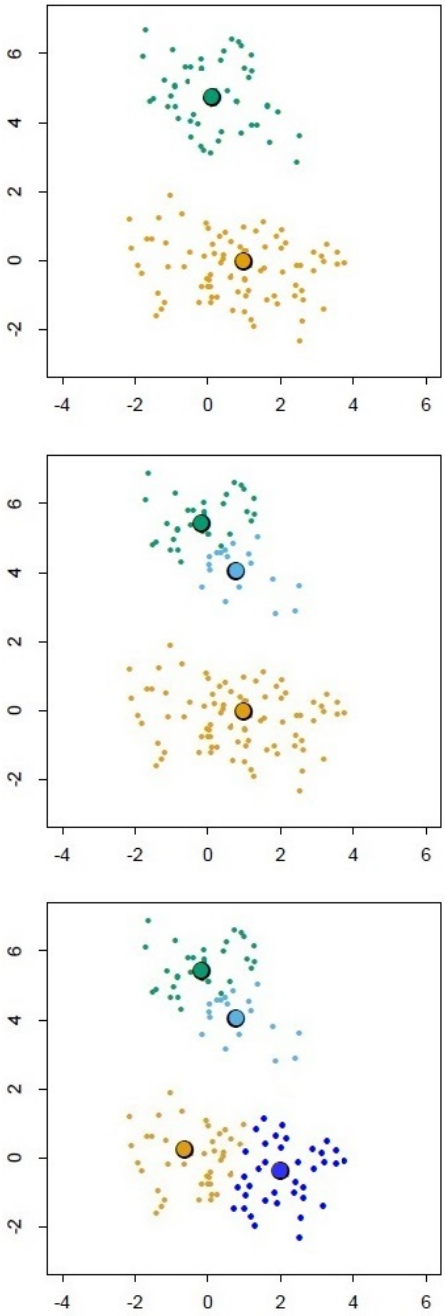

FIGURE 3: **Example of Bisecting K-means Clustering**
This figure shows an example of clustering through the bisecting K-means algorithm in a two-dimensional space, broken down into different steps.

