

# Neighbouring Assets\*

Sina Seyfi<sup>†</sup>

July, 2023

## ABSTRACT

Firms with similar characteristics display similar expected returns. Defining neighbouring assets as those with the most similar set of characteristics, I show that past returns of an asset's neighbours predict its future expected returns. If a majority of an asset's neighbours have performed poorly (well) in the past, it is likely that this asset also performs poorly (well) in the future. By classifying each asset into a decile portfolio based on the past performance of its neighbours, with 94 characteristics, a long-short portfolio generates an out-of-sample annualized Sharpe ratio of 1.15 with a monthly alpha of 2.72% ( $t = 8.86$ ).

**Keywords:** Asset pricing, portfolios, cross-section of expected returns, stock characteristics, machine learning

---

\*I am thankful to Peter Nyberg, Matthijs Lof, Matti Keloharju, Petri Jylhä, Cristian Tiu (discussant), Theis Ingerslev Jensen, Anders Löflund (discussant), Mikko Leppämäki, Renato Lazo Paz (discussant), Markku Kaustia, Aleksi Pitkäljärvi, Ville Rantala, participants at the FMA European conference, Nordic Finance Network (NFN) workshop, Finance Brown Bag (Aalto University), and GSF Winter Workshop for their useful comments. I thank OP Group Research Foundation for the financial support. First version: December 2022

<sup>†</sup>PhD student of Finance, Aalto University, School of Business, Department of Finance  
Email: sina.seyfi@aalto.fi

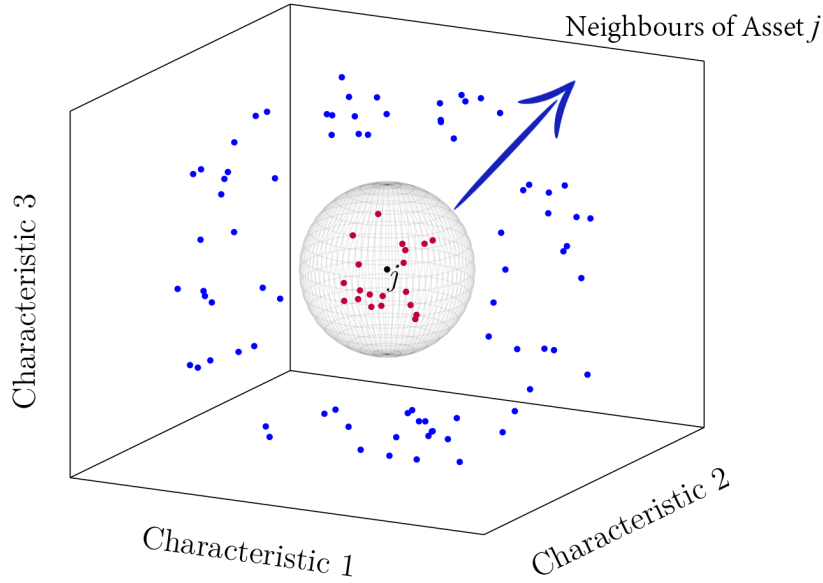
If firm-level characteristics are determinants of expected stock returns, then firms with similar characteristics should have similar expected rates of returns. This suggests that past returns of firms which share similar characteristics should be similar to the future returns of a firm which is alike in characteristics. The empirical asset pricing literature, however, has discovered quite a large number of firm characteristics which relate to the expected stock returns.<sup>1</sup> With this zoo of characteristics, it is unclear how the similarity of firms should be measured and according to which characteristics assets should be sorted into portfolios (Cochrane (2011)). In this paper, I tackle this large-dimensionality challenge and link similar firms through many characteristics.

For recognizing analogous firms, I start by finding *neighbouring assets*, identified as those assets with the most similar set of characteristics. I measure this similarity between two assets as the distance of their characteristics and define an asset's *neighbours* as those whose characteristics have the closest distance to this very asset. For instance, without loss of generality suppose that there exist only three characteristics; In Figure 1, I show the closest assets to asset  $j$  with red dots and define them as neighbouring assets for asset  $j$ . More distant assets (shown with blue dots) are not considered neighbours of asset  $j$ . By defining neighbouring assets in this way, each asset has almost the same characteristics as its neighbours, and neighbouring assets display fundamental linkages in many dimensions such as the similarity of accounting variables or financial ratios. If the cross-sectional variation in expected stock returns roots in the cross-sectional variation in asset-specific characteristics, naturally neighbouring assets should have similar expected returns.

Recognizing neighbouring assets has several attractive applications in asset pricing. To begin with, grouping (separating) assets with similar (different) characteristics must generate cross-sectional dispersion in expected returns. This assumption provides a new method for cross-sectional return predictability through a large set of predictors in a flexible setting. In order to produce cross-sectional dispersion in average returns through portfolios of neighbouring assets, I classify each asset into one of the decile portfolios based on the past performance of its neighbours, wherein decile 1 (10) is indicative of the loser (winner) portfolio. For example, if a majority of an asset's neighbours in the past, which have had the same characteristics as the most recent characteristics of the asset in question now, have belonged to decile portfolio 1, the loser portfolio, it is most likely that this asset itself also belongs to the loser portfolio, decile 1. Intuitively, if assets with a specific set of characteristics in the past have performed badly (well), it is likely that the assets in the future with the same set of characteristics will perform badly (well), as well. Considering the case with only two characteristics as an example, if assets with high momentum and low size have produced a high expected return in the in-sample data, all the assets with high momentum and low size in out-of-sample data are then grouped into a decile portfolio which is supposed to generate a high expected return. This portfolio reflects the behaviour of high momentum and low size stocks over time. Similarly, each decile portfolio is representative of a potentially large set of characteristics, generating the corresponding expected returns.

---

<sup>1</sup>See, among many others, Green, Hand, and Soliman (2011), Harvey, Liu, and Zhu (2016), Hou, Xue, and Zhang (2017), Green, Hand, and Zhang (2017), Li and Rossi (2020), and Chen and Zimmermann (2021).



**Figure 1.** Definition of neighbouring assets in the characteristics space

Without loss of generality, if there exist only three characteristics, the  $k$  closest assets to the asset  $j$  (shown as red dots) are defined as the neighbours of asset  $j$ . If expected returns are a function of firm characteristics, regardless of the functional form, neighbouring assets display similar expected returns.

