Bagnara, Matteo; Goodarzi, Milad

Working Paper

Clustering-based sector investing

SAFE Working Paper, No. 397

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Bagnara, Matteo; Goodarzi, Milad (2023) : Clustering-based sector investing, SAFE Working Paper, No. 397, Leibniz Institute for Financial Research SAFE, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/273735

Matteo Bagnara | Milad Goodarzi

Clustering-Based Sector Investing

SAFE Working Paper No. 397 | August 2023

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Clustering-based Sector Investing

MATTEO BAGNARA[§] AND MILAD GOODARZI[¶]

ABSTRACT

Industry classification groups firms into finer partitions to help investments and empirical analysis. To overcome the well-documented limitations of existing industry definitions, like their stale nature and coarse categories for firms with multiple operations, we employ a clustering approach on 69 firm characteristics and allocate companies to novel economic sectors maximizing the within-group explained variation. Such sectors are dynamic yet stable, and represent a superior investment set compared to standard classification schemes for portfolio optimization and for trading strategies based on within-industry mean-reversion, which give rise to a latent risk factor significantly priced in the cross-section. We provide a new metric to quantify feature importance for clustering methods, finding that size drives 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 and Allianz Global Investors, Bockenheimer Landstraße 42-44, 60323 Frankfurt, Germany, Milad.goodarzi@allianzgi.com.

We thank Christian Schlag, Florian Heider, Alexander Hillert, Max Riedel, Jiri Woschitz and seminar and conference participants at the 12^{th} International Research Meeting in Business and Management, the 2023 Research Symposium on Finance and Economics, the 2023 Global Finance Conference, the 6^{th} Dauphine Finance PhD Workshop, the 2023 Financial Markets and Corporate Governance Conference and the Leibniz Institute for Financial Research SAFE for useful comments. We gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE.

This work represents the authors' personal opinions and does not necessarily reflect the views of the Allianz Global Investors.

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) that extends its influence into the realm of financial markets. In the domain of portfolio allocation, many investors adopt a "top-down" approach where they first identify broad asset classes and only afterwards decide how to allocate their funds among assets within a class (Barberis and Shleifer, 2003). Within this framework, the formation of economic sectors or industries holds paramount significance, dating back to the 1930s with the introduction of the Standard Industrial Classification (SIC) codes in the U.S. for statistical purposes across governmental agencies. Since then, this topic has been subject of intense scrutiny by both data vendors and academics. The sheer number of alternative classifications that have been proposed is a testament to the fact that existing ones either present substantial limitations or do not satisfy everybody's needs.

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 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.

Industries represent 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 decisions are influenced by industry categorization both at the institutional (Busse and Tong, 2012) and at the retail level (Jame and Tong, 2014). Furthermore, some financial phenomena often have a relevant industry-wide component, such as the *dot-com bubble* 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 70 papers in the Journal of Finance and 467 in the Journal of Financial Economics (Weiner, 2005). In Asset Pricing, 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 a model's validity.[1]

[1]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 French's website at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

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).[2] To summarize their investment performance, we focus on the maximum SR portfolio built using industries as base assets. During the period considered, it earns a monthly 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. Economic sectors have therefore a strong potential to deliver profitable investment strategies out of a contained number of assets.

Firms are usually classified according to four main systems: 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. Furthermore, managers might be able to manipulate industry categories to realize tangible benefits (Chen et al., 2016). 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 account for information from financial statements and investment research reports to assign firms to economic sectors that better satisfy the needs of investment professionals.

In our work, we follow and develop further this logic using firm characteristics to find economic sectors that offer better investment perspectives compared to existing systems. More in detail, we use bisecting K-means 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), similar also to the approach used in the return comovement literature (Piotroski and Roulstone, 2004). This approach, which is the main responsible for

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

our findings beyond the specific clustering algorithm employed, allows us to interpret our clusters as new economic sectors. Said differently, using firm characteristics as a 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 our new clustering-based sectors, we design a metric to describe how their structure changes over time and introduce a novel approach to quantify the relative importance of features in determining differences across clusters.

Our results are five-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 FF 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, cluster sectors deliver more attractive assets for mean-variance investors. The maximum Out-Of-Sample (OOS) SR portfolio from $K = 10$ sectors earns an annualized SR of 1.23, outperforming what one can obtain investing in any other set of industries (e.g. 0.81 with SIC codes, 0.73 with FF industries and 0.84 with GICS codes). This holds for all Ks considered. By creating more uniform firm groups, our method finds portfolios whose returns spread much more widely, resulting in better investment opportunities.

Third, trading strategies based on the mean-reversion of stock returns belonging to the same group are remarkably profitable for cluster sectors and give rise to a new risk factor that is priced in the cross-section. Averaging mean-reversion portfolios across 10 clusters delivers a monthly mean excess return of 0.46% and an alpha of 0.39% with respect to Fama and French (2015) plus momentum, both of which are highly statistically significant. The market price of risk of the corresponding factor portfolio estimated on individual stocks is 0.24%, i.e. roughly half the one of the market factor. Similar strategies based on any other industrial classification exhibit instead average returns close to zero and statistically insignificant. Clustering-based classification has therefore both a practical investment appeal and important implications for pricing.

Fourth, clustering-based sectors strike a good compromise between variability and stability. We propose a *Stability Index* aimed at describing changes over time in the relations linking firms across clusters and find that with 10 groups the factor structure remains stable roughly three quarters of the time. Existing industries, instead, are completely stale, a

feature that reveals responsible for their disappointing performance at capturing the return variation of firms belonging to the same economic sector.

Fifth, we provide a new metric that quantifies the contribution of firm characteristics to distinguishing clusters from each other, the *Proportion of (A)Cross-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 follows. After reviewing 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. Robustness tests are carried out in Section 7. Section 8 concludes.

2 Relation to the Literature

Our paper relates to various strands of literature. One concerns the validity and the goodness of different industry classification schemes. 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 regarding return covariation at increasingly finer levels of industry partitioning. Kile and Phillips (2009) argue 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 with 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 the 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) show 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. Jame and Tong (2014) 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 As-

set Pricing.[3] 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. 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, which has been recently applied also in Asset Pricing. Greengard et al. (2020) employ t-distributed stochastic neighborhood embedding (t-SNE) to cluster risk factors into 6 groups. In similar 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 represents an attractive set of investment assets accounting for the information contained in many firm characteristics. Specifically, we do not look for an optimal number of clusters based on prediction performance; rather, we fix K to match the number of sectors in 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

[3]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.

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 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, the NAICS, 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.[4] 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)).

[4]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

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.[5] 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.[6]

3.2 Limitations of Standard Classification Schemes

The classification schemes discussed above present several drawbacks. First of all, there might be some discrepancies across different databases. Guenther and Rosman (1994) find 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 shortcut of designating a primary industry for conglomerates determined by the product segment with the highest percentage of sales, which is also used by the Securities and Exchange Commission (SEC), lead investors to neglect a relevant part of firms' underlying economic operations, and managers exploit this fact to get into more "favorable" industries (Chen et al., 2016). 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 certain 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

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

[6] 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. 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).

contained in a large number of return predictors from the "factor zoo" literature (Cochrane, 2011), *de facto* treasuring the results that classification schemes based merely on the firm's 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.[7]

4 Data

Our dataset coincides with the updated version of Gu et al. (2020).[8] The original data includes 94 firm characteristics for CRSP stocks in the NYSE, AMEX and NASDAQ, that we merge with CRSP monthly 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.[9] 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.[10] The sample spans July 1984 to June 2019 for a total of 7052 firms. On 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), firm characteristics are cross-sectionally rank-transformed and mapped 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 two and the first four digits of GICS result in 11 and 24 sectors, respectively.[11][12] 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

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

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

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

[10] As we repeat our procedure 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.

[11] 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.

[12] Notice that some studies report an outdated number of industries for 2-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.

all the methods considered and to avoid having groups with very few firms.[13] 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 within-cluster commonality. Further, we design a measure to describe how much clusters change over time and 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.

5.1.1 K-means

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 straightforward way to formalize the notion of similarity between two observations is to use the Euclidean distance over all 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 cluster dissimilarity as the average of the Euclidean distance between every point belonging to a cluster and its centre (or *centroid*). 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

[13] For example, 2-digit SIC codes result in 67 industries; 2-digit NAICS 97 and 3-digit GICS 76. Notice that K. French provides a classification into 49 industries, too. We stop at 48 as this is the original categorization provided in Fama and French (1997) and thus likely the most frequently used one among the two. Adding one further industry does not convey any new finding and only exacerbates the problem of having few firms in some industries. Results are available on request.

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 ideal approach would be using combinatorial optimization, which evaluates every possible arrangement of the data into K clusters. Operationally, this is infeasible unless the dataset is very small. Therefore, more parsimonious 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 a limited fraction of all the possible assignments is examined, ensuring computational feasibility. To avoid getting stuck on local optima and ensure robustness, changing the initial cluster configuration becomes pivotal.

Among these iterative descent 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.

5.1.2 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.

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 with 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.[14] 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 more likely to produce groups with wildly different numbers 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 that of conditional portfolio sorts or regression trees, which facilitates its interpretation.

5.2 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 bisecting K-means requires a one-dimensional vector of inputs for each firm (column). To comply with this restriction, we aggregate the information using time-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.[15] 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

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

[15] We address this point later in Section 5.3.

[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, book-to-market and in general those derived from balance sheet information.

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 over the past year:

$$r^i_{t,\tau} = \alpha^i + \beta^i r^{Mkt}_{t,\tau} + \varepsilon^i_{t,\tau}$$
$$\hat{\varepsilon}_{i,t,\tau} = r^i_{t,\tau} - \hat{\alpha}^i - \hat{\beta}^i r^{Mkt}_{t,\tau}$$

where $\hat{}$ denotes OLS estimates of the coefficient of interest, t represents the year, τ the month in year t, $r^i_{t,\tau}$ are returns are in excess of the risk-free rate, and $r^{Mkt}_{i,t,\tau}$ is the return to the market factor.[17] 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", which 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.[18] We record the average R^2 across the firms in C_k. Finally, 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. Since 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), the resulting clusters can be thought of as new economic sectors.[19]

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 bisecting 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,

[17] Available at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. For the rest of the section, we remove the time index for readibility and use just $\hat{\varepsilon}^i$.

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

[19] 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. A similar procedure is used in the return comovement literature, e.g. Piotroski and Roulstone (2004) and Drake et al. (2017).

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.

5.3 Cluster Time-stability

In this Section we address the stability of clustering-based sectors over time. Do clusters change substantially in terms of composition from one period to another, i.e. do firms jump in and out of different clusters frequently or do the groups remain similar?[20]

While standard industry systems are known to be stale, we perform classification every year t based on firm characteristics known up to that point. *A priori* there is no strong statistical argument to assert whether clusters should remain very similar or change wildly over time. From an economic perspective, a case can be made that, if there exists persistent relations linking firms to each other, meaningful economic sectors should be relatively stable. We have reason to believe this holds for clusters based on the nature of the optimization process for determining the optimal cluster configuration. First, using a comprehensive set of 69 firm characteristics ensures that the algorithm is consistently based on features with the strongest association with returns, even if the predictive power of some covariates is time-varying, *ceteribus paribus*. Second, clusters are chosen in order to maximize the average within-cluster R^2, a well-established criterion for assessing the effectiveness of classifications systems that allows clusters to be interpreted as economic sectors. Therefore, we expect our approach to be able to uncover persistent latent structures over different iterations.

Beyond its economic meaning, investigating cluster time-stability is interesting also from a purely statistical point of view. To the best of our knowledge, no method has been proposed to measure it, mainly because clustering is *per se* an algorithm that abstracts from the time dimension.[21] Therefore, the technique we illustrate in the following is another relevant

[20] In cluster analysis, cluster stability refers stability to input randomization (Ben-David et al., 2006), i.e. the ability of an algorithm to provide groups that do not change much from one sample to another provided the data belong to the same latent clusters. In this paper, we focus instead on the temporal evolution of clusters.

[21] Clustering for time-series data makes use of methods such as dynamic time warping (Aghabozorgi et al., 2015). Applications to panel data have a much shorter history in the literature, with approaches that require a considerable amount of model structure compared to classical unsupervised methods (e.g. Ando and Bai (2017)). Rather than employing more intricate techniques, we opt for the conventional cluster analysis due to its comprehensibility, but we enhance its efficacy by introducing novel

contribution of this paper.

A challenge to face when investigating time-stability is the fact that cluster labels are inconsistent when the method is repeated, with no clear way to track them, as we explained above. Imagine the case where clusters have comparable compositions across time but different sizes, such that when 10% of firms in cluster C_1 move to cluster C_2 without C_1 gaining any new firms. Comparing the list of firms within C_1 between the two periods is impractical because of the change in size. To address this, we should fix a specific point in time to decide which is the "right" dimension of each cluster and make assumptions about the development of this unobservable quantity, which would be very arbitrary. We propose instead to focus on a stability measure that accounts for how strong is the relation linking one firm to the others across clusters.