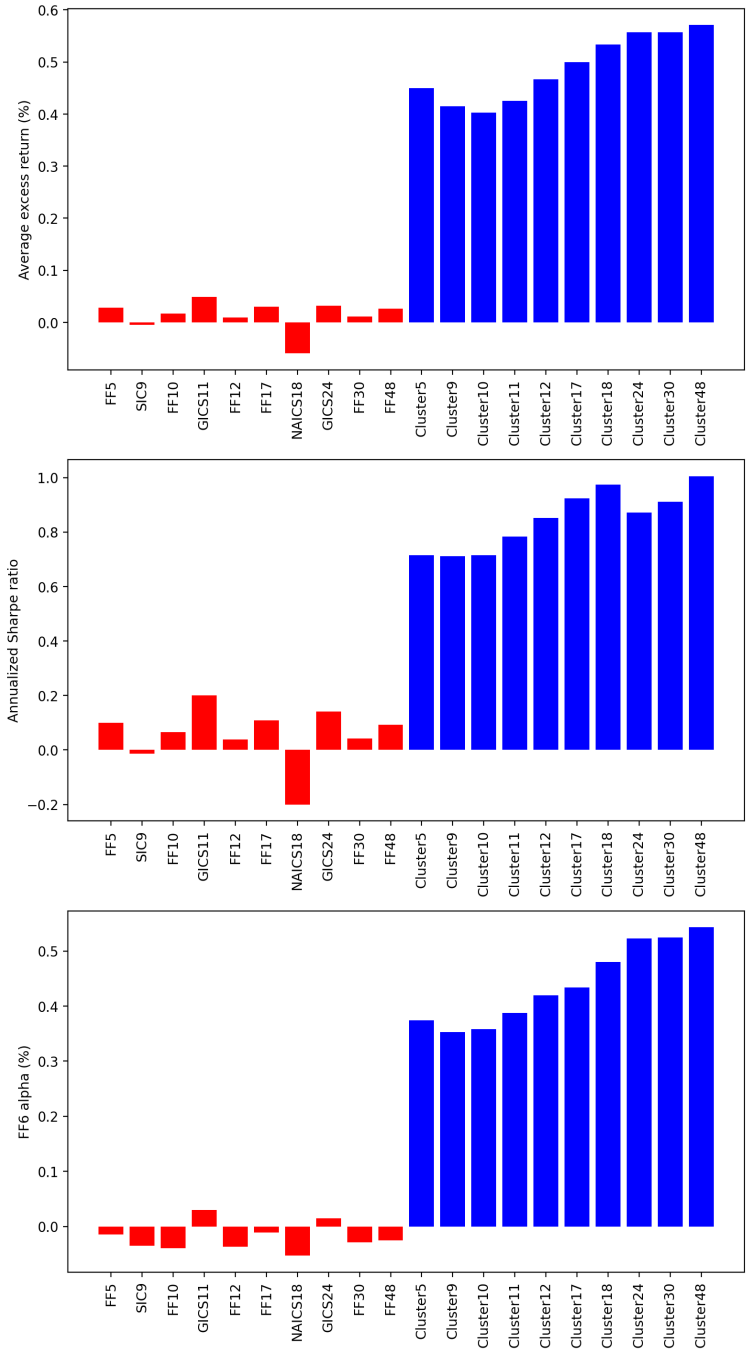

FIGURE 4: **Mean-reversion strategies**
This figure shows average excess returns, annualized Sharpe Ratios and the alpha with respect to the Fama and French (2015) plus momentum ('FF6') for mean-reversion strategies for different industrial classification schemes and different number of industries K. Data refer to the period July 1984 - June 2019.