More formally, suppose that an asset  $j$  at time  $t$  has a set of most recent characteristics  $\mathbf{x}_{j,t-1}$  and an expected return  $E(r_{j,t})$ . All assets at time  $t-t^0$ , the in-sample period, with a set of characteristics  $\mathbf{x}_{t-t^0-1}$ , are assigned into one of the decile portfolios based on their mean returns at time  $t-t^0$ ,  $E(r_{t-t^0})$ , which is observable to the investor at time  $t$ . Hence, each asset in the in-sample data is assigned to a decile portfolio. For classifying an asset  $j$  at time  $t$ , I first find the  $k$  nearest neighbours of this asset in the in-sample data, i.e. those  $k$  assets with the most similar set of characteristics in the in-sample data, and put asset  $j$  into the decile portfolio in which the majority of its  $k$  neighbours belonged to in the past. Following [Ali and Hirshleifer \(2020\)](#), I also assume that *closer* neighbours, having a smaller distance of characteristics, must contribute more in decile prediction, so I weigh all of the  $k$  nearest neighbouring assets by the inverse of their distance to the asset  $j$ .<sup>2</sup> The in-sample period is considered a rolling window consisting of  $t-120$  months before time  $t$ , which allows for the time-varying relationship between characteristics and expected returns. Figure 2 panel (a) visualizes this concept.

I find that this strategy generates a fairly large dispersion in the cross-section of stock returns. Recognizing neighbouring assets according to 94 firm characteristics,<sup>3</sup> a value-

<sup>2</sup>[Ali and Hirshleifer \(2020\)](#) define the connection between two firms as the number of shared analyst coverage. For the return predictability, they assume that stocks which have more co-covered analysts are more likely to be similar, and hence they weigh stocks by the number of co-covered analysts (see equation 1 in their paper). In my analysis, although I follow [Ali and Hirshleifer \(2020\)](#) by weighting neighbouring assets with the inverse of their distance, the results are robust if I weigh all neighbouring assets equally.

<sup>3</sup>These characteristics are listed in Table A1. I thank Dacheng Xiu for providing the characteristics data on

and equally-weighted long-short strategy of extreme decile portfolios generate a monthly Fama-French three (FF3) alpha of 1.77% ( $t = 8.89$ ) and 1.77% ( $t = 12.30$ ), respectively, even after excluding the smallest 5% of stocks.<sup>4</sup> The corresponding annualized Sharpe ratios in the period 1980-2021 inclusively are 1.34 and 1.87 for value- and equally-weighted portfolios. Even after excluding micro caps (those which are below 20% NYSE percentile), the FF3 monthly alphas are 1.02% ( $t = 6.68$ ) and 1.15% ( $t = 9.43$ ), with Sharpe ratios of 0.95 and 1.26. These results suggest that the performance is not derived from the microcap stocks.

If past returns of an asset's neighbours at an individual level have predictive powers, then portfolios consisting of each asset's neighbours should also predict future returns of this asset. To test this, for each asset  $j$  at time  $t$ , I create a portfolio at each time  $t - 1$ ,  $t - 2$ , ...,  $t - t$  ( $t = 120$  months) which contains the nearest neighbours of asset  $j$  at each month in the past. Then I predict the future return of asset  $j$  based on the average returns of its neighbouring portfolios. Figure 2 panel (b) provides a visualisation. Forming decile portfolios based on this method, I find that an out-of-sample value-weighted long-short portfolio generates a monthly FF3 alpha of 1.78% ( $t = 6.56$ ) with an annualized Sharpe ratio of 0.95. The monthly FF3 alpha for an equally-weighted counterpart long-short portfolio increases to 2.87% ( $t = 13.98$ ) with an annualized Sharpe ratio of 2.12.

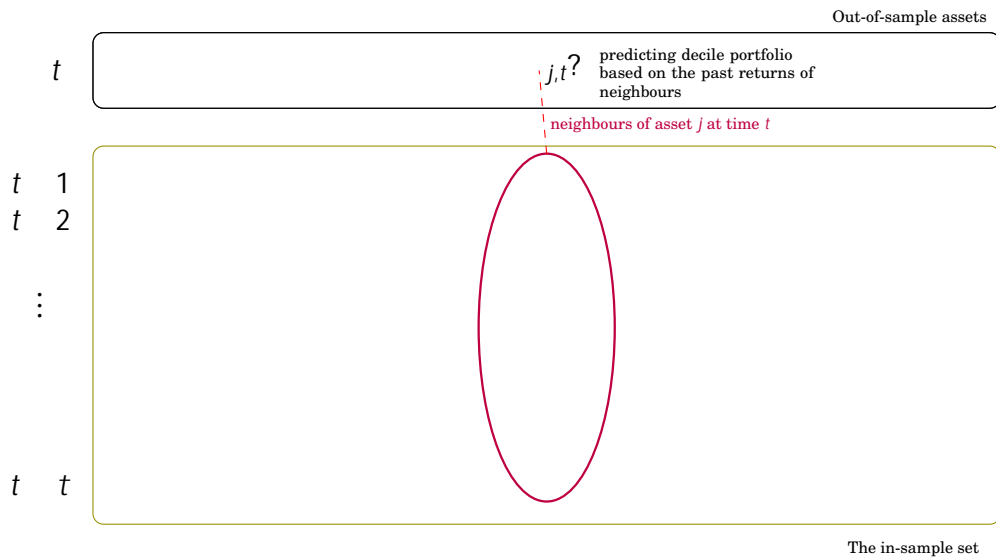
What mechanism drives the performance of neighbouring assets strategy? If exposures to systematic risks are a function of firm characteristics (such as in Kelly, Pruitt, and Su (2019) or Gu, Kelly, and Xiu (2021)), then stocks with similar characteristics have similar exposure to the systematic risk factors. In turn, if the cross-sectional variation in expected stock returns is caused by exposure to systematic risk factors, then neighbouring assets, because of similar exposures, have similar expected returns. This holds regardless of what the true sources of risks are and what the functional form between characteristics and covariances is: as long as the compensation for exposure to risk factors—observed or latent—drives the relationship between characteristics and expected returns, neighbouring assets display similar expected returns and a long-short portfolio of neighbouring stocks strategy is profitable.