Each year, when clusters are formed, we build all potential pairs among the firms in our sample, which amounts to $(N^2 - N)/2$. For each couple (i, j), for $i, j = 1, ..., N$ and $i \neq j$, we create a variable $S_t(i, j)$

$$S_t(i,j) = \begin{cases} +1, & \text{if } i \text{ and } j \text{ belong to the same cluster} \\ -1, & \text{otherwise} \end{cases} \tag{2}$$

Since clustering is performed annually, $S_t(i, j)$ does not change within the same year, hence we use only the year-index t. Then, we average this over time, and take its absolute value, $|\bar{S}(i,j)| = |1/T \sum_{t=1}^{T} S_t(i,j)|$. In this way, the resulting variable ranges between 0 and 1, with 1 indicating perfect stability (always together, $S_t(i,j) = +1 \ \forall \ t = 1, ..., T$ or never together, $S_t(i,j) = -1 \ \forall \ t = 1, ..., T$) and 0 total instability (on average, half the time together and half the time not together). For each firm $i = 1, ..., N$, we take the average across all firms $j = 1, ..., N$, $j \neq i$, obtaining $\mathcal{S}^i = 1/N \sum_{j \neq i, j=1}^{N} |\bar{S}(i,j)|$. Notice that since to measure stability being always together or never together is equally good, $\mathcal{S}^i$ measures how strong is the relation between i and the rest of the cross-section over time. The cross-sectional average of $\mathcal{S}^i$ also varies between 0 and 1 and captures the overall stability across all firms and clusters. We refer to this as *Stability Index (SI)*: $SI = 1/N \sum_{i=1}^{N} \mathcal{S}^i$.[22]

Let us consider a simple example to understand how the stability index works. Consider

approaches to explain its underlying mechanics, such as a measure for cluster time-stability and metric for feature importance as we show later on.

[22]The stability index could be computed also taking the average over all $|\bar{S}(i,j)|$, for $i \neq j, i, j = 1, ..., N$. We prefer to average this at the firm i level so we are able to assign a measure to each stock, thus preserving the dimensionality of our cross-section. We believe this feature makes $\mathcal{S}^i$ easier to interpret.

there are $N = 5$ firms, 2 clusters C_1, C_2 and 2 time periods, $t = 1, 2$. At $t = 1$,

$$\begin{aligned} \text{cluster}_1 &= \{\text{firm}_1, \text{firm}_2\} \\ \text{cluster}_2 &= \{\text{firm}_3, \text{firm}_4\, \text{firm}_5\} \end{aligned}$$

For $t = 2$, imagine the following scenarios.

1. In scenario 1, the clusters remain exactly the same:

$$\begin{aligned} \text{cluster}_1 &= \{\text{firm}_1, \text{firm}_2\} \\ \text{cluster}_2 &= \{\text{firm}_3, \text{firm}_4\, \text{firm}_5\} \end{aligned}$$

   We have perfectly stable clusters, hence $SI = 1$.

2. In scenario 2, the clusters change considerably as firm_2 moves to C_2 and firm_3 to C_1:

$$\begin{aligned} \text{cluster}_1 &= \{\text{firm}_1, \text{firm}_3\} \\ \text{cluster}_2 &= \{\text{firm}_2, \text{firm}_4\, \text{firm}_5\} \end{aligned}$$

   In this case, SI will be indeed lower: 0.4.

3. In the third scenario both firm_1 and firm_2 move to C_2 and $\text{firm}_3, \text{firm}_4$ move to C_1:

$$\begin{aligned} \text{cluster}_1 &= \{\text{firm}_3, \text{firm}_4\} \\ \text{cluster}_2 &= \{\text{firm}_1, \text{firm}_2\, \text{firm}_5\} \end{aligned}$$

   In this case $SI = 0.8$. Although at first we would be tempted to conclude that the migration of two couples represents great instability, the final result is that the two migrating firms are still together in the respective couple: even if clusters have "reshuffled", in scenario 3 they are more similar to the clusters at $t = 1$ than those of scenario 2. In fact, SI is lower in the second case than in the third case, because single-firm migrations are more disruptive when considering the relation across firms than seeing couple of firms migrating together.[23]

This example shows that SI is a simple and easy-to-grasp metric that nonetheless is sensitive enough to precisely describe a wide range of situations.

[23] Triplets or quadruplets could also be used. We prefer pairs as they represent the smallest group possible involving more than one firm, which is also appealing for interpretation.

An alternative way to capture stability of clustering methods is looking at how many times the relation between pairs of firms *changes*. Imagine the case where firm_1 and firm_2 are in the same cluster for, say, 5 years, and then not anymore for the following 5 years. Here $|\bar{S}(1,2)| = 0$, but one might argue that even if the link between the two firms breaks down after the first 5-year period for example due to structural changes affecting one firm, clusters are again perfectly stable afterwards. This looks very different from the situation in which firm_1 and firm_2 are in the same cluster only every other year, i.e. their relation keeps changing, where $|\bar{S}(1,2)| = 0$, too. Therefore, to complement *SI* we build another measure aimed at capturing changes in the relation among firms or "instability". For every couple (i,j), for $i \neq j, i, j = 1, ..., N$ we compute

$$G_t(i,j) = \begin{cases} 1, & \text{if } S_t(i,j) - S_{t-1}(i,j) \neq 0 \\ 0, & \text{if } S_t(i,j) - S_{t-1}(i,j) = 0 \end{cases} \tag{3}$$

The quantity $S_t(i,j) - S_{t-1}(i,j)$ can assume three values: 0 if firm i and j remain in the same cluster between $t-1$ and t; 2 (from not together to together) or -2 (from together to not together). The normalization in (3) maps $G_t(i,j)$ back to the $[0,1]$ interval, with 0 denoting stability and 1 instability. Then, we follow the same procedure used for the previous stability measure, which means we first compute the time-average of $G_t(i,j)$ for every couple and after that, for each firm i, we take the average across all other firms obtaining $\mathcal{G}^i = 1/N \sum_{j \neq i, j=1}^{N} \Big(1/(T-1) \sum_{t=2}^{T} G_t(i,j)\Big)$. Finally, the cross-sectional average of $\mathcal{G}^i$ is an overall measure of instability in the relations among firms, that we call *Instability Index (II)*: $II = 1/N \sum_{i=1}^{N} \mathcal{G}^i$.

5.4 Feature Importance in Clustering: A novel Metric

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 for cluster analysis, in contrast to other unsupervised ML paradigms like dimension reduction (e.g. Principal Component Analysis). 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 (A)Cross-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:

$$PAC\text{-}FS_p = \frac{S_p}{\sum_{p=1}^{P} S_p} \tag{4}$$

PAC-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 assess 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 (Cluster10). The middle panel shows figures for FF17, NAICS18, and $K = 18$ (Cluster18) clusters, and the right panel refers to GICS24 and 24 clusters (Cluster24). 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. Throughout the paper, whenever possible, we compute the results for each other intermediate K for clusters to make the comparisons more precise. 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 unequal 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 GICS11 and 331 for SIC9 in the left panel, 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.

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 from our clustering algorithm at the end of June of each year (i.e. when clusters are formed), and report time-series averages over the entire sample for different K in Table 2. When calculating the within-sector R^2, only firms with at least 10 available observations over the past 12 months are included to ensure meaningful statistics. Furthermore, only "functional" industries are considered.

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.[24] When K is higher, all classification schemes are better at summarizing the within-industry variation, as one might expect. The R^2 related to Cluster18 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 schemes, 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)). Further expanding the number of industries, we observe that the variation captured by our approach in excess of existing industries ("incremental variation") diminishes: for K sufficiently high, the groups of firms become so small and so homogeneous that differences among algorithms are less crucial than with fewer clusters. Nonetheless, the main takeaway is that informing the clustering procedure based on firm characteristics deliver tremendous improvements over static product-based classifications that overlook valuable firm-level information, yielding more accurate and nuanced economic sectors.

[24] 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.

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. Our analysis has demonstrated that firms comove more tightly within clustering-based sectors than within classical industries. Additionally, our cluster formation relies on a substantial number of return predictors from the literature. Incorporating characteristics with predictive power for stock returns offers a closer link to portfolio performance, suggesting potential 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 maximum SR portfolio and a trading strategy that exploits within-cluster mean-reversion.

6.3.1 Maximum Sharpe Ratio portfolio

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

$$\begin{aligned} \max_{\delta} \left\{ \frac{\delta' \mu}{\sqrt{\delta' \Sigma \delta}} \right\} \\ s.t.\ \delta' \mathbf{1} = 1 \\ 0 \leq \delta^i \leq 1 \ \forall \ i = 1, ..., K \end{aligned} \tag{5}$$

where μ represents a $K \times 1$ 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.[25] The second constraint imposes short-sale restrictions to avoid trading strategies with extreme positions that might be infeasible in practice. We solve the optimization problem each year after performing the clustering. In Table 3 we report OOS results for both standard industrial classifications and clustering-based sectors, where maximum SR portfolio returns are calculated over the next 12 months keeping the classification and the optimal weights δ unchanged as of the end of June of year t. This is an important exercise as it represents the outcome of feasible investment strategies beyond mere backtesting. Clustering surpasses all other standard industries for any number of sectors K.

[25] From here on we use excess returns of cluster portfolios, not CAPM residuals.

For instance, with $K = 10$, clustering-based sectors can be combined into a maximum SR portfolio that earns an annualized SR of 1.23, against 0.81, 0.73 and 0.84 for SIC9, FF10 and GICS11, respectively. Economic intuition would suggest that the portfolio SR should increase in K thanks to a larger investment universe. However, while this should be the case in-sample, it need not be necessarily so OOS due to estimation errors. It is comforting to acknowledge that the performance of clustering-based sector is strikingly stable with respect to K, varying between 1.20 ($K = 30$) and 1.36 ($K = 48$), whereas for example FF industries it varies more wildly (between 0.59 and 1 for $K = 30$ and $K = 5$, respectively). Even the maximum SR attainable using any set of classical industries is only 1, well below the minimum SR from clustering-based sectors.

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 pairs trading (Gatev et al., 2006) with the difference that the trading signal is built at the cluster level instead of at the pair level. 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 average excess return over the last year t, $\bar{r}_t^i$, is below the corresponding value-weighted cluster excess return $\bar{r}_t^k$ and that shorts stocks with returns above it. In other words, we form a value-weighted portfolio of the type

$$r_{t+1,\tau}^{k,MR} = \sum_{i \in C_k} w^i D_t(i,k) r_{t+1,\tau}^{i,k} \tag{6}$$

where

$$D_t(i,k) = \begin{cases} +1, & \text{if } \bar{r}_t^i < \bar{r}_t^k \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

and MR stands for Mean-Reversion. As before, τ denotes the month and t the year. Eq. (6) says that the portfolios uses the mean-reversion signal built at the end of the previous year, i.e. it is an implementable strategy.

Operationally, this strategy is formed at 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 4 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 yields between 0.42% and 0.51%. From Table 4, 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 4 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. Even controlling for several important risk factors, mean-reversion strategies that use clusters instead of standard industry classification remain highly profitable with alphas that generally 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 4. 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.

6.3.3 A within-cluster Mean-Reversion Risk Factor: WMR

Empirical evidence shows that the strategy described in Section 6.3.2 produces large and significant excess returns that cannot be explained by traditional Asset Pricing factors when built using the information contained in characteristics-based clusters. In this section, we take one step further and test the hypothesis that the within-cluster mean-reversion strategy represents a new risk factor that impacts the cross-section of stock returns. Since the strategy

is proxied by a single portfolio (the equal-weighted average of single-sector mean-reversion portfolios), we refer to this as a "factor", more specifically the *Within-cluster Mean-Reversion (WMR)* factor.

As test assets, we use the entire universe of individual stocks employed to construct economic sectors. Individual stock returns are among the highest hurdles for Asset Pricing models.[26] We have reason to believe that WMR is priced even among single stocks because it relies on a firm-level signal, i.e. the variable $D_t(i,k)$. This covariate measures whether firm i is relatively "cheap" or "expensive" if we interpret returns simply as firm i's appreciation from the previous period.[27] By leveraging stock i's position relative to its peers in the same economic sector, $D_t(i,k)$ encapsulates a potential large amount of information and simplifies it into a directional trading signal. This signal remains agnostic about the strength of the temporary deviation of firm i's performance in its industry, which is advantageous considering that individual stock returns typically exhibit low signal-to-noise ratios. Since WMR is the average mean-reversion strategy over all clusters, it essentially represents a portfolio that goes long all the stocks that are temporarily outperforming and short those that are momentarily underperforming. As such, WMR encompasses signals coming from the entire economy and thus potentially relevant for each firm, in similar spirit to what happens for the industry momentum (Moskowitz and Grinblatt, 1999), which is widely accepted as a robust anomaly.