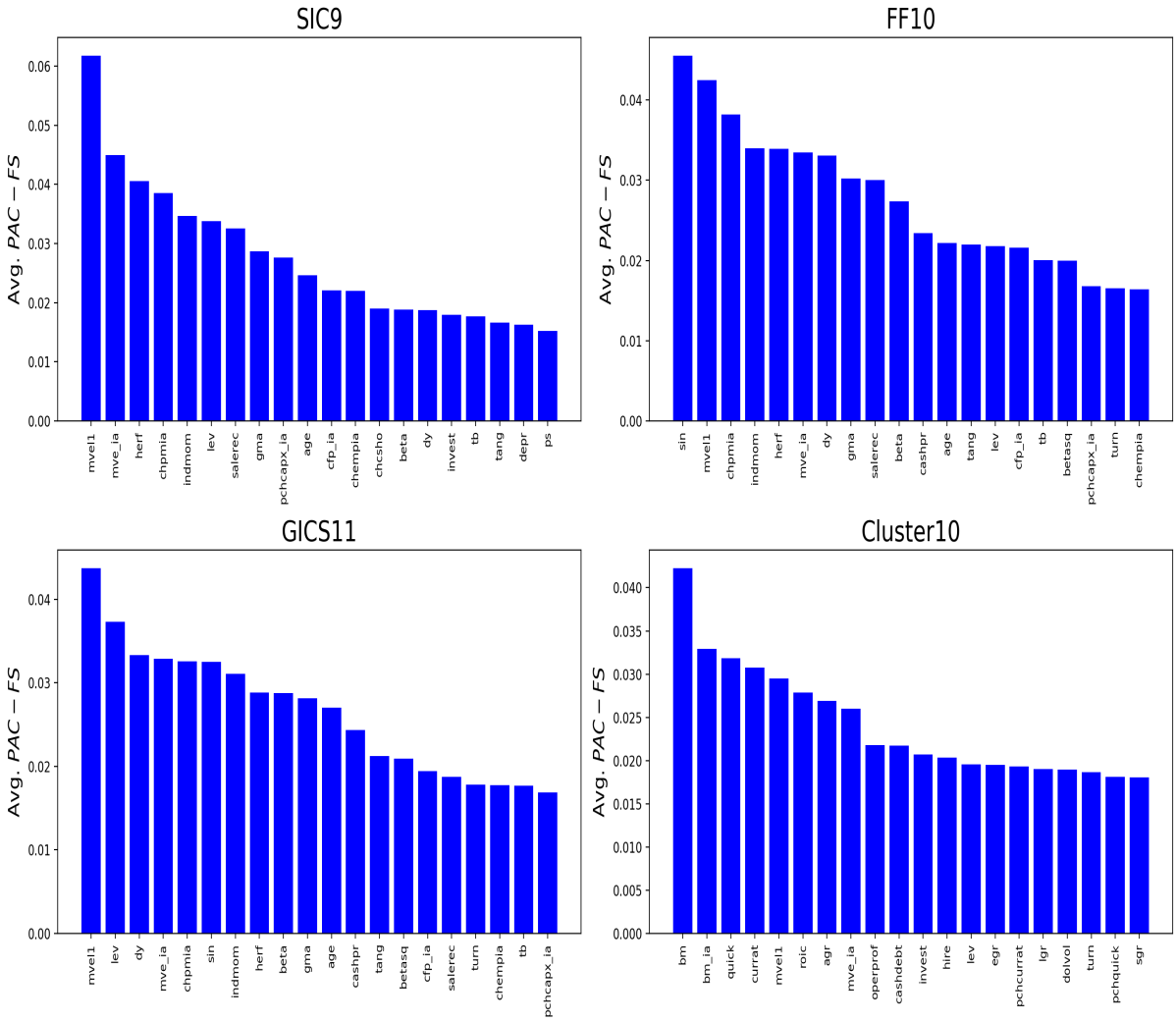

FIGURE 5: **Feature Importance,** $K = 10$
This figure shows the time-series average $PAC\text{-}FS_p$ for the twenty characteristics with the highest values, in descending order, for different K corresponding to each industry classification that yield a number of economic sectors between 9 and 11, as report in the titles above each panel. Data refer to the period July 1984 - June 2019.