Grouping adjacent assets into decile portfolios according to their past neighbours has several economic interpretations. One interpretation is that an asset's neighbours' past returns predict its future returns. From this point of view, my paper contributes to the literature on finding peer firms by introducing a new way of defining related firms based on the similarity of characteristics. The literature has defined various types of connections between firms, including same industry linkage (Moskowitz and Grinblatt (1999)), same principal customers (Cohen and Frazzini (2008)) and suppliers (Menzly and Ozbas (2010)), same technology (Lee, Ma, and Wang (2016)), common active mutual fund owners (Anton and Polk (2014)), and same analyst coverage (Ali and Hirshleifer (2020)), among others, all yielding to cross-sectional predictability of returns. In this regard, I introduce a new

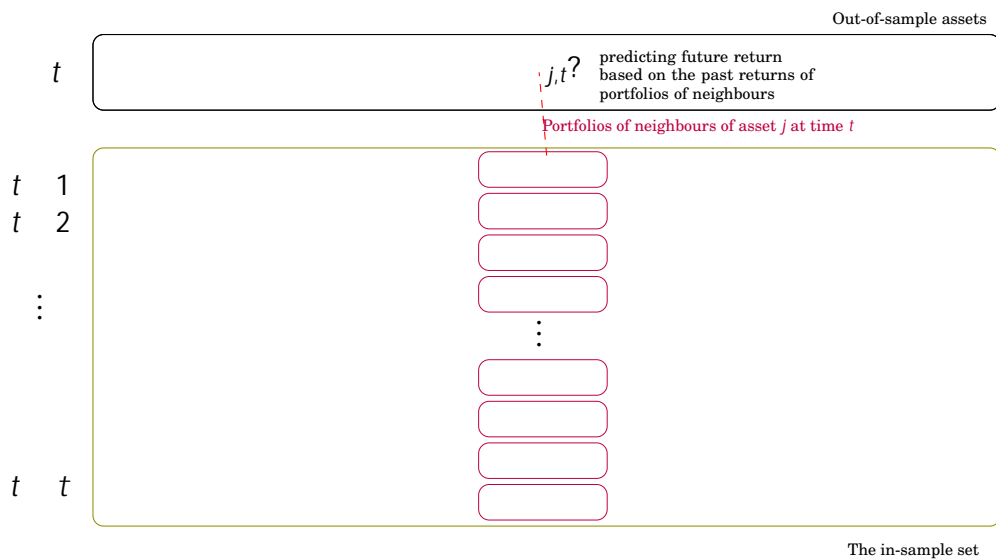
---

his website.

<sup>4</sup>The smallest 5% stocks are usually very illiquid and have many missing characteristics. I exclude them from my analysis to make sure they do not distort the results, although they do not affect the results as long as the portfolios are value-weighted.



(a) Individual neighbours



(b) Portfolios of neighbouring assets

**Figure 2.** Neighbouring assets classification algorithm

Panel (a) shows the classification framework based on the neighbouring assets. For an asset  $j$  at time  $t$ , I find its  $k$  neighbours in the in-sample data, from time  $t - t$  to  $t - 1$ , which have had the closest distance of characteristics to asset  $j$  at time  $t$ . If the majority of neighbours have performed poorly (well), it is likely that asset  $j$  at time  $t$  also performs poorly (well). Hence, I put asset  $j$  at time  $t$  into the decile portfolio to which the majority of its neighbours belonged in the past. The in-sample set is considered a rolling window. In panel (b), for each asset  $j$  at time  $t$ , I create a portfolio of neighbouring assets in each month in the in-sample data. Blue dots show the assets which have had the most similar characteristics to asset  $j$  at time  $t$ . Then I predict the return of asset  $j$  at time  $t$  based on the average of in-sample neighbouring portfolios.

linkage between firms, that is, *closeness of many characteristics* at the same time.

Indeed, the economic link between firms with similar characteristics based on one or two characteristics has been studied in the literature. For instance, [Fama and French \(1995\)](#) document that earnings of firms with similar size and book-to-market are explained

by common factors. Or, [Hirshleifer, Hou, and Teoh \(2012\)](#) argue that firms with similar accrual co-move. Considering seventeen characteristics one-by-one, [Müller \(2019\)](#) shows that firms which are in the same quantile based on one characteristic have information spillovers and can be considered economically linked firms. [He, Wang, and Yu \(2021\)](#) argue that, because of investor attention, stocks with similar price, size, book-to-market, return on assets, and investment-to-assets have spillover effects. I generalize all these approaches by linking the firms based on many characteristics. If two firms are neighbours according to my definition, they share similar characteristics, and their values are sensitive to the same type of information. Also, other types of connections such as being in the same industry or having the same supplier can lead firms to share similar characteristics, such as same amount of sales, growth, industry momentum, etc. As an extreme case, suppose that two firms are identical in all aspects: same products, same customers, same industry, etc; In this case, there is no reason to expect that these two firms would have different expected rates of returns. In fact, the similarity of characteristics is indicative of fundamental linkages. Therefore, I show that the similarity and dissimilarity of characteristics account for the cross-predictability of returns, and hence, using a neighbourhood definition based on the characteristics is an effective way of recognizing related firms.

The neighbourhood connection, importantly, becomes stronger when I use more characteristics for finding neighbouring assets. For instance, when only three characteristics, namely, size, book-to-market, and momentum are used to find neighbouring assets, an out-of-sample value-weighted long-short portfolio generates a monthly average returns of 1.64% ( $t = 4.49$ , and Sharpe Ratio  $SR = 0.69$ ) between 1980-2021. When adding 9 more characteristics to the previous ones,<sup>5</sup> a long-short portfolio generates 1.96% ( $t = 5.57$ ,  $SR = 0.86$ ) monthly average returns. Finally, by considering 94 characteristics, the monthly average performance of this long short-portfolio increases to 2.36% ( $t = 7.42$ ,  $SR = 1.15$ ), apparently suggesting that the predictive power of neighbouring assets increases when the neighbourhood definition is expanded to many dimensions. This finding indicates that a large set of characteristics *jointly* predict future expected returns and the factor structure of expected returns is remarkably high-dimensional, which is also in line with the emerging literature on the high-dimensionality of the factor structure of expected returns. For instance, [Kozak, Nagel, and Santosh \(2020\)](#) and [Jensen, Kelly, and Pedersen \(2021\)](#) show that there is little redundancy among a large set of documented anomalies, while [Bryzgalova, Huang, and Julliard \(2022\)](#) find that a large set of characteristics are needed to explain the variations in the cross-section of stock returns.