To test our hypothesis, we perform Fama-MacBeth regressions (Fama and MacBeth (1973), FMB) on all stocks in our sample. The exercise is repeated for different industrial classification systems and different K, as in the previous section. Fama and French (2015) factors plus momentum are included as controls. In the first FMB step, we estimate factor betas using rolling windows of 60 months to reflect potential changes over time in the risk profiles of the stocks considered. We report the market price of risk (mpr) and the corresponding FMB t-statistic in Table 5. The second-to-last column refers to the WMR factor built from the economic sectors on the rows. The last column indicates the average adjusted R^2 for the second FMB step, i.e. the cross-sectional regressions. Among traditional factors, the market (Mkt-RF) and the size factors (SMB) are consistently priced in all the cases considered. The value (HML), the investment (CMA) and the momentum factors (MOM) are never priced, while the profitability factor (RMW) exhibits a negative mpr which is barely significant only in few cases. The column of interest regards WMR: similarly to what happens for the trading strategies in Section 6.3.2, the factor is never priced if built

[26] Typically, empirical Asset Pricing factors are tested against well-known groups of portfolios such as the 25 size-and-book-to-market-sorted portfolios from Fama and French (1992), which are known to have a strong factor structure. Individual stocks are instead rarely used as test assets. Modern methods often perform poorly in this case compared to portfolios, such as in Lettau and Pelger (2020). We use individual stocks to set a higher hurdle to test the WMR factor.

[27] Which is trivially the case if there are no dividends paid. Another advantage of using returns instead of prices to measure if a stock is cheap or expensive is that it mitigates the effect of upward drifts in prices and already existing price discrepancies among stocks belonging to the same group (Kakushadze, 2015).

from standard industry codes (the only exception is FF48, where it is significant just at the 10% level) but has always a significant mpr for clustering-based sectors. More specifically, this generally rises with K, ranging from 0.24% (Cluster10) to 0.33% (Cluster24), with a t-statistic always significant at the 5% level and significant at the 1% in 4 cases. The mpr reaches economically significant levels as high as 60% of the mpr of the market factor (0.52) and 70% of the size factor (0.44).

This Section shows that clustering-based economic sectors that exploit information from firm characteristics are relevant not only for investment purposes but also from a broader economic perspective. The WMR factor represents an aggregate measure of within-group relative performance that captures a novel risk factor relevant at the individual firm level.

6.4 Stability and Instability Index

What is the optimal outcome regarding the stability index described in Section 5.3? One the one hand, total instability would imply excessive dynamism: if firms were constantly shifting groups, we would conclude that there exists no strong link among firms over time, rendering clustering interpretation cumbersome. On the other hand, perfect stability is not desirable either because it would lead to staleness, which is one of the limitations of existing classification schemes. As there is a trade-off between the two extremes, the ideal outcome would be clusters that exhibit some variation over time, but not too much, because it would indicate that we achieved the goal of uncovering stable latent relations among firms akin to traditional industries. We expect this to happen because our clustering approach is not completely unsupervised but rather guided by the maximization of the explained variation within economic sectors, a strong indicator of "good" industries. We show that this is indeed what occurs in the subsequent analysis.

We calculate the variable $\mathcal{S}^i$ for each firm in the sample for each number of clusters K used in the previous exercises. Figure 5 shows the respective empirical density in a histogram. The red vertical dashed bar is the mean of the distribution, i.e. the stability index SI and the black bar refers to existing industries. In our sample, firms always belong to the same classical industry, such that the distribution of $\mathcal{S}^i$ collapses to 1. This happens for every classification different from clusters. The top-left panel illustrates the case $K = 5$. The stability index is 0.58 with a standard deviation of 0.11. On average, the relation between firm i and the rest of the cross-section remains the same more between half and two thirds of the time, which means that clustering-based sectors are quite dynamic but also far from being completely unstable. Increasing the number of clusters K shifts the entire density to the right towards one with a contemporaneous decrease in the dispersion around the

mean. With $K = 10$, $SI = 0.72$ (0.08 standard deviation). With $K = 24$, $SI = 0.87$ (0.04 standard deviation). The maximum occurs at $K = 48$, where $SI = 0.93$ (0.02 standard deviation). Intuitively, this pattern is easy to understand: when K rises, clusters become smaller and thus more homogeneous, such that the relations linking firms with each other also get stronger and it is less likely that firm couples are split during the cluster assignment. This development is also in line with the increase in the within-cluster R^2 documented in Table 2, which tended to converge across classification systems for large K, similarly to what the stability index does in Figure 5. These two results together hint at an interesting finding: the evolution of stability is inversely related to the incremental explained variation of our approach, because larger R^2 differences between clustering-based sectors and existing industries occur with more dynamic (i.e. less stable) clusters, e.g. with $K = 5$. Hence, the ideal level of stability that one wishes to achieve is eventually influenced by the number of clusters used, too.[28]

To complement these results, Figure 6 shows the empirical density of $\mathcal{G}^i$. The red vertical line is once again the average, i.e. the instability index II and the black line at zero represents any other existing classification system. For $K = 5$, $II = 0.29$ with 0.08 standard deviation, which means the relation between firm i and the rest of the cross-section changes sign around one fourth of the times. For $K = 48$, $II = 0.06$ with 0.02 standard deviation. The distribution shifts to the left towards zero and tightens up around the mean when K increases, for the same reasons that result in the shift to the right in the previous figure. The instability measure $\mathcal{G}^i$ is not the simple complement to 1 of the stability measure $\mathcal{S}^i$, as we explained in detailed above in the text. It is therefore comforting to observe patterns that can be interpreted in the same way, i.e. cluster instability decreases when the number of clusters increases, as within-cluster explained variation in excess of existing industries does.

Overall, the findings of this Section shows that clusters provide a favorable equilibrium between variability and stability, offering a level of time-variation that is meaningful but not excessive. This characteristic contributes to the perception of clusters as reliable representations of economic sectors.

6.5 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 one. In Figure 7 we report the time-series average of the $PAC\text{-}FS_p$ for the twenty variables with the largest values, in descending order. Following the same approach as above, we compare the results across different industry

[28] Notice that clustering-based sectors retain a certain level of dynamism even for the highest K. In that case, SI is high but still statistically significant from 1.

classifications for comparable K.[29] 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`) belongs to the top-6 features in all three cases. The amount of overall cross-cluster cross-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 for Cluster10. The book-to-market ratio (with the industry-adjusted variant) explain the highest portion of cross-cluster differences, i.e. more than 12%. The two next most important characteristics are financial liquidity ratios, in particular quick and current ratio (`quick` and `currat`). The importance of size is considerably downsized compared to standard classification systems. Furthermore, it is noteworthy that after the 9th characteristic, the $PAC\text{-}FS_p$ 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 9 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 8 we show results for FF17, NAICS18, GICS24 and Cluster24.[30] Once again, size (and its industry-adjusted version) plays one of the biggest roles in the cross-cluster cross-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, this is much less important for FF17. In fact, the first 5 positions for the latter change evidently with a larger number of industries. Something similar 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 is now the third most important feature whereas it did not even appear in Figure 7. `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,

[29] Of course, firm characteristics do not impact existing industries which are product-oriented. However, *PAC-FS* is still a useful metric to pin down distinguishing traits among economic sectors regardless the way in which they are formed.

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

followed by quick-and current ratio. Focusing on these 4 features, after which the $PAC\text{-}FS_p$ 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 ones: 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 8 is that the most salient features do not necessarily have lower $PAC\text{-}FS_p$ 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.

To sum up, 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. Together with the time-varying nature illustrated in the previous sections, these marked discrepancies are potentially responsible for 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 Robustness Tests

7.1 Short-term reversal, Long-term Reversal and Industry Momentum

One might be concerned that the trading strategy we describe in Section 6.3.2 might at least partially overlap with three well-known risk factors: short-term reversal, long-term reversal and industry momentum (Moskowitz and Grinblatt, 1999). We show here that this is not the case.

We augment the Fama and French (2015) model plus momentum with the short-term and long-term reversal factors from K. French's webpage together with the industry-momentum factor from Chen and Zimmermann (Forthcoming).[31] We thus obtain a 9-factor model, against which we regress the equally-weighted average of the within-cluster mean-reversion portfolios of above. We report alphas and adjusted R^2 in Table B.1 in the Appendix B. The results remain essentially unchanged compared to before, with significant alphas arising only with clustering-based sectors. Moreover, the magnitude increases with K, as economic intuition would suggest. Trading strategies that exploit within-cluster mean-reversion capture an effect that cannot be explained even adding three additional factors that specifically control for phenomena of reversal or momentum at the industry level.

We repeat the same test for the FMB regressions. We report the results in Table B.2 in the Appendix. The increase in the average cross-sectional R^2 compared to the previous case (from 25% to 30%) shows that including the additional controls helps capturing a higher portion of the return variation in the cross-section. The market and the size factors remain strongly significant in all the cases considered, with mpr of around 0.50% and 0.40%, respectively. The other factors, including the newly added ones, are not priced over the period considered, except for WMR. The latter has a significant mpr only for clustering-based sectors, as above, that tend to increase with K and is between 0.2% and 0.3%, an economically important value when compared to other factors.

In sum, controlling for short- and long-term reversal and industry momentum confirms the results observed earlier that within-cluster mean-reversion strategies are profitable only for clustering-based sectors and the corresponding WMR factor is priced in the cross-section.

7.2 Standard K-means

We have provided empirical evidence that our clustering technique boosts sector-investing profitability for both mean-variance investors (Section 6.3.1) and within-cluster mean rever-

[31] We thank the authors for making the data available at https://www.openassetpricing.com/data/.

sion strategies (Section 6.3.2). We attribute these improvements to the fact that we enhance an unsupervised ML technique (clustering) with the objective of finding the clusters that explain the highest proportion of return variation of the constituent firms. To support this claim, in this Section we conduct the same analysis as before using the standard K-means algorithm instead of bisecting K-means. If our pseudo-supervised clustering approach is the main driver of high sector-investing performance, we expect that changing the clustering algorithm will not change substantially the profitability of the investment strategies introduced earlier. We show that this is indeed the case.

Except for using K-means instead of bisecting K-means, every other choice concerning the data remains the same, including the firm characteristics, the frequency of cluster formation and so on. We start by reporting descriptive statistics related to the distribution of the number of firms per cluster in Table B.3. For comparability with Table 1, we show results for the same K and repeat the figures for classical industries. The average number of firms for Cluster10, Cluster18 and Cluster24 is 306, 172 and 130, respectively, which is quite similar to the average of the other classification systems with comparable number of sectors. Compared to bisecting K-means, the distribution now is more skewed to the right (e.g. skewness is 1.66 for Cluster10 against 1.14 for bisecting K-means) and the kurtosis is also higher, which dovetails with the known fact that bisecting K-means tends to produce clusters of more similar sizes. However, as in the previous case, all industrial classifications share a high degree of similarity regarding the number of firms per economic sector.

Table B.4 shows average within-cluster R^2 for different K. Also here we report results for standard industries for ease of comparison. Clustering achieves the highest R^2 regardless the number of industries. For instance, with $K = 10$, the average R^2 is 10.73%, which is even slightly higher than for bisecting K-means (9.31%). For larger K, the performance of the two clustering algorithms is nearly indistinguishable. Moreover, as with bisecting K-means, the amount explained rises with K and the incremental explanatory power over existing industries shrinks. Firm characteristics contain relevant information to build meaningful economic sectors beyond the specific the clustering algorithm used.

Let us now evaluate the investment performance of clusters formed using standard K-means. In Table B.5 we show the annualized OOS SR of the maximum SR portfolio built using economic sectors as base assets. Also in this case, clustering consistently outmatches existing industries across all K, with results that are very close to the bisecting K-means case: for example, the annualized SR for Cluster5 is 1.29 (1.23 for bisecting K-means) whereas it is 1.08 for Cluster30 (1.20 for bisecting K-means).

The second investment strategy we propose exploits within-cluster mean-reversion, as illustrated in Section 6.3.2. Table B.6 presents results for standard K-means for several

number of clusters. The tradings strategy is profitable in all cases, with average excess returns that are sizeable, ranging from 0.26% (Cluster5) to 0.49% (Cluster10), and strongly statistically significant. Mean-reversion strategies grow in K at first, peak at Cluster10, and afterwards alternate between increase and decrease, differently from the bisecting K-means case where average returns display a more pronounced growing pattern in K. Moreover, bisecting K-means tend to earn higher average returns overall. Similar patterns occur also for the annualized SR, which varies from 0.44 for $K = 5$ to 0.79 for $K = 48$, reaching its apex at 0.83 for $K = 30$, and for alphas with respect to Fama and French (2015) model plus momentum: the bisecting K-means is overall slightly better for these two metrics, too. Table B.7 shows mpr for the WMR factors corresponding to each mean-reversion strategy. Unsurprisingly, the conclusion are once again similar to those for the other clustering algorithm: standard K-means produces a risk factor that is priced in the cross-section, with an mpr that is roughly 0.30% per month, very similar to the bisecting K-means case, and that rises with K at first but then decreases and increases again. Here, the factor based on $K = 5$ sectors is not priced. Adding short-term and long-term reversal plus industry momentum to the model leaves the results substantially unchanged (see Table B.8). Together with Table B.5, the fact that all results are strongly significant and close in values to the bisecting K-means case brings solid support for the claim that it is not the clustering algorithm that matters for sector-investing, but rather the choice of guiding the approach through the objective of maximizing the within-cluster R^2.

As a final step to confirm this conjecture, we also investigate whether the choice of the clustering algorithm impacts cluster time-stability and feature importance.