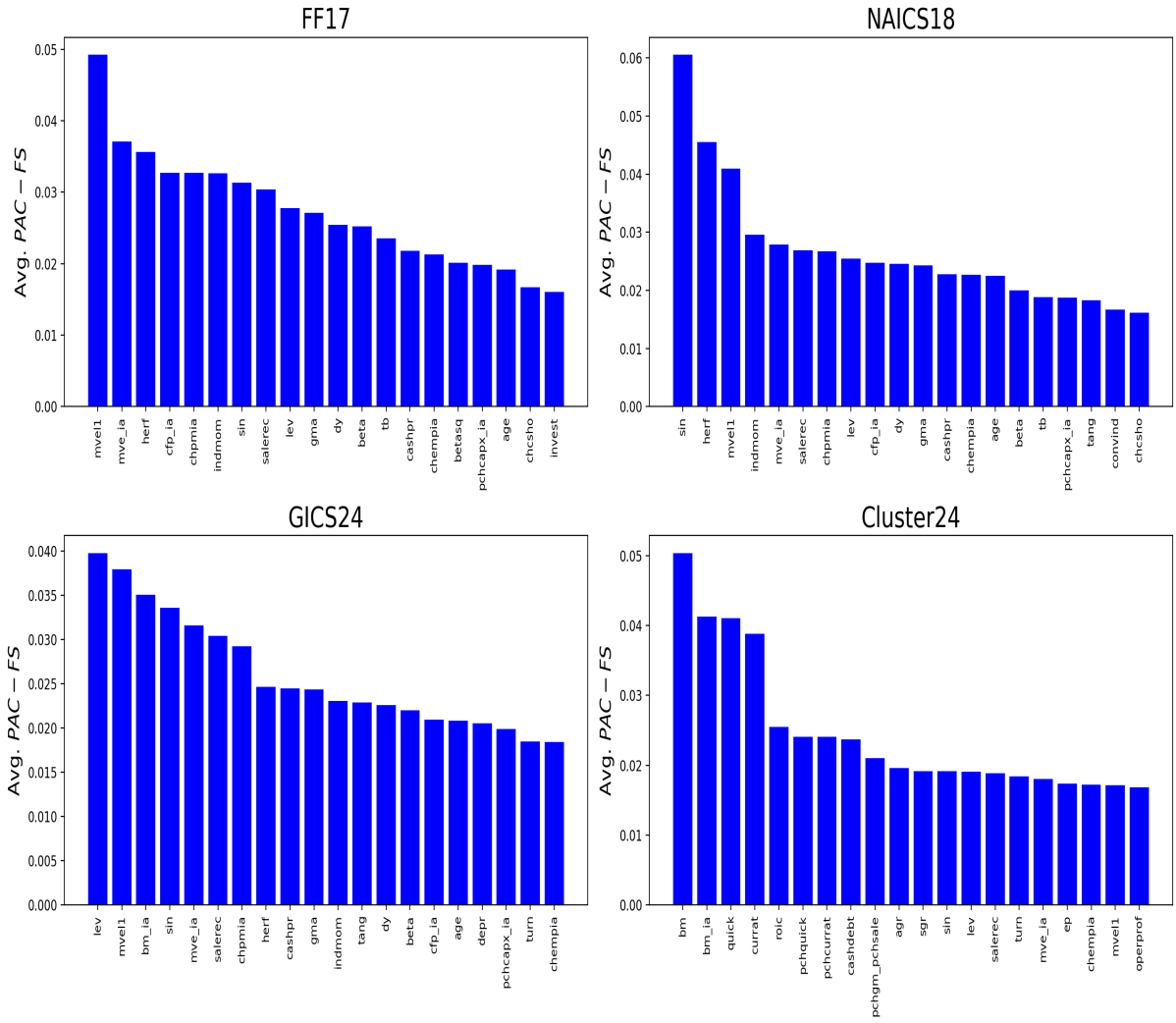

FIGURE 6: **Feature Importance,** $K = 18, 24$
This figure shows the time-series average $PAC\text{-}FS_p$ for the twenty characteristics with the highest values, in descending order, for different K corresponding to each industry classification that yield a number of economic sectors between 17 and 24, as report in the titles above each panel. Data refer to the period July 1984 - June 2019.

A Further Data Details

Table A.1: **Firm characteristics**

The table reports the 69 firm characteristics employed in the clustering algorithm. Data are obtained from Dacheng xiu's website (https://dachxiu.chicagobooth.edu/#research). Data refer to the period July 1984 - June 2019.