Return predictability through past returns, such as using a momentum strategy, has been well documented in the literature to be robust and pervasive ([Jegadeesh and Titman \(2001\)](#), [Goyal and Wahal \(2015\)](#)). However, a momentum strategy uses each asset's own past history to predict its future returns. In contrast, and unlike [Kelly, Malamud, and](#)

---

<sup>5</sup>These 9 characteristics are documented to be among the most important features in the literature. They include 1-month momentum ([Jegadeesh and Titman \(1993\)](#)), change in momentum ([Gettleman and Marks \(2006\)](#)), maximum daily return ([Bali, Cakici, and Whitelaw \(2011\)](#)), industry momentum ([Moskowitz and Grinblatt \(1999\)](#)), return volatility ([Ang, Hodrick, Xing, and Zhang \(2006\)](#)), dollar trading volume ([Chordia, Subrahmanyam, and Anshuman \(2001\)](#)), sales to price ([Barbee Jr, Mukherji, and Raines \(1996\)](#)), share turnover ([Datar, Naik, and Radcliffe \(1998\)](#)), and asset growth ([Datar et al. \(1998\)](#)).

Pedersen (2020) who use all asset's signals to predict each individual returns, I use only each asset's neighbours past performance for return predictability of individual securities. It is important to check, nevertheless, if neighbouring assets' strategy is not just a proxy for momentum. While a momentum factor which goes long (short) in the winner (loser) stocks generates a monthly average of 0.54% ( $t = 2.69$ ) in the period 1980-2021, it obtains a negative alpha of -0.20% ( $t = -1.18$ ) when regressed on a long-short portfolio from neighbouring assets. On the other hand, this long-short portfolio of neighbouring assets still generates an alpha of 0.92% ( $t = 5.70$ ) with respect to the four-factor Carhart model. This clearly indicates that my long-short portfolio spans the momentum factor.

Inspired by Martin and Nagel (2022) who discriminate between in-sample and out-of-sample predictability, the focus of my paper is fully on out-of-sample prediction. My decile portfolios generate a wide range of expected returns, namely from -0.91% ( $t = -2.01$ ) in the lowest decile to 1.44% ( $t = 3.55$ ) in the highest decile in excess of the risk-free rate. This spread persists over time and does not disappear after excluding tiny stocks or when considering only large stocks. Most importantly, the profitability comes from both the long leg and the short side. I track the pattern of characteristics in my decile portfolios in over 40 years of out-of-sample study (from 1980 to 2021), and find that most of these patterns are either monotonically linear or a U-shape or an inverse of a U-shape or an M-shape. This finding suggests that the characteristics-returns relationship could be up to the order of 4, whereas most of the literature has focused on the linear part. These patterns seem to be invariant over time and they are stronger when the universe of data includes tiny stocks.

**Closely Related Literature.** My paper contributes to the literature that tries to model the future risk premium as a function of lagged stock-level characteristics, such as Fama and French (2008). Traditional approaches for mapping characteristics to the expected returns include employing cross-sectional regressions used by Haugen and Baker (1996), Lewellen (2015) and Light, Maslov, and Rytchkov (2017) or portfolio sorts such as Daniel, Grinblatt, Titman, and Wermers (1997). Newer methods include machine learning methods surveyed by Gu, Kelly, and Xiu (2020). Gu et al. (2021), Feng, Polson, and Xu (2018b) and Feng, He, and Polson (2018a) use deep neural networks to map characteristics to the risk premiums, while Chen, Pelger, and Zhu (2019) estimate Stochastic Discount Factor with generative adversarial networks. In Seyfi (2023), I develop a new method to find the most eminent characteristics. Li and Rossi (2020) employ boosted regression trees for predicting mutual funds returns based on their characteristics. My method deviates from the literature, however, by (1) the way the classification is defined, that is, forming decile portfolio sorts and (2) the tool that I pick for solving this classification problem, which is analogous to *k-Nearest Neighbours*<sup>6</sup> (*k*NN) borrowed from machine learning and has clear economic interpretations. Whereas machine learning methods often suffer from high computational costs, over-fitting, and lack of transparency, my proposed approach is highly transparent, simple, and non-parametric. There is no risk of over-fitting in my approach because *k*NN does not need any training procedure.

My paper also alleviates the curse-of-dimensionality problem in the cross-sections of

---

<sup>6</sup>Cover and Hart (1967)

returns. For example, Giglio, Liao, and Xiu (2021) argue that the curse of dimensionality brings statistical inference problems. Of the most famous methods for handling the curse of dimensionality is to apply dimension reduction approaches. Among others, Kozak, Nagel, and Santosh (2018), Kozak et al. (2020), Lettau and Pelger (2020) use PCA-based methods and Rapach, Strauss, and Zhou (2013), Feng, Giglio, and Xiu (2020) and Freyberger, Neuhierl, and Weber (2020) opt LASSO-type approaches in order to pick the most important variables. While these methods are based on handpicking potentially the most useful features, I deviate from the literature by unifying the joint effect of characteristics zoo in the neighbouring assets, a novel method for working with many characteristics at once.

My paper also contributes to the literature of building basis assets, that is, how test portfolios should be created. Ahn, Conrad, and Dittmar (2009) use return correlations to sort assets into portfolios, while Moritz and Zimmermann (2016) and Bryzgalova, Pelger, and Zhu (2020) use Trees to group similar assets into portfolios. As stated by Nagel and Singleton (2011), managed portfolios are the optimal tools for hypothesis testing in conditional asset pricing. Inspired by Lewellen, Nagel, and Shanken (2010) who suggest that using double-sort portfolios provides a low hurdle for asset pricing models, I use 94 characteristics to form portfolios of neighbouring assets which is a new way of creating test assets based on many characteristics. The portfolios I create span the investment opportunity set, and can be used for testing asset pricing models.

Lastly, my paper proposes a new way of defining connected firms through neighbouring assets, complementary to the literature trying to find linked firms, such as Barberis and Shleifer (2003), Bouwman (2011), Shue (2013), Leary and Roberts (2014), and Kaustia and Rantala (2015), among others.