Figure B.1 and B.2 report the distribution of the stability measure $\mathcal{S}^i$ and the instability measure $\mathcal{G}^i$, respectively, for different K. As above, the red dashed vertical line represents the mean, which is the Stability (Instability) index, and the black vertical line denotes existing industries, which are completely stale and thus collapse onto 1 (0). It is quite clear that there are no striking differences between the two clustering algorithm: the distribution of the stability (instability) measure shifts to the right (left) towards 1 (0) and tightens up as the number of economic sectors increases, in accordance to economic rationale. For instance, for $K = 5$, $SI = 0.58$ ($II = 0.29$) and for $K = 48$, $SI = 0.93$ ($II = 0.06$). The only noticeable fact is that now the stability (instability) measure can become essentially 1 (0) for a few firms, which did not happen for bisecting K-means. In other words, the latter offer clusters that are more dynamic than K-means. Overall, however, the main message remains unchanged: clustering-based economic sectors are relatively stable but at the same time offer variation that is absent in other classification systems.

Figure B.3 shows the time-series average of the $PAC - FS_p$ relative to the twenty features

with the highest values, in descending order. Results are displayed for the same K used in the bisecting K-means case. Two facts are worth noticing. First, for Cluster 10, the most important features are book-to-market (with its industry-adjusted version), which explains alone almost 7% of cross-cluster differences, and financial liquidity ratios (quick and current ratio). The top-4 features are the same as for bisecting K-means. Second, the main drivers of differences across clusters are remarkably stable when $K = 18$, a feature that was not found for classical industries. From Figure B.3 we can therefore conclude that the algorithm used to form clusters does not impact the role played by firm characteristics in differentiating clusters among each other.

In conclusion, using K-means instead of bisecting K-means generate very modest differences regarding the within-cluster return variation that can be explained by cluster-portfolios, the performance of maximum SR portfolios and mean-reversion-based strategies and factors, cluster time-stability and feature importance. As mentioned earlier in the text, bisecting K-means is often considered an improvement over the standard K-means for a variety of reasons. Our analysis finds indeed a marginally better performance in terms of investments strategies and cluster time-stability. However, the takeaway of this Section is that when we are interested in creating new economic sectors using return predictors, the mathematical approach employed to cluster firms is not decisive. Instead, the pivotal ingredient to the impressive results we document is the idea of augmenting cluster analysis with the objective of maximizing a standard measure for the goodness of industries, the within-cluster R^2.

8 Conclusion

Existing industry classifications present several drawbacks that have led the profession to look for new schemes over time. We treasure the result from the literature that industry codes accounting for the market perception of firms outperform those that are solely product-oriented. Using the information contained in a large number of stock return predictors, we propose a new classification method that links the power of clustering analysis with an easy-to-interpret Asset Pricing criterion, namely the maximum within-cluster explained variation, a natural metric to asses the effectiveness of industry assignments. This paradigm permits interpreting the resulting clusters as new economic sectors regardless the specific clustering algorithm employed. Results reveal strong potential both for research and investment purposes. Clustering surpasses all existent classification schemes in capturing the return variation of stocks in the same cluster, 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. Significant gains from clustering classification arise also when exploiting mean-reversion trading strategies, which deliver a latent risk factor that captures firms' performance relative to their peers and is significantly priced in the cross-section. We develop a simple yet effective measure to describe how the cluster structure evolves with each iteration, and show that clustering-based sectors strike a great balance between dynamism and time-stability. 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

Aghabozorgi, S., A. S. Shirkhorshidi, and T. Y. Wah. 2015. Time-series clustering–a decade review. *Information systems* 53:16–38.

Ando, T., and J. Bai. 2017. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association* 112:1182–1198.

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.

Ben-David, S., U. Von Luxburg, and D. Pál. 2006. A sober look at clustering stability. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pp. 5–19. Springer.

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, A. Y., and T. Zimmermann. Forthcoming. Open Source Cross Sectional Asset Pricing. *Critical Finance Review* .

Chen, H., L. Cohen, and D. Lou. 2016. Industry window dressing. *The Review of Financial Studies* 29:3354–3393.

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.

Drake, M. S., J. Jennings, D. T. Roulstone, and J. R. Thornock. 2017. The comovement of investor attention. *Management Science* 63:2847–2867.

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. 1992. The cross-section of expected stock returns. *the Journal of Finance* 47:427–465.

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.

Fama, E. F., and J. D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy* 81:607–636.

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

Gatev, E., W. N. Goetzmann, and K. G. Rouwenhorst. 2006. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies* 19:797–827.

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.

Jame, R., and Q. Tong. 2014. Industry-based style investing. *Journal of Financial Markets* 19:110–130.

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.

Lettau, M., and M. Pelger. 2020. Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies* 33:2274–2325.

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

Piotroski, J. D., and D. T. Roulstone. 2004. The influence of analysts, institutional investors, and insiders on the incorporation of market, industry, and firm-specific information into stock prices. *The accounting review* 79:1119–1151.

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: **Sector Investing: Out-of-sample SR**

This table shows the OOS Sharpe Ratio of the maximum SR 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 yearand the classification into clusters is kept fixed over the next 12 months. Data refer to the period July 1984 - June 2019.

K	SIC	NAICS	FF	GICS	Clustering
5	-	-	1.00	-	1.23
9	0.81	-	-	-	1.25
10	-	-	0.73	-	1.23
11	-	-	-	0.84	1.21
12	-	-	0.83	-	1.34
17	-	-	0.74	-	1.27
18	-	0.98	-	-	1.27
24	-	-	-	0.82	1.24
30	-	-	0.59	-	1.20
48	-	-	0.79	-	1.36

TABLE 4: **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	0.42	**3.9**	0.67	**4.3**	0.37	**4.12**	0.38
Cluster9	0.46	**4.41**	0.76	**4.67**	0.39	**4.38**	0.35
Cluster10	0.46	**4.86**	0.83	**5.12**	0.39	**4.61**	0.3
Cluster11	0.51	**5.65**	0.97	**6**	0.44	**5.58**	0.31
Cluster12	0.5	**4.72**	0.81	**5.39**	0.47	**5.36**	0.38
Cluster17	0.49	**5.14**	0.88	**5.85**	0.48	**6.09**	0.41
Cluster18	0.52	**5.71**	0.98	**6.28**	0.49	**6.46**	0.41
Cluster24	0.42	**4.19**	0.72	**4.68**	0.4	**5.1**	0.46
Cluster30	0.48	**4.6**	0.79	**5.26**	0.42	**5.35**	0.48
Cluster48	0.47	**4.87**	0.83	**5.44**	0.45	**6.38**	0.52

TABLE 5: **Sector Investing: WMR factor**

This table shows market prices of risk (in percent) estimated through FMB regressions (Fama and MacBeth, 1973) for the five Fama and French (2015) plus momentum plus WMR, the within-cluster mean-reversion portfolio, for different industrial classification schemes and different number of industries K. The corresponding t-statistic is reported in brackets, with bold numbers for values above conventional significance levels. The second-to-last column refers to the WMR factors built from the economic sectors on the rows. The last column refers to the average adjusted R^2 from the second step of the FMB regressions. WMR factors are rebalanced at the end of June of each year between 1984 and 2019.

	Mkt-RF	SMB	HML	RMW	CMA	Mom	WMR	Adj.R2
FF5	**0.53 (2.29)**	**0.43 (2.6)**	-0.04 (-0.24)	**-0.24 (-1.69)**	0 (-0.05)	-0.13 (-0.47)	0.03 (0.57)	0.25
SIC9	**0.52 (2.26)**	**0.43 (2.57)**	-0.04 (-0.25)	**-0.24 (-1.66)**	-0.01 (-0.05)	-0.15 (-0.57)	0.02 (0.29)	0.24
FF10	**0.53 (2.28)**	**0.42 (2.55)**	-0.04 (-0.22)	**-0.24 (-1.7)**	0 (0.04)	-0.15 (-0.55)	0.06 (1.11)	0.25
GICS11	**0.52 (2.26)**	**0.42 (2.55)**	-0.04 (-0.27)	-0.23 (-1.63)	0 (0.01)	-0.13 (-0.46)	0.05 (1.01)	0.25
FF12	**0.52 (2.25)**	**0.43 (2.59)**	-0.04 (-0.23)	**-0.24 (-1.68)**	0 (0.01)	-0.16 (-0.59)	0.06 (1.14)	0.25
FF17	**0.52 (2.28)**	**0.43 (2.58)**	-0.04 (-0.23)	**-0.24 (-1.7)**	-0.01 (-0.09)	-0.16 (-0.61)	0.09 (1.54)	0.25
NAICS18	**0.52 (2.25)**	**0.44 (2.63)**	-0.05 (-0.29)	**-0.25 (-1.74)**	0 (0)	-0.14 (-0.5)	0.04 (0.57)	0.24
GICS24	**0.51 (2.24)**	**0.43 (2.6)**	-0.05 (-0.28)	-0.23 (-1.62)	0 (-0.01)	-0.15 (-0.57)	0.06 (1.29)	0.25
FF30	**0.52 (2.27)**	**0.42 (2.51)**	-0.05 (-0.28)	-0.22 (-1.54)	0 (-0.01)	-0.16 (-0.58)	0.07 (1.29)	0.25
FF48	**0.53 (2.29)**	**0.43 (2.56)**	-0.05 (-0.32)	-0.23 (-1.63)	-0.01 (-0.05)	-0.16 (-0.59)	**0.1 (1.73)**	0.25
Cluster5	**0.52 (2.25)**	**0.43 (2.57)**	-0.04 (-0.26)	-0.22 (-1.59)	0 (-0.04)	-0.16 (-0.57)	**0.28 (2.06)**	0.25
Cluster9	**0.53 (2.3)**	**0.42 (2.51)**	-0.05 (-0.33)	-0.25 (-1.71)	-0.01 (-0.12)	-0.15 (-0.55)	**0.33 (2.65)**	0.25
Cluster10	**0.52 (2.27)**	**0.43 (2.54)**	-0.05 (-0.32)	**-0.24 (-1.69)**	-0.01 (-0.09)	-0.15 (-0.56)	**0.24 (2.11)**	0.25
Cluster11	**0.51 (2.23)**	**0.44 (2.59)**	-0.05 (-0.33)	**-0.24 (-1.67)**	-0.02 (-0.14)	-0.14 (-0.51)	**0.24 (2.17)**	0.25
Cluster12	**0.51 (2.24)**	**0.43 (2.57)**	-0.05 (-0.31)	-0.23 (-1.63)	-0.01 (-0.08)	-0.14 (-0.53)	**0.3 (2.31)**	0.25
Clusetr17	**0.51 (2.22)**	**0.42 (2.49)**	-0.06 (-0.33)	-0.21 (-1.5)	0 (-0.03)	-0.12 (-0.44)	**0.3 (2.69)**	0.25
Cluster18	**0.51 (2.22)**	**0.42 (2.51)**	-0.05 (-0.28)	-0.22 (-1.56)	0 (0.04)	-0.13 (-0.48)	**0.24 (2.11)**	0.25
Cluster24	**0.52 (2.29)**	**0.41 (2.46)**	-0.07 (-0.39)	-0.23 (-1.63)	0 (-0.04)	-0.12 (-0.45)	**0.33 (2.67)**	0.25
Cluster30	**0.52 (2.27)**	**0.42 (2.5)**	-0.06 (-0.38)	**-0.24 (-1.66)**	-0.01 (-0.09)	-0.15 (-0.54)	**0.31 (2.44)**	0.25
Cluster48	**0.52 (2.24)**	**0.4 (2.41)**	-0.04 (-0.25)	-0.22 (-1.55)	0 (0.04)	-0.12 (-0.44)	**0.31 (2.65)**	0.25

Figures

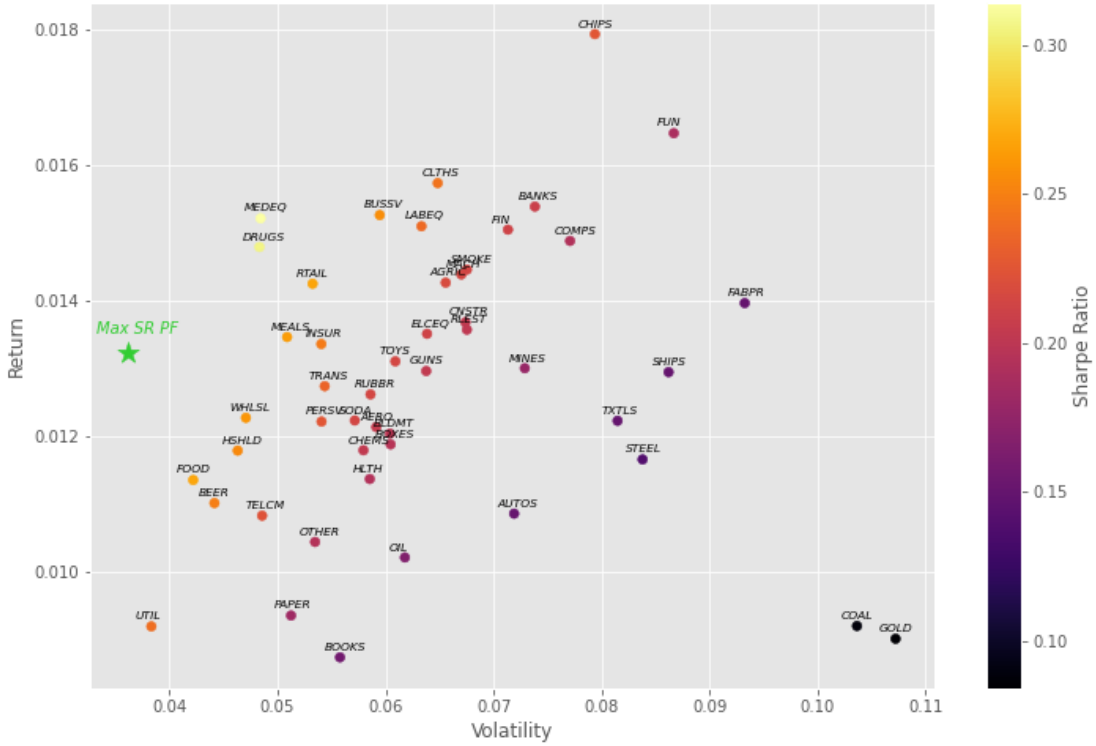

FIGURE 1: **Average Excess Returns and Standard Deviation, replicated Fama and French (1997) Industries**

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 maximum SR portfolio that results from using the industries as base assets. Data refer to the period July 1984 - June 2019.