Acronym	Full name	Paper
absacc	Absolute accruals	Bandyopadhyay, Huang and Wirjanto (2010)
acc	Working capital accruals	Sloan (1996)
age	# years since first Compustat coverage	Jiang, Lee and Zhang (2005)
agr	Asset growth	Cooper, Gulen and Schill (2008)
baspread	Bid-ask spread	Amihud and Mendelson (1989)
beta	Beta	Fama and MacBeth (1973)
betasq	Beta squared	Fama and MacBeth (1973)
bm	Book-to-market	Rosenberg, Reid and Lanstein (1985)
bm ia	Industry-adjusted book to market	Asness, Porter and Stevens (2000)
cashdebt	Cash flow to debt	Ou and Penman (1989)
cashpr	Cash productivity	Chandrashekar and Rao (2009)
cfp	Cash flow to price ratio	Desai, Rajgopal and Venkatachalam (2004)
cfp ia	Industry-adjusted cash flow to price ratio	Asness, Porter and Stevens (2000)
chatoia	Industry-adjusted change in asset turnover	Soliman (2008)
chcsho	Change in shares outstanding	Pontiff and Woodgate (2008)
chempia	Industry-adjusted change in employees	Asness, Porter and Stevens (1994)
chinv	Change in inventory	Thomas and Zhang (2002)
chmom	Change in 6-month momentum	Gettleman and Marks (2006)
chpmia	Industry-adjusted change in profit margin	Soliman (2008)
convind	Convertible debt indicator	Valta (2016)
currat	Current ratio	Ou and Penman (1989)
depr	Depreciation / PP&E	Holthausen and Larcker (1992)
divi	Dividend initiation	Michaely, Thaler and Womack (1995)
divo	Dividend omission	Michaely, Thaler and Womack (1995)
dolvol	Dollar trading volume	Chordia, Subrahmanyam and Anshuman (2001)
dy	Dividend to price	Litzenberger and Ramaswamy (1982)
egr	Growth in common shareholder equity	Richardson, Sloan, Soliman and Tuna (2005)
ep	Earnings to price	Basu (1977)
gma	Gross profitability	Novy-Marx (2013)
herf	Industry sales concentration	Hou and Robinson (2006)
hire	Employee growth rate	Bazdresch, Belo and Lin (2014)
idiovol	Idiosyncratic return volatility	Ali, Hwang and Trombley (2003)
ill	Illiquidity	Amihud (2002)
indmom	Industry momentum	Moskowitz and Grinblatt (1999)
invest	Capital expenditures and inventory	Chen and Zhang (2010)
lev	Leverage	Bhandari (1988)
lgr	Growth in long-term debt	Richardson, Sloan, Soliman and Tuna (2005)
maxret	Maximum daily return	Bali, Cakici and Whitelaw (2011)
mom12m	12-month momentum	Jegadeesh (1990)
mom1m	1-month momentum	Jegadeesh and Titman (1993)
mom36m	36-month momentum	Jegadeesh and Titman (1993)
mom6m	6-month momentum	Jegadeesh and Titman (1993)
mvel1	Size	Banz (1981)
mve ia	Industry-adjusted size	Asness, Porter and Stevens (2000)
operprof	Operating profitability	Fama and French (2015)
pchcapx ia	Industry adjusted % change in capital expenditures	Abarbanell and Bushee (1998)
pchcurrat	% change in current ratio	Ou and Penman (1989)
pchdepr	% change in depreciation	Holthausen and Larcker (1992)
pchgm pchsale	% change in gross margin - % change in sales	Abarbanell and Bushee (1998)
pchquick	% change in quick ratio	Ou and Penman (1989)
pchsale pchrect	% change in sales - % change in A/R	Abarbanell and Bushee (1998)
pctacc	Percent accruals	Hafzalla, Lundholm and Van Winkle (2011)
pricedelay	Price delay	Hou and Moskowitz (2005)
ps	Financial statements score	Piotroski (2000)
quick	Quick ratio	Ou and Penman (1989)
rd	R&D increase	Eberhart, Maxwell and Siddique (2004)
retvol	Return volatility	Ang, Hodrick, Xing and Zhang (2006)
roic	Return on invested capital	Brown and Rowe (2007)
salecash	Sales to cash	Ou and Penman (1989)
salerec	Sales to receivables	Ou and Penman (1989)
sgr	Sales growth	Lakonishok, Shleifer and Vishny (1994)
sin	Sin stocks	Hong and Kacperczyk (2009)
sp	Sales to price	Barbee, Mukherji, and Raines (1996)
std dolvol	Volatility of liquidity (dollar trading volume)	Chordia, Subrahmanyam and Anshuman (2001)
std turn	Volatility of liquidity (share turnover)	Chordia, Subrahmanyam, andAnshuman (2001)
tang	Debt capacity/firm tangibility	Almeida and Campello (2007)
tb	Tax income to book income	Lev and Nissim (2004)
turn	Share turnover	Datar, Naik and Radcliffe (1998)
zerotrade	Zero trading days	Liu (2006)