## I. Neighbouring Assets and Expected Returns

### A. Motivation

The primitive intuition behind the idea of classifying assets based on their characteristics lies in one key assumption. That is, the future risk premium of assets is a time-varying function of their lagged characteristics. My cross-sectional classification of assets into decile portfolios maps a large set of firm characteristics to the expected returns in a non-linear setting. To highlight the importance of this application, consider a case where hundreds of characteristics are available and non-linearities and interactions between these return predictors are important. If non-linearities and interactions are naively added to the set of return predictive signals, the number of predictors easily exceeds the number of observations in each cross-section (Kozak (2020)).<sup>7</sup> This large-dimensionality of predictors exacerbated by complex functional forms makes standard approaches considered in Fama and French (2008) practically infeasible: neither does it allow investors to create charac-

---

<sup>7</sup>If there are 100 firm characteristics available, for example, and one aims to consider all of the interactions and non-linearities only up to the second order for predicting expected returns (such as size, momentum, size,  $b^2$ , ...), then the number of predictors exceeds 5000, which is even more than the number of stocks in each cross-section. In this case, even running cross-sectional regressions is no longer feasible.

teristic sorted portfolios nor to run cross-sectional regressions for deriving the joint, and possibly complex, relationship between characteristics and expected returns (Green, Hand, and Zhang (2014)). For this reason, Karolyi and Van Nieuwerburgh (2020) argue that "new methods for the cross-section of returns" are needed to deal with the large-dimensionality challenge of characteristics. In this paper, I show that through neighbouring assets, investors can map a large set of stock return predictors to the expected returns at once. Or equivalently, one can derive the relationship between predictors, jointly, and expected returns.

Indeed, the time-varying functional form of the characteristics-risk premium relationship can be complex. To capture this complexity, researchers recently started using state-of-the-art machine learning tools pioneered by Gu et al. (2020). I model this functional form based on only the training sample: I do not consider any closed functional form for the relationship between characteristics and future returns and let past data derive this relationship instead. This data-driven functional form addresses the aforementioned concerns regarding if non-linearities and interactions matter, and if so, how they should enter the model. I assume that every form of relationship a set of characteristics have had with expected returns in the past will hold for the future, allowing for this relationship to vary with time in the long term.

Modelling future risk premiums as a function of characteristics is a prediction problem. I convert this prediction problem to a classification one in a novel way. I argue that the problem of explaining/predicting variations in the expected returns through their characteristics is equivalent to creating decile portfolios in which their expected (mean) returns line up monotonically. Predicting a decile portfolio which is a function of rank of expected returns is a multiclass classification which also reduces the noise. For the out-of-sample study, I predict the decile portfolio that each asset belongs to, instead of predicting the returns directly. These decile portfolios should generate dispersion in the long run.

My approach for cross-sectional returns prediction, therefore, can be seen as a supervised classification problem wherein I predict a decile portfolio (class) for each asset based on its characteristics (features). One can analogize my methodology to the  $k$ -Nearest Neighbours algorithm in machine learning. Basically,  $k$ NN is a simple classifier which classifies each observation to a class based on the majority of its neighbours' labels in the training sample. In my approach, the training sample is the in-sample rolling window of panel data, labels are decile portfolios and characteristics serve as features. From this perspective, this paper is the first to suggest a pure machine learning supervised classification approach for predicting the cross-sectional variation in stock returns. Machine learning papers in asset pricing, so far, have focused on the regression, and not classification, wing to tackle prediction problems (Kelly and Xiu (2021)). Classification methods, thus far, have been considered useful mostly for binary types of problems like a corporate default<sup>8</sup> and consequently, asset pricing researchers neglected them in the cross-sections and inclined to regression methods (Nagel (2021)). By defining features as lagged-characteristics and

---

<sup>8</sup>As an example, see Lessmann, Baesens, Seow, and Thomas (2015).

target variables as decile portfolios based on the future expected returns, in this paper, I introduce a new way of employing machine learning classifiers for cross-sections of returns.

Rather than clear economic interpretations that  $k$ NN has in this context, I briefly motivate employing this classifier from a statistical point of view. The desired tool for classifying assets into portfolios must have several features. First and foremost, it must be transparent and the logic behind that should be interpretable. There are strong machine learning classifiers, among which deep neural networks, but if the mapping function is not interpretable, it does not teach us any lesson about the economy behind the result. Sometimes, the situation is even worse; that is, the mapping function cannot even be seen by the investor, which is referred to as black-box machines. Second, everything else equal, we are interested in the simpler methods, as the complexity comes with a cost. Third, the classifier must be able to handle multi-class data, as the behaviour of expected returns cannot be summarized in only two portfolios or any other binary set. Forth, because we know the asset returns are affected by both themselves and other assets' past returns and characteristics, the classifier must be able to combine time-series and cross-sectional information at the same time. More importantly, the classifier must be able to handle large datasets, either in the sample size or in the dimension. Fifth, the classifier must be able to work with different types of distributions. I do not impose any prior assumption on the distribution of characteristics. We are not necessarily interested in linear classifiers, as we want to study the relationship between characteristics and expected returns from a broader perspective. Hence, functional form flexibility also matters. Sixth, the classifier must be supervised, because each asset has a target variable.

Among machine learning classifiers,  $k$ NN congregates all the above-mentioned features at once. Specifically,  $k$ NN is reputed for being quite intuitive and simple. It is a non-parametric classification technique which makes it very flexible when the relationship between characteristics and expected returns is complex. It also handles unbalanced panel data where the number of assets varies in each cross-section. The return predictability for each asset comes from its neighbours that share similar characteristics in many dimensions. This transparent intuition is a generalization of portfolio sorts in the literature where grouping similar assets together in only two or three dimensions. Also defining neighbouring assets in a novel way unifies the joint effect of characteristics. There is one key message in using this classifier: past returns of an asset's neighbours predict its future expected returns.

Another reason why the neighbouring assets algorithm should generate cross-sectional dispersion is that there exists a strong *clustering effect* between neighbouring assets, which allows me to group these assets into portfolios sorted based on the expected returns. Basically, the clustering effect occurs when observations in the same cluster or group have a similar behaviour. In the asset pricing context, it can be interpreted as, all assets in each portfolio/cluster (which have the same features) have similar expected returns (behaviour). So, the same behaviour of each cluster/portfolio is similarity of expected returns. By forming portfolios of assets with similar characteristics, in fact, I group assets with similar expected returns together.