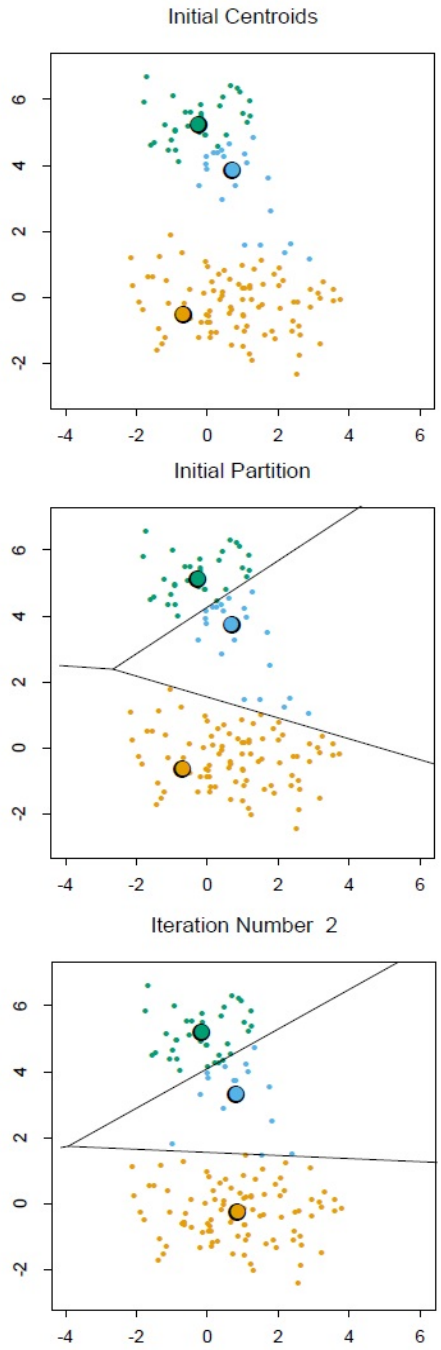

FIGURE 2: **Example of K-means Clustering**
This figure is adapted from Hastie et al. (2009) and shows an example of clustering through the K-means algorithm in a two-dimensional space, broken down into different steps.

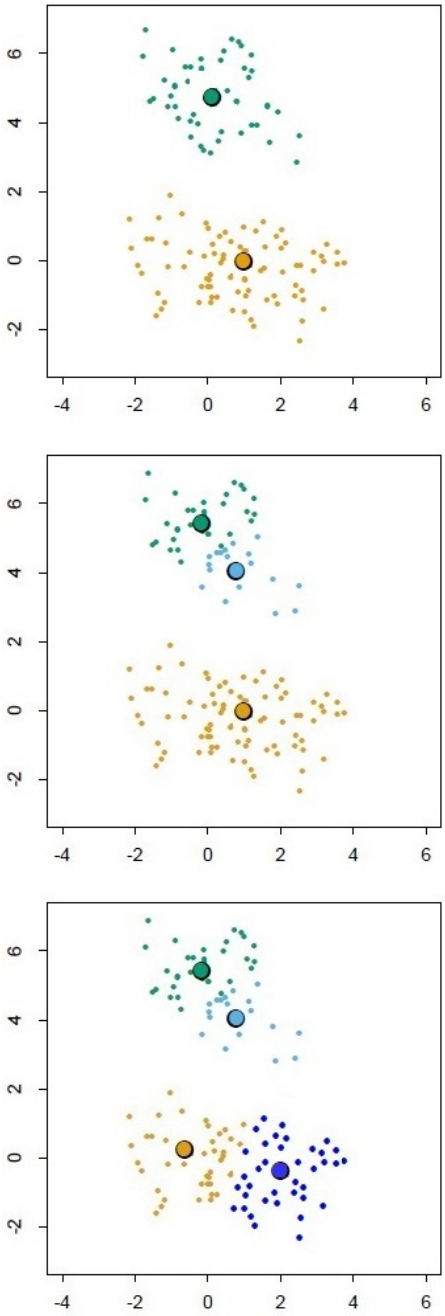

FIGURE 3: **Example of Bisecting K-means Clustering**
This figure shows an example of clustering through the bisecting K-means algorithm in a two-dimensional space, broken down into different steps.

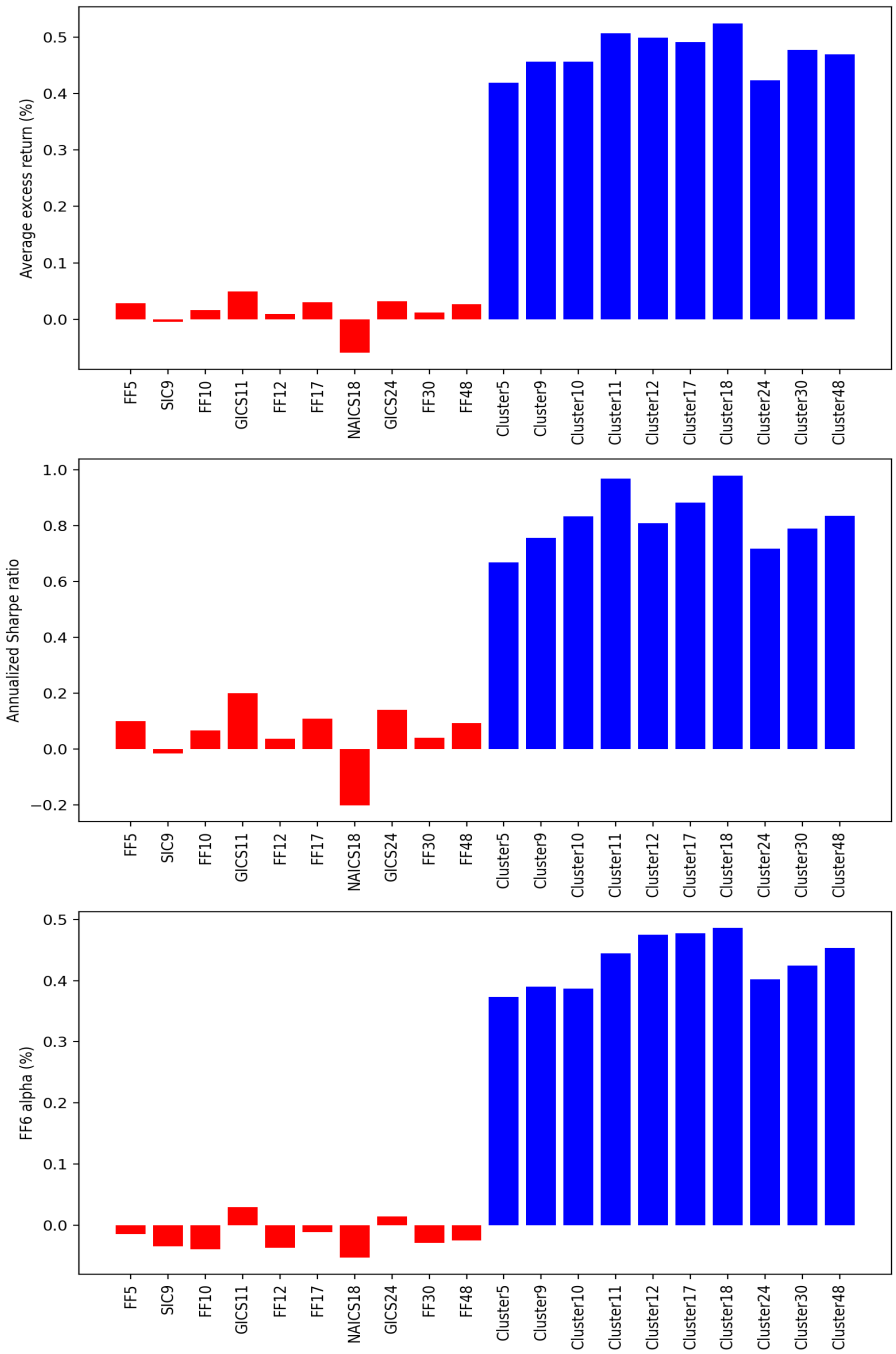

FIGURE 4: **Mean-reversion strategies**
This figure shows average excess returns, annualized Sharpe Ratios and the alpha with respect to the Fama and French (2015) plus momentum ("FF6") for mean-reversion strategies for different industrial classification schemes and different number of industries K.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

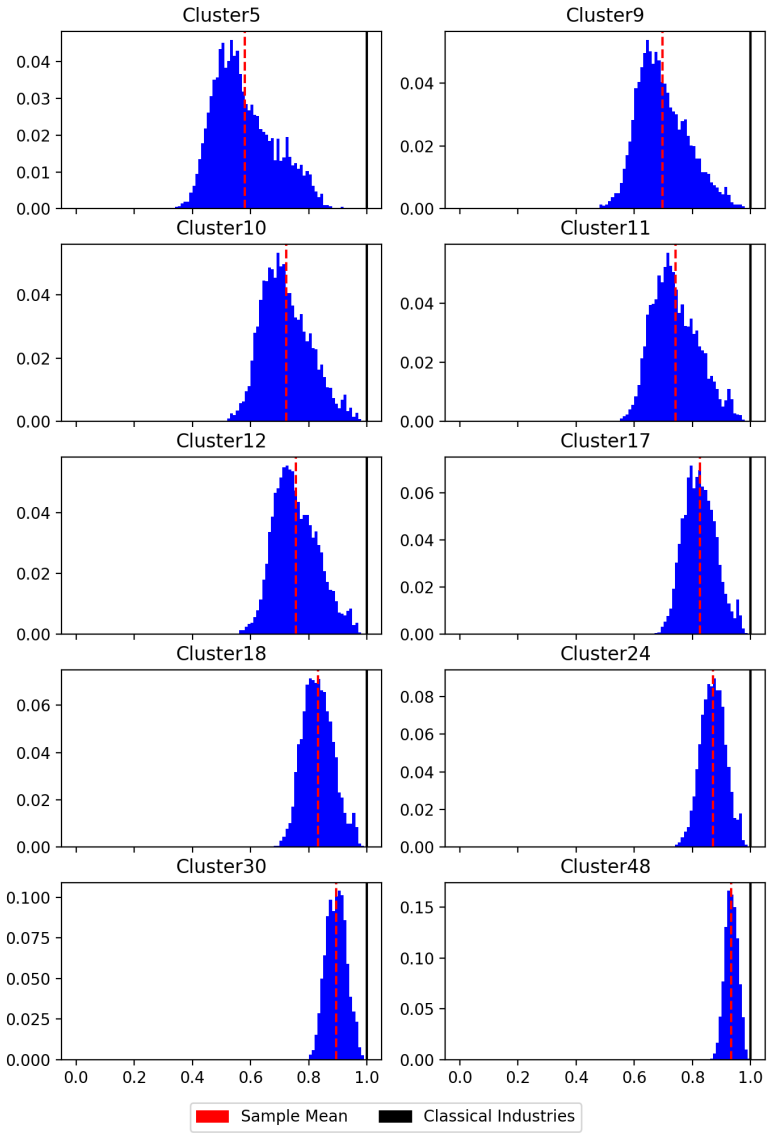

FIGURE 5: **Cluster Time-stability and Stability Index**
This figure shows the empirical density of the firm-level stability measure $\mathcal{S}^i$ for $i = 1, ..., N$ for different number of clusters K. The red vertical bar denotes the cross-sectional mean, i.e. the stability index *SI*. The black bar refers to any other classification systems (the density collapses onto the value 1). Data refer to the period July 1984 - June 2019.