## B. Estimation of Expected Stock Returns

In this section, I explain how to predict the future return premium of each asset given a large set of lagged characteristics through neighbouring stocks. This return prediction requires first defining the target variable for both in-sample and out-of-sample sets. I define target variables in the in-sample set as the decile portfolio that each asset belonged to. To do so, I sort all assets at time  $t_0 < t$  into one of the decile portfolios based on their expected returns at time  $t_0$ . Realized returns at each time consist of two components. One is the idiosyncratic part which is the noisy constituent, and the other is the systematic element which we are interested in predicting.

I model  $r_{j,t}$ , the realized return for asset  $j$  at time  $t$ , as

$$r_{j,t} = E(r_{j,t}) + e_{j,t}, \quad (1)$$

where the first part is the expected return at time  $t$  and the second part is the idiosyncratic noise. I aim for predicting classes in out-of-sample based on  $E(r_{j,t})$ , that is I assume that the systematic part is a function of most recent characteristics:

$$E(r_{j,t}) = F_{t-1}(\mathbf{x}_{j,t-1}). \quad (2)$$

Here  $F_{t-1}(\cdot)$  is a function that maps the most recent characteristics to the expected returns through neighbouring assets and allows for time-varying relationship between characteristics and expected returns. I need then to estimate  $E(r_{j,t})$  to be used as my target variable in the in-sample set. I proxy this unconditional expected return in the in-sample data as

$$E(r_{j,t}) = \frac{1}{T} \sum_{t^0=0}^{T-1} r_{j,t-t^0}. \quad (3)$$

Here  $T$  determines how much information from past we want to include in the target variable. If  $T = 1$ , then I assume that realized returns are a function of lagged characteristics:

$$r_{j,t} = F_{t-1}(\mathbf{x}_{j,t-1}) + e_{j,t}. \quad (4)$$

For  $T > 1$ , while  $E(r_{j,t})$  is capturing the information at time  $t$ , it assumes that previous information in the past  $T - 1$  months is also important for predicting the expected returns. However, if  $T$  is becoming larger than enough, very far information might be irrelevant for the estimation of future expected returns, and hence worsens the accuracy of the model for predicting the true decile. Finally, I rank each asset based on its target variables to a decile portfolio and in the out-of-sample set predict the portfolio to which each asset should belong.

## C. Methodology

In this section, I formally describe the framework of asset classification through neighbouring assets. Forming managed portfolios is, indeed, a classification problem. I deviate

from the literature by predicting the decile portfolio that each asset belongs to, instead of predicting the expected returns directly. I denote decile portfolio  $i$  by  $c_i$ , such that  $c_1$  and  $c_{10}$  are loser and winner portfolios respectively. Suppose that asset  $j$  at time  $t_0$ ,  $a_{j,t_0}$ , has the most recent characteristics  $\mathbf{x}_{j,t_0-1}$ , with an expected return  $E(r_{j,t_0})$ . At each month in the in-sample period, I sort all assets based on their expected returns to one of the decile portfolios, implying to have a balanced dataset in the in-sample period. The decile portfolio  $C_{t_0}$  for asset  $a_{j,t_0}$  at each time is a monotonic function of its expected returns:

$$C_t(a_{j,t_0}) = h(E(r_{j,t_0})/R_{t_0}), \quad (5)$$

where  $R_{t_0}$  is the whole cross-section of returns at time  $t_0$ . Apparently, as  $h(\cdot)$  is a decile maker function, all the assets in the cross-sections belong to one of  $c_1, c_2, \dots, c_{10}$ , with the same number of assets in each one. These decile portfolios are the target variable which the model aims to predict, while the set of lagged-characteristics  $\mathbf{x}_{j,t_0-1}$  are the explanatory variables for asset  $j$  at time  $t_0$ . Ultimately, I aim to map lagged-characteristics  $\mathbf{x}_{j,t-1}$  to the decile portfolios  $C_t(a_{j,t}) \in \{c_1, c_2, \dots, c_{10}\}$  at each time  $t$  in the out-of-sample period.

For classifying each asset at time  $t$ , the in-sample period includes all available information up to time  $t-1$ . I consider  $t$  months for the length of the rolling window of in-sample data. Therefore, the in-sample set potentially includes all data points in

$$S_t = \left( \mathbf{x}_{s,t-t^\theta-1}, C_{t-t^\theta}(a_{s,t-t^\theta}) \right) \quad (6)$$

where  $t^\theta = 1, \dots, t$  and for all assets  $a_{j,t-t^\theta}$  in this time period, where target variables  $C_{t-t^\theta}(a_{j,t-t^\theta})$  are already known to the investors at time  $t$ . Finally,  $S_t$  is the in-sample set I use for time  $t$ , and the out-of-sample includes the whole cross-section of assets at time  $t$ .

In this paper, I use a distance-weighted scheme for decile prediction. I assume that *closer* assets would contribute more than further assets even after identifying the  $k$  nearest ones. I define the distance weights for asset  $a_{s,t-t^\theta}$  when comparing to the asset  $a_{j,t}$  as:

$$w(a_{s,t-t^\theta}|a_{j,t}) = \begin{cases} \frac{1}{l(\mathbf{x}_{s,t-t^\theta-1}, \mathbf{x}_{j,t-1})} & \text{if } a_{s,t-t^\theta} \in k \text{ neighbours of } a_{j,t} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

for some distance measure  $l$ . Clearly, in equation 7 assets which are among the  $k$  neighbours of  $a_{j,t}$  get a weight proportional to the inverse of their distance, while other assets get a weight of zero, i.e., they are ignored for decile prediction. I show that the model performance is robust to different measures of  $l$ . I consider a euclidean distance measure in this context and treat all the features equally, that is, I weigh characteristics evenly assuming that they contribute equally to separating assets with different expected returns.

Finally, for classifying asset  $a_{j,t}$ , I sum over the weights of each class and put  $a_{j,t}$  into the decile portfolio which has the highest summation. Mathematically,

$$C_t(a_{j,t}) = \arg \max_{c \in \{c_1, \dots, c_{10}\}} \sum_{s \in S_t} w(a_{s,t-t^\theta}|a_{j,t}) d(c, C_{t-t^\theta}(a_{s,t-t^\theta})), \quad (8)$$

where  $d(\cdot, \cdot)$  is a Kronecker delta function:

$$d(x_1, x_2) = \begin{cases} 1, & \text{if } x_1 = x_2 \\ 0, & \text{otherwise} \end{cases}. \quad (9)$$

Intuitively, asset  $a_{j,t}$  goes to decile portfolio  $c \in \{c_1, \dots, c_{10}\}$  if the summation of weights of assets in portfolio  $c$  in the in-sample data is the highest. For an equally weighted (versus distance weighted) scheme, the equation 8 will simply become a mode function. Therefore equation 8 can be seen as a weighted mode function. Indeed, this algorithm is a local classifier so that its viewing angle stretches only to the  $k$ th neighbour. However, if  $k$  approaches  $|S_t|$ , the number of entire assets in the training set, the method becomes a global classifier.

#### D. Bayesian Perspective of Neighbouring Assets Algorithm

Now, I estimate the posterior probability that asset  $a_{j,t}$  belongs to portfolio  $c_i$  for  $i = 1, \dots, 10$ . For the sake of simplicity, I first consider the case that each neighbouring asset in the in-sample data has equal weight, that is, uniform weighting. For each asset  $a_{j,t}$  with characteristics  $\mathbf{x}_{j,t-1}$ , consider a region  $R$  in characteristics hyperspace with the center  $\mathbf{x}_{j,t-1}$  such that  $R$  captures a set of  $k$  assets from set  $S_t$ , namely  $a_{s^\theta, t-1}$  for  $\theta = 1, \dots, k$ , with characteristics  $\mathbf{x}_{s^\theta, t-1}$ , and  $s^\theta = 1, \dots, k$  (without loss of generality). It is clear that

$$l(\mathbf{x}_{s^\theta, t-1}, \mathbf{x}_{j,t-1}) = l(\mathbf{x}_{s, t-1}, \mathbf{x}_{j,t-1})$$

for all  $s$  which is not among the set  $s^\theta \in \{1, \dots, k\}$ , and for some distance measure  $l$ . I denote the volume of  $R$  in characteristic hyperspace as  $V$  and suppose that it contains  $n_i$  assets which belong to decile portfolio  $c_i$ . Moreover, I denote the whole number of assets belonging to portfolio  $c_i$  in the entire set  $S_t$  by  $N_i$ , so that

$$\hat{\alpha}_i N_i = |S_t|.$$

Then the conditional decile portfolio estimation for asset  $a_{j,t}$  with set of characteristics  $\mathbf{x}_{j,t-1}$  is:

$$p(a_{j,t}|c_i; \mathbf{x}_{j,t-1}) = \frac{n_i}{N_i V}. \quad (10)$$

Likewise, the unconditional probability would be

$$p(a_{j,t}|\mathbf{x}_{j,t-1}) = \frac{\hat{\alpha}_i n_i}{|S_t| V} = \frac{k}{|S_t| V}. \quad (11)$$

And the prior labels of portfolios are

$$p(c_i) = \frac{n_i}{|S_t|}. \quad (12)$$

Now using Bayes' theorem we observe that the posterior probability that asset  $a_{j,t}$  belongs to portfolio  $i$  at time  $t$  is given by:

$$p(c_i/a_{j,t}; \mathbf{x}_{j,t-1}) = \frac{p(a_{j,t}/c_i; \mathbf{x}_{j,t-1})p(c_i)}{p(a_{j,t}; \mathbf{x}_{j,t-1})} = \frac{n_i}{k}. \quad (13)$$

Next, I weigh all  $n_i$  points in the  $R$  by the inverse of their distance from asset  $a_{j,t}$ , that is,

$$w(a_{s,t}^{\theta}/a_{j,t}) = \frac{1}{l(\mathbf{x}_{s,t}^{\theta-1}, \mathbf{x}_{j,t-1})}$$

for all assets  $j$  in the region  $R$  centered by lagged-characteristics of asset  $a_{j,t}$ . In this case, the posterior probability of asset classification is

$$p(c_i/a_{j,t}; \mathbf{x}_{j,t-1}) = \frac{\hat{a}_{s^{\theta}=1}^k w(a_{s^{\theta},t}^{\theta}/a_{j,t})d(c_i, C_t^{\theta}(a_{s^{\theta},t}^{\theta}))}{\hat{a}_{s^{\theta}=1}^k w(a_{s^{\theta},t}^{\theta}/a_{j,t})}, \quad (14)$$

where  $d(.,.)$  is defined according to equation 9. Equation 14 means that the posterior probability of asset  $a_{j,t}$  belongs to portfolio  $c_i$  given a set of characteristics is proportional to the sum of the weights of its neighbours which belonged to portfolio  $c_i$ . Therefore, in order to maximise the probability of true classifications, one should put  $a_{j,t}$  to portfolio  $c_i$  that has the highest summation of weights,  $\hat{a}_{s^{\theta}=1}^k w(a_{s^{\theta},t}^{\theta}/a_{j,t})d(c_i, C_t^{\theta}(a_{s^{\theta},t}^{\theta}))$ :

$$C_t(a_{j,t}) = \arg \max_{c \in \{c_1, \dots, c_{10}\}} p(c/a_{j,t}; \mathbf{x}_{j,t-1}). \quad (15)$$

In a case where all weights are equal, equation 14 becomes as equation 13. Intuitively, it is most likely that asset  $a_{j,t}$  belongs to portfolio  $c_i$  if the majority of its *closest* neighbours have belonged to this portfolio in the past.

### E. Portfolios of Neighbouring assets

In this section, I consider another case wherein portfolios of neighbouring assets would predict future returns of individual assets. Considering in-sample panel data from the past  $t$  months, for each asset  $j$  at time  $t$ , I create a portfolio which contains asset  $j$  neighbours in each month. Formally, without loss of generality, assume that assets  $a_{1,t-t^{\theta}}, \dots, a_{k,t-t^{\theta}}$  are the  $k$  nearest neighbours of asset  $j$  at time  $t-t^{\theta}$ . For this month, I create a portfolio of these  $k$  assets:

$$r_{p_j,t-t^{\theta}} = \frac{1}{k} \hat{a}_{s=1}^k r_{s,t-t^{\theta}}, \quad (16)$$

where  $r_{p_j,t-t^{\theta}}$  is the return of portfolio of neighbouring assets of asset  $j$  at time  $t-t^{\theta}$ , for  $t^{\theta} = 1, 2, \dots, t$ .<sup>9</sup> Then I predict a return for asset  $j$  at time  $t$  (out-of-sample) as:

$$\hat{r}_{j,t} = \frac{1}{t} \hat{a}_{t^{\theta}=1}^t r_{p_j,t-t^{\theta}}. \quad (17)$$

<sup>9</sup>I find that the model is robust to distance-weighting or equally-weighting portfolios of neighbours.