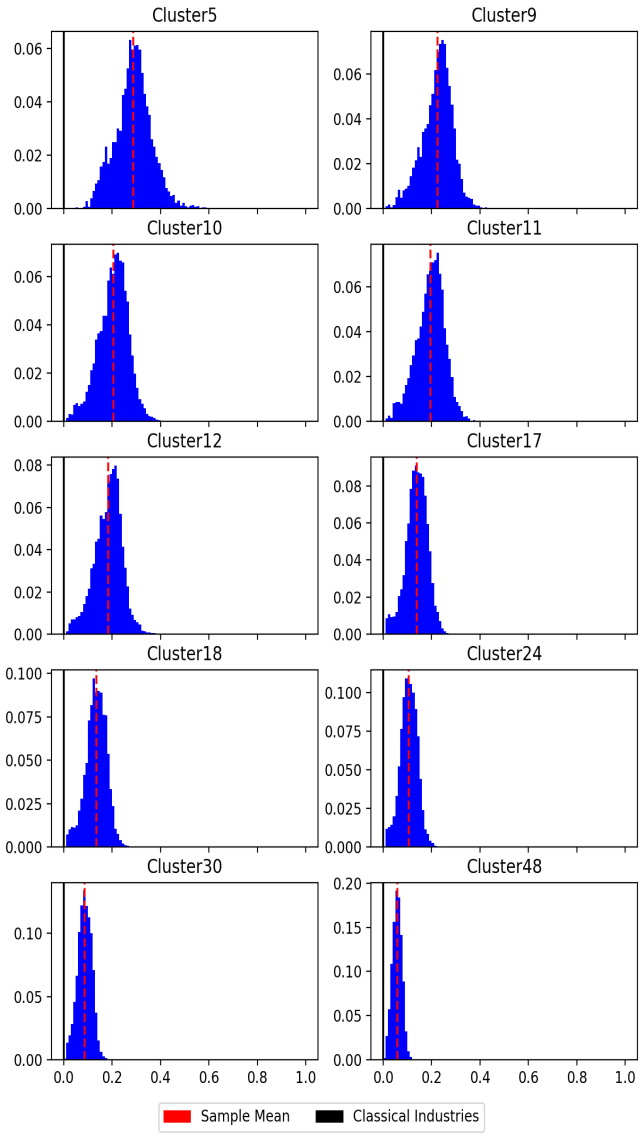

FIGURE 6: **Cluster Time-Instability and Instability Index**
This figure shows the empirical density of the firm-level instability measure $\mathcal{G}^i$ for $i = 1, ..., N$ for different number of clusters K. The red vertical bar denotes the cross-sectional mean, i.e. the instability index *II*. The black bar refers to any other classification systems (the density collapses onto the value 0). Data refer to the period July 1984 - June 2019.

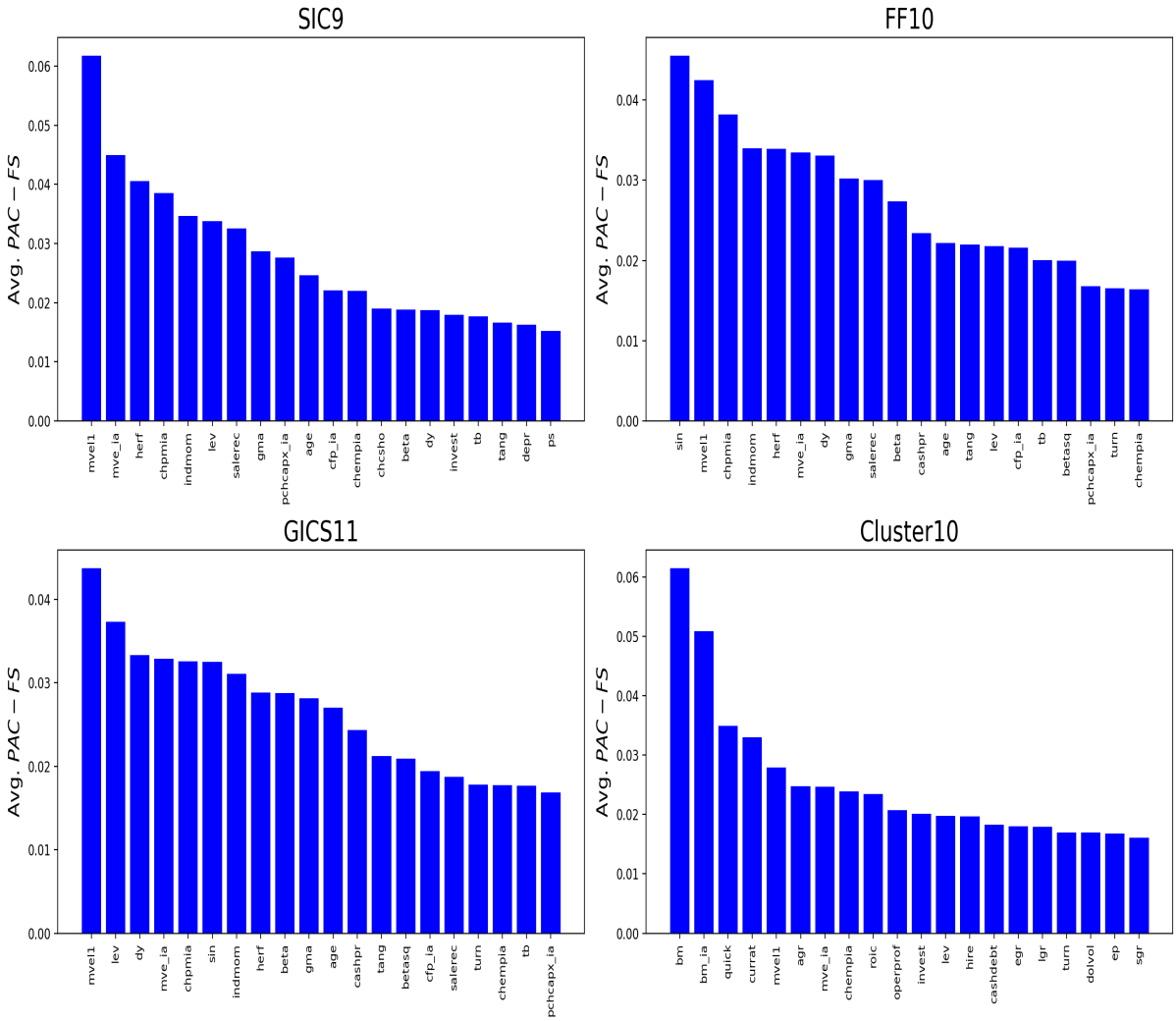

FIGURE 7: **Feature Importance,** $K = 10$
This figure shows the time-series average $PAC\text{-}FS_p$ for the twenty characteristics with the highest values, in descending order, for different K corresponding to each industry classification that yields a number of economic sectors between 9 and 11, as report in the titles above each panel. Data refer to the period July 1984 - June 2019.

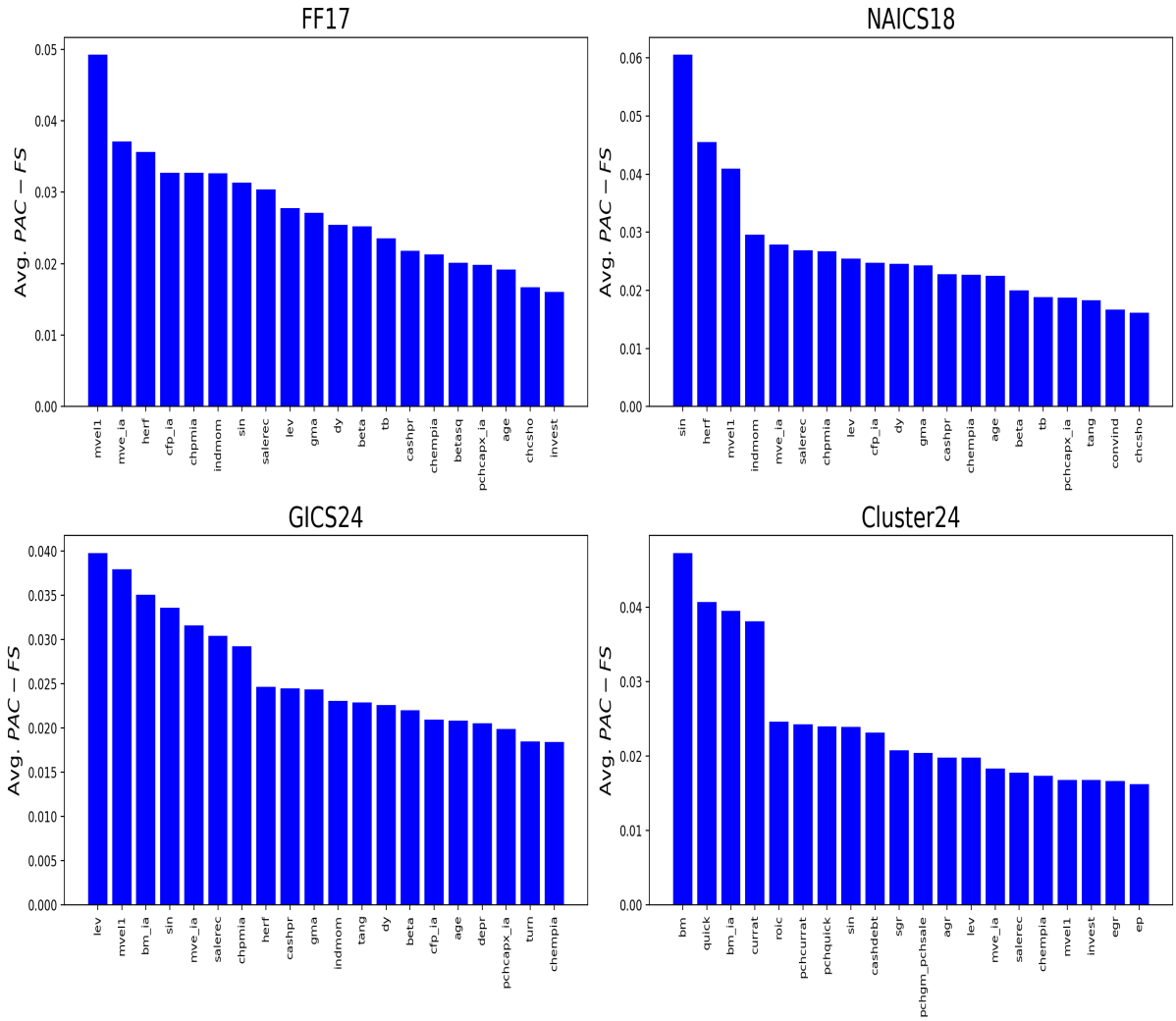

FIGURE 8: **Feature Importance,** $K = 18, 24$
This figure shows the time-series average $PAC\text{-}FS_p$ for the twenty characteristics with the highest values, in descending order, for different K corresponding to each industry classification that yields a number of economic sectors between 17 and 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)

B Robustness Tests

TABLE B.1: **Mean-Reversion strategies with additional controls**

This table shows alphas (in percent) with respect to the Fama and French (2015) plus momentum plus short-term and long-term reversal plus industry-momentum (Moskowitz and Grinblatt, 1999) 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 adjusted R^2 is reported in the last column. Strategies are rebalanced at the end of June of each year between 1984 and 2019.

	Alpha (%)	t-stat	Adj. R^2
FF5	-0.02	-0.46	0.5
SIC9	-0.04	-0.89	0.46
FF10	-0.04	-1.32	0.55
GICS11	0.03	0.76	0.38
FF12	-0.04	-1.2	0.55
FF17	-0.01	-0.27	0.43
NAICS18	-0.05	-1.1	0.34
GICS24	0.01	0.45	0.4
FF30	-0.03	-0.67	0.38
FF48	-0.02	-0.6	0.41
Cluster5	0.39	**4.27**	0.38
Cluster9	0.41	**4.68**	0.37
Cluster10	0.4	**4.83**	0.32
Cluster11	0.46	**5.86**	0.33
Cluster12	0.49	**5.52**	0.39
Cluster17	0.48	**6.18**	0.41
Cluster18	0.49	**6.42**	0.41
Cluster24	0.4	**5.06**	0.46
Cluster30	0.43	**5.4**	0.49
Cluster48	0.45	**6.39**	0.53

TABLE B.2: **Sector Investing: WMR factor with additional controls**

This table shows market prices of risk (in percent) estimated through FMB regressions (Fama and MacBeth, 1973) for the five Fama and French (2015) plus momentum plus short-term and long-term reversal plus industry momentum (Moskowitz and Grinblatt, 1999) plus WMR, the within-cluster mean-reversion portfolio, for different industrial classification schemes and different number of industries K. The corresponding t-statistic is reported in brackets, with bold numbers for values above conventional significance levels. The second-to-last column refers to the WMR factors built from the economic sectors on the rows. The last column refers to the average adjusted R^2 from the second step of the FMB regressions. WMR factors are rebalanced at the end of June of each year between 1984 and 2019.

	Mkt-RF	SMB	HML	RMW	CMA	Mom	STR	LTR	INDMOM	Ind	Adj. R^2
FF5	**0.54 (2.39)**	**0.41 (2.45)**	-0.02 (-0.1)	-0.22 (-1.56)	0.01 (0.12)	-0.06 (-0.24)	0.13 (0.69)	0.2 (1.47)	0.24 (1.08)	0.03 (0.54)	0.3
SIC9	**0.53 (2.34)**	**0.4 (2.43)**	-0.02 (-0.09)	-0.22 (-1.55)	0.01 (0.12)	-0.08 (-0.29)	0.14 (0.71)	0.19 (1.46)	0.23 (1.04)	0.02 (0.36)	0.3
FF10	**0.54 (2.37)**	**0.4 (2.42)**	-0.02 (-0.11)	-0.23 (-1.57)	0.02 (0.16)	-0.08 (-0.28)	0.13 (0.65)	0.19 (1.47)	0.24 (1.08)	0.05 (0.99)	0.3
GICS11	**0.54 (2.37)**	**0.4 (2.39)**	-0.02 (-0.15)	-0.22 (-1.53)	0.01 (0.1)	-0.06 (-0.21)	0.1 (0.54)	0.18 (1.34)	0.25 (1.13)	0.05 (0.98)	0.3
FF12	**0.54 (2.37)**	**0.41 (2.45)**	-0.02 (-0.11)	-0.23 (-1.59)	0.02 (0.14)	-0.08 (-0.3)	0.12 (0.64)	0.2 (1.49)	0.23 (1.06)	0.05 (1.05)	0.3
FF17	**0.55 (2.41)**	**0.4 (2.41)**	-0.02 (-0.1)	-0.22 (-1.56)	0.01 (0.1)	-0.07 (-0.28)	0.12 (0.6)	0.19 (1.42)	0.24 (1.12)	0.08 (1.48)	0.3
NAICS18	**0.54 (2.35)**	**0.42 (2.5)**	-0.03 (-0.17)	-0.23 (-1.61)	0.02 (0.14)	-0.06 (-0.24)	0.12 (0.64)	0.19 (1.45)	0.24 (1.1)	0.02 (0.38)	0.3
GICS24	**0.53 (2.32)**	**0.41 (2.47)**	-0.03 (-0.18)	-0.22 (-1.56)	0.01 (0.08)	-0.08 (-0.29)	0.11 (0.56)	0.18 (1.37)	0.23 (1.08)	0.05 (1.1)	0.3
FF30	**0.54 (2.37)**	**0.39 (2.37)**	-0.02 (-0.14)	-0.21 (-1.45)	0.02 (0.14)	-0.08 (-0.29)	0.11 (0.56)	0.18 (1.39)	0.22 (1)	0.07 (1.24)	0.3
FF48	**0.54 (2.37)**	**0.4 (2.4)**	-0.03 (-0.19)	-0.21 (-1.49)	0.01 (0.1)	-0.07 (-0.25)	0.12 (0.65)	0.19 (1.43)	0.23 (1.05)	0.09 (1.51)	0.3
Cluster5	**0.54 (2.34)**	**0.41 (2.46)**	-0.03 (-0.17)	-0.21 (-1.5)	0.01 (0.06)	-0.06 (-0.23)	0.11 (0.6)	0.19 (1.44)	0.22 (0.98)	**0.26 (1.95)**	0.3
Cluster9	**0.55 (2.38)**	**0.41 (2.43)**	-0.03 (-0.21)	-0.24 (-1.63)	0 (-0.02)	-0.06 (-0.23)	0.12 (0.61)	0.17 (1.31)	0.25 (1.13)	**0.3 (2.54)**	0.3
Cluster10	**0.54 (2.37)**	**0.41 (2.43)**	-0.03 (-0.21)	-0.23 (-1.62)	0 (0)	-0.06 (-0.23)	0.12 (0.62)	0.18 (1.36)	0.24 (1.1)	**0.23 (2.03)**	0.3
Cluster11	**0.54 (2.35)**	**0.41 (2.46)**	-0.04 (-0.22)	-0.23 (-1.6)	0 (-0.02)	-0.05 (-0.2)	0.12 (0.65)	0.18 (1.31)	0.25 (1.11)	**0.23 (2.19)**	0.3
Cluster12	**0.53 (2.32)**	**0.42 (2.48)**	-0.03 (-0.2)	-0.22 (-1.57)	0 (0.04)	-0.07 (-0.25)	0.12 (0.62)	0.19 (1.39)	0.24 (1.08)	**0.29 (2.31)**	0.3
Clusetr17	**0.53 (2.32)**	**0.41 (2.42)**	-0.04 (-0.24)	-0.22 (-1.49)	0.01 (0.06)	-0.05 (-0.19)	0.12 (0.65)	0.18 (1.35)	0.23 (1.04)	**0.28 (2.6)**	0.3
Cluster18	**0.54 (2.34)**	**0.4 (2.39)**	-0.04 (-0.22)	-0.22 (-1.51)	0.01 (0.1)	-0.05 (-0.18)	0.12 (0.62)	0.19 (1.41)	0.23 (1.02)	**0.22 (2.03)**	0.3
Cluster24	**0.54 (2.37)**	**0.4 (2.39)**	-0.05 (-0.31)	-0.23 (-1.59)	0 (0)	-0.04 (-0.14)	0.12 (0.63)	0.17 (1.28)	0.25 (1.11)	**0.31 (2.58)**	0.3
Cluster30	**0.54 (2.36)**	**0.41 (2.42)**	-0.05 (-0.29)	-0.24 (-1.63)	0 (-0.01)	-0.06 (-0.24)	0.13 (0.66)	0.18 (1.36)	0.23 (1.04)	**0.29 (2.38)**	0.3
Cluster48	**0.53 (2.32)**	**0.39 (2.31)**	-0.03 (-0.17)	-0.22 (-1.53)	0.02 (0.16)	-0.04 (-0.15)	0.12 (0.63)	0.18 (1.36)	0.23 (1.06)	**0.29 (2.56)**	0.3

TABLE B.3: **Number of Firms per Sector (K-means): Descriptive Statistics**

This table reports the distribution of the number of firms within each economic sector for different classification methods. For clusters, K-means is used instead of bisecting K-means. Only "functional" sectors with $N \geq 5$ firms are considered. The first panel groups schemes that yields 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	306	186	164	172	126	130
Std.	275	209	195	370	274	331	204	90	151
Skewness	1.48	0.73	0.81	1.66	3.1	3.76	1.80	1.01	1.90
Kurtosis	1.81	-0.74	-0.43	1.96	9.18	13.4	2.69	0.39	3.16
Min	7	28	21	5	22	5	5	5	5
1st	9	41	22	5	24	6	5	10	5
50th	248	237	196	161	103	76	93	104	69
99th	1180	784	740	1482	1383	1686	832	384	650
Max	1209	830	805	1670	1438	1832	991	404	807

TABLE B.4: **Within-sector Explained Variation (K-means)**

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. K-means is used instead of bisecting K-means. Data refer to the period July 1984 - June 2019.

K	SIC	NAICS	FF	GICS	Clustering
5	-	-	4.17%	-	8.50%
9	5.98%	-	-	-	10.09%
10	-	-	8.51%	-	10.73%
11	-	-	-	8.80%	10.70%
12	-	-	8.00%	-	10.57%
17	-	-	9.98%	-	10.60%
18	-	8.98%	-	-	10.87%
24	-	-	-	9.00%	10.55%
30	-	-	9.88%	-	10.75%
48	-	-	11.20%	-	11.84%

TABLE B.5: **Sector Investing (K-means): Out-of-sample SR**

This table shows the OOS Sharpe Ratio of the maximum SR 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 yearand the classification into clusters is kept fixed over the next 12 months. Data refer to the period July 1984 - June 2019.

K	SIC	NAICS	FF	GICS	Clustering
5	-	-	1.00	-	1.29
9	0.81	-	-	-	1.28
10	-	-	0.73	-	1.39
11	-	-	-	0.84	1.39
12	-	-	0.83	-	1.35
17	-	-	1.74	-	1.21
18	-	0.98	-	-	1.17
24	-	-	-	0.82	1.30
30	-	-	0.59	-	1.08
48	-	-	0.79	-	1.38

TABLE B.6: **Sector Investing (K-means): 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 clustering-based sectors when K-means is used instead of bisecting K-means, for 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
Cluster5	0.26	**2.58**	0.44	**2.65**	0.21	**2.46**	0.37
Cluster9	0.48	**4.33**	0.74	**4.73**	0.47	**4.95**	0.34
Cluster10	0.49	**4.11**	0.7	**5.06**	0.51	**5.09**	0.36
Cluster11	0.41	**4.44**	0.76	**4.59**	0.39	**4.88**	0.31
Cluster12	0.39	**3.85**	0.66	**4.1**	0.38	**4.47**	0.37
Cluster17	0.38	**3.7**	0.63	**4.1**	0.34	**4.02**	0.4
Cluster18	0.4	**4.55**	0.78	**4.7**	0.39	**4.8**	0.23
Cluster24	0.37	**4.24**	0.73	**4.61**	0.39	**4.93**	0.28
Cluster30	0.43	**4.82**	0.83	**5.17**	0.45	**5.83**	0.32
Cluster48	0.42	**4.63**	0.79	**5.06**	0.41	**5.43**	0.39

TABLE B.7: **Sector Investing (K-means): WMR factor**

This table shows market prices of risk (in percent) estimated through FMB regressions (Fama and MacBeth, 1973) for the five Fama and French (2015) plus momentum plus WMR, the within-cluster mean-reversion portfolio, for different industrial classification schemes and different number of industries K. Standard K-means is used instead of bisecting K-means. The corresponding t-statistic is reported in brackets, with bold numbers for values above conventional significance levels. The second-to-last column refers to the WMR factors built from the economic sectors on the rows. The last column refers to the average adjusted R^2 from the second step of the FMB regressions. WMR factors are rebalanced at the end of June of each year between 1984 and 2019.

	Mkt-RF	SMB	HML	RMW	CMA	Mom	WMR	Adj. R^2
Cluster5	**0.54 (2.33)**	**0.42 (2.52)**	-0.07 (-0.4)	-0.23 (-1.64)	-0.02 (-0.15)	-0.13 (-0.49)	0.16 (1.2)	0.25
Cluster9	**0.52 (2.28)**	**0.41 (2.44)**	-0.07 (-0.42)	**-0.24 (-1.65)**	-0.02 (-0.14)	-0.12 (-0.45)	**0.33 (2.49)**	0.25
Cluster10	**0.54 (2.33)**	**0.4 (2.43)**	-0.07 (-0.4)	**-0.24 (-1.69)**	-0.01 (-0.08)	-0.1 (-0.39)	**0.34 (2.42)**	0.25
Cluster11	**0.53 (2.3)**	**0.42 (2.5)**	-0.07 (-0.39)	**-0.24 (-1.67)**	-0.02 (-0.16)	-0.12 (-0.45)	**0.27 (2.48)**	0.25
Cluster12	**0.53 (2.3)**	**0.42 (2.53)**	-0.06 (-0.38)	**-0.25 (-1.72)**	-0.02 (-0.16)	-0.12 (-0.43)	**0.29 (2.45)**	0.25
Cluster17	**0.52 (2.27)**	**0.42 (2.48)**	-0.06 (-0.38)	**-0.24 (-1.69)**	-0.01 (-0.07)	-0.14 (-0.52)	**0.33 (2.62)**	0.25
Cluster18	**0.53 (2.29)**	**0.41 (2.47)**	-0.05 (-0.33)	**-0.24 (-1.68)**	-0.01 (-0.09)	-0.14 (-0.52)	**0.25 (2.47)**	0.25
Cluster24	**0.52 (2.25)**	**0.42 (2.5)**	-0.05 (-0.33)	-0.22 (-1.52)	-0.01 (-0.11)	-0.15 (-0.56)	**0.2 (1.91)**	0.25
Cluster30	**0.53 (2.32)**	**0.41 (2.44)**	-0.07 (-0.41)	**-0.24 (-1.66)**	-0.02 (-0.15)	-0.14 (-0.53)	**0.31 (2.89)**	0.25
Cluster48	**0.52 (2.26)**	**0.42 (2.54)**	-0.05 (-0.3)	-0.23 (-1.59)	0 (-0.01)	-0.13 (-0.48)	**0.23 (2.15)**	0.25

TABLE B.8: **Sector Investing (K-means): WMR factor with additional controls**

This table shows market prices of risk (in percent) estimated through FMB regressions (Fama and MacBeth, 1973) for the five Fama and French (2015) model plus momentum plus short-term and long-term reversals plus industry momentum (Moskowitz and Grinblatt, 1999) plus WMR, the within-cluster mean-reversion portfolio, for different industrial classification schemes and different number of industries K. Standard K-means is used instead of bisecting K-means. The corresponding t-statistic is reported in brackets, with bold numbers for values above conventional significance levels. The second-to-last column refers to the WMR factors built from the economic sectors on the rows. The last column refers to the average adjusted R^2 from the second step of the FMB regressions. WMR factors are rebalanced at the end of June of each year between 1984 and 2019.

	Mkt-RF	SMB	HML	RMW	CMA	Mom	STR	LTR	INDMOM	WMR	Adj. R^2
Cluster5	**0.55 (2.41)**	**0.39 (2.37)**	-0.04 (-0.26)	-0.22 (-1.53)	0 (-0.03)	-0.06 (-0.21)	0.11 (0.59)	0.18 (1.36)	0.23 (1.06)	0.14 (1.11)	0.3
Cluster9	**0.54 (2.36)**	**0.4 (2.37)**	-0.04 (-0.23)	-0.22 (-1.55)	0 (0)	-0.04 (-0.17)	0.13 (0.65)	0.17 (1.3)	0.25 (1.14)	**0.3 (2.38)**	0.3
Cluster10	**0.56 (2.45)**	**0.38 (2.3)**	-0.04 (-0.25)	-0.22 (-1.56)	0 (0.02)	-0.03 (-0.12)	0.12 (0.61)	0.17 (1.3)	0.25 (1.14)	**0.3 (2.17)**	0.3
Cluster11	**0.55 (2.41)**	**0.39 (2.35)**	-0.04 (-0.24)	-0.22 (-1.55)	0 (-0.04)	-0.05 (-0.17)	0.11 (0.59)	0.17 (1.29)	0.25 (1.12)	**0.23 (2.2)**	0.3
Cluster12	**0.55 (2.39)**	**0.4 (2.4)**	-0.04 (-0.21)	-0.23 (-1.59)	0 (-0.04)	-0.04 (-0.16)	0.12 (0.61)	0.18 (1.31)	0.25 (1.12)	**0.28 (2.35)**	0.3
Cluster17	**0.55 (2.39)**	**0.39 (2.33)**	-0.03 (-0.21)	-0.23 (-1.56)	0 (0.04)	-0.07 (-0.26)	0.12 (0.62)	0.17 (1.28)	0.23 (1.03)	**0.31 (2.56)**	0.3
Cluster18	**0.55 (2.4)**	**0.38 (2.28)**	-0.03 (-0.19)	-0.22 (-1.51)	0 (0.03)	-0.06 (-0.24)	0.12 (0.62)	0.17 (1.27)	0.24 (1.08)	**0.25 (2.43)**	0.3
Cluster24	**0.53 (2.35)**	**0.41 (2.42)**	-0.03 (-0.2)	-0.21 (-1.46)	0.01 (0.06)	-0.07 (-0.25)	0.11 (0.59)	0.18 (1.37)	0.24 (1.07)	**0.19 (1.93)**	0.3
Cluster30	**0.54 (2.38)**	**0.4 (2.41)**	-0.05 (-0.3)	-0.24 (-1.64)	0 (-0.01)	-0.06 (-0.21)	0.12 (0.63)	0.18 (1.39)	0.24 (1.09)	**0.29 (2.72)**	0.3
Cluster48	**0.54 (2.35)**	**0.4 (2.42)**	-0.04 (-0.21)	-0.21 (-1.49)	0.01 (0.08)	-0.05 (-0.18)	0.12 (0.61)	0.18 (1.32)	0.25 (1.11)	**0.22 (2.04)**	0.3

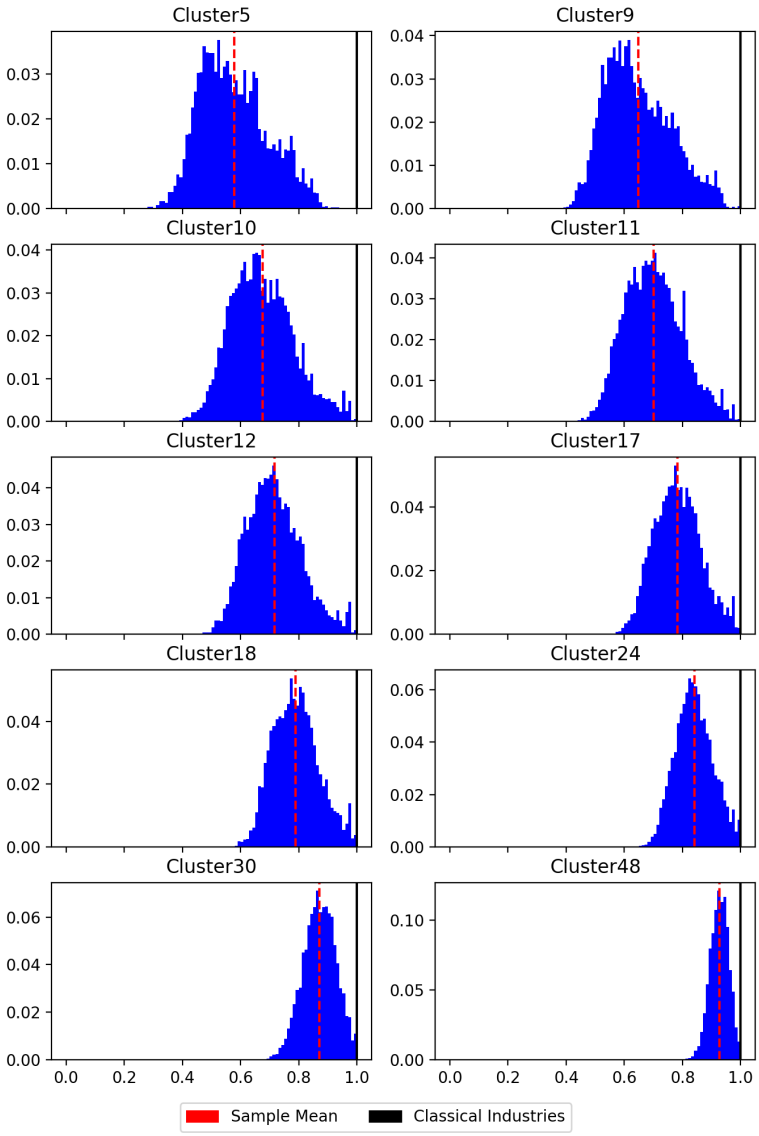

FIGURE B.1: **Cluster Time-stability and Stability Index (K-means)**
This figure shows the empirical density of the firm-level stability measure $\mathcal{S}^i$ for $i = 1, ..., N$ for different number of clusters K. Standard K-means is used instead of bisecting K-means. The red vertical bar denotes the cross-sectional mean, i.e. the stability index *SI*. The black bar refers to any other classification systems (the density collapses onto the value 1). Data refer to the period July 1984 - June 2019.

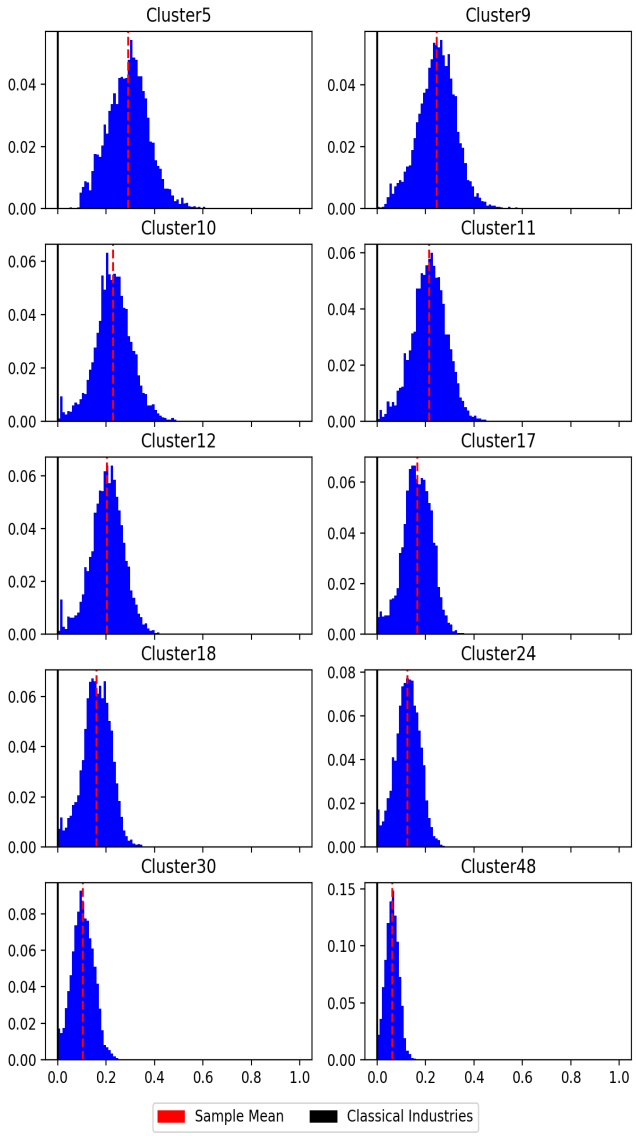

FIGURE B.2: **Cluster Time-Instability and Instability Index (K-means)**
This figure shows the empirical density of the firm-level instability measure $\mathcal{G}^i$ for $i = 1, ..., N$ for different number of clusters K. Standard K-means is used instead of bisecting K-means. The red vertical bar denotes the cross-sectional mean, i.e. the instability index *II*. The black bar refers to any other classification systems (the density collapses onto the value 0). Data refer to the period July 1984 - June 2019.

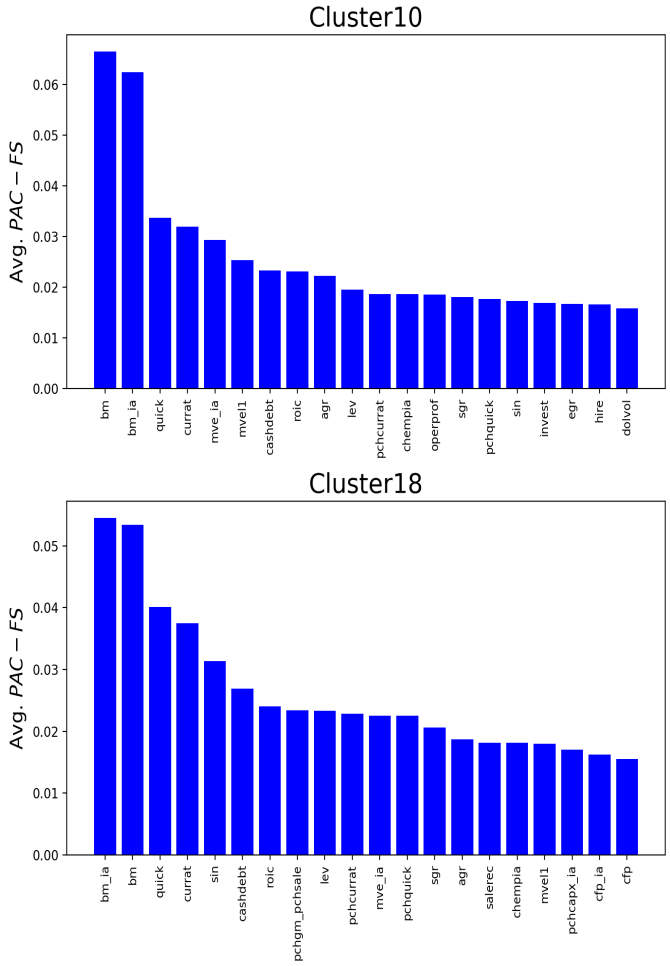

FIGURE B.3: **Feature Importance,** $K = 10$ **and** $K = 18$
This figure shows the time-series average $PAC\text{-}FS_p$ for the twenty characteristics with the highest values, in descending order, for different K corresponding to Cluster10 and Cluster 18, as report in the titles above each panel. Standard K-means is used instead of bisecting K-means. Data refer to the period July 1984 - June 2019.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 396	Nils Grevenbrock, Alexander Ludwig, Nawid Siassi	Homeownership Rates, Housing Policies, and Co-Residence Decisions
No. 395	Ruggero Jappelli, Loriana Pelizzon, Marti Subrahmanyam	Quantitative Easing, the Repo Market, and the Term Structure of Interest Rates
No. 394	Kevin Bauer, Oliver Hinz, Moritz von Zahn	Please Take Over: XAI, Delegation of Authority, and Domain Knowledge
No. 393	Michael Kosfeld, Zahra Sharafi	The Preference Survey Module: Evidence on Social Preferences from Tehran
No. 392	Christian Mücke	Bank Dividend Restrictions and Banks' Institutional Investors
No. 391	Carmelo Latino, Loriana Pelizzon, Max Riedel	How to Green the European Auto ABS Market? A Literature Survey
No. 390	Kamelia Kosekova, Angela Maddaloni, Melina Papoutsi, Fabiano Schivardi	Firm-Bank Relationships: A Cross-Country Comparison
No. 389	Stefan Goldbach, Philipp Harms, Axel Jochem, Volker Nitsch, Alfons J. Weichenrieder	Retained Earnings and Foreign Portfolio Ownership: Implications for the Current Account Debate
No. 388	Gill Segal, Ivan Shaliastovich	Uncertainty, Risk, and Capital Growth
No. 387	Michele Costola, Katia Vozian	Pricing Climate Transition Risk: Evidence from European Corporate CDS
No. 386	Alperen Afşin Gözlügöl, Wolf-Georg Ringe	Net-Zero Transition and Divestments of Carbon-Intensive Assets
No. 385	Angela Maddaloni	Liquidity Support and Distress Resilience in Bank-Affiliated Mutual Funds
No. 384	Julian Greth, Alperen Afşin Gözlügöl, Tobias Tröger	The Oscillating Domains of Public and Private Markets
No. 383	Satyajit Dutt, Jan Wedigo Radermacher	Age, Wealth, and the MPC in Europe - A Supervised Machine Learning Approach
No. 382	Jan Wedigo Radermacher	Mamma Mia! Revealing Hidden Heterogeneity by PCA-Biplot - MPC Puzzle for Italy's Elderly Poor