Finally, I sort assets into decile portfolios based on the predicted returns and update portfolios monthly. Intuitively, when a set of assets with the same set of characteristics as asset  $j$  have generated an average return of  $r_{p_j,t}^{\theta}$  in the past, it is likely, then, that asset  $j$  at time  $t$  will have similar performance to the average performance of its neighbours in the past.

## II. Empirical Analysis

In this section, I present the empirical analysis which contains 42 years of out-of-sample study.

### A. Data

I obtain monthly individual stock returns from CRSP for all firms in the NYSE, AMEX, and NASDAQ for the period starting from January 1970 until the end of 2021. This universe of data contains more than 3.7 million individual observations.<sup>10</sup> For labelling the data, each firm must have at least  $T = 12$  past returns. Considering this removes the samples without labels, yielding more than 3.4 million observations. However, in the case that I consider the rank based on the realized returns for the in-sample data, I do not remove these observations. In the beginning, I remove the smallest 5% stocks based on their lagged market capitalization. These stocks are mostly very illiquid and have many missing characteristics. In fact, because of their small sizes, including them does not distort the results as long as we use a value-weighting scheme. However, removing them assures that the results are not affected by very tiny stocks. After removing them still the data set includes more than 3.2 million observations with 28,611 unique assets. Next, I consider three sets of data for my analysis. First is the universe of all data except the 0.05 tiniest which I refer to all stocks. Second, is the stocks which have a market capitalization above 20% of NYSE, which I refer to this as all-but-tiny stocks. Third, I consider only large stocks which include only stocks with a market cap higher than the NYSE median. These three data sets are useful for studying the clustering effect and the relationship between characteristics and expected returns while ensuring that these relationships are not derived from a specific set of stocks.

For the out-of-sample study, I start the analysis in January 1980 and consider the whole data set before that for the in-sample analysis. Of course, the in-sample set is a rolling window which moves before the out-of-sample set. For the out-of-sample study, the average number of stocks per month for all stocks exceeds 5600 assets, while more than 2800 and 1400 for all but tiny and large stocks, respectively.

For the characteristics, I use data provided by [Green et al. \(2017\)](#) and [Gu et al. \(2020\)](#) who build a large-collection of 94 stock predictors.<sup>11</sup> I use the same acronyms used by [Gu](#)

---

<sup>10</sup>I follow [Gu et al. \(2020\)](#) by including stocks with prices below \$5, share codes beyond 10 and 11, and financial firms, to make the data as large as possible. However, applying common filters, such as using only share codes 10 and 11, does not change the results.

<sup>11</sup>I download data from Dacheng Xiu website.

et al. (2020) and list them in Table A1 in the Appendix. One issue is the missing characteristics, which I follow Gu et al. (2020) by imputing them by the cross-sectional median. For finding neighbouring assets, all the features must have the same scale, otherwise, some features with larger scales dominate others, and the results will be extracted based on just those features. Therefore, I cross-sectionally demean all characteristics and make them unite variance. In order to alleviate the effect of outliers, I also winsorize data in 0.01 and 0.99 levels. I obtain risk-free rate data from Fama and French library, as well as market returns, size, value, operating profitability, and investment factors. I download  $q$  and augmented  $q$  model ( $q^5$ ) from the work based on Hou, Mo, Xue, and Zhang (2019) and Hou, Mo, Xue, and Zhang (2021).

### B. Model Parameters

The neighbouring assets framework requires only one hyper-parameter to be fixed before the analysis starts, which is a small number compared to other machine learning tools. This parameter is  $k$ , the number of neighbours for each asset. As I use a distance-weighted scheme, the role of  $k$  becomes less important, as the further neighbours receive lower weights, sometimes closer to zero. I show that results in the distance-weighted scheme are robust for large enough  $k$ s. I choose  $k = 1000$  for the main analysis and later consider different  $k$  in the robustness checks and show that the model is not sensitive to the amount of  $k$ .

Moreover, in the data processing stage, there are three hyper-parameters which must be tuned (fixed) for defining the target classes and explanatory variables. One is  $T$  in equation 3 which determines the labels. Intuitively,  $T$  captures the information I want to include in the target class of my in-sample data. The case  $T = 1$  is equivalent to mapping the most recent characteristics to the realized returns. I consider two cases where  $T = 1$ , and  $T = 12$ . For the latter case, I consider one year including the current month for predicting the labels based on the average of last year's returns. I show that the model performs well in both cases.

Another parameter which is also related to defining the set  $S_t$  is  $t$  in equation 6. Intuitively,  $t$  determines how much information from the past must be considered in the in-sample set,  $S_t$ . I assume that past data in the neighbourhood of asset  $a_{j,t}$  carry predictive information. However, very far data might be irrelevant as the pattern of characteristics might change over time. Although I find that the characteristic patterns are almost constant and do not vanish over time, the model still has the flexibility to recognize structural changes in characteristic patterns when creating decile portfolios. I show that model performance is robust over the various amounts of  $t$  when it captures at least 3 years of in-sample data. For the main analysis, I set  $t = 120$ , which means I include 10 previous years in  $S_t$  that can be used in the out-of-sample predictions.

Lastly, I consider different distance metrics rather than Euclidean distance and show the results do not change with respect to the distance measure.

### C. Out of Sample Study of Cross-Sections

I begin the out-of-sample study from January 1980 with a 10-years in-sample rolling window ( $t = 120$ ). The class labels are observable only in the in-sample set. I consider the whole 94 characteristics at once for finding neighbouring assets. Based on the neighbouring assets I predict the decile portfolio for each asset in each period, and update portfolios month by month.

Table 1 shows the results for value-weighted portfolios when 94 characteristics are considered for finding the neighbouring assets. The first four columns show the case when the decile portfolios are based on the realized returns. For example, if an asset's realized return has been among the lowest 10%, then its rank (class) is 1. In this case, the target variables are predicted based on  $r_t$ , only the realized returns. The last four columns, on the other hand, are the counterpart results when assets labels are determined based on the rank of last year's average returns,  $\frac{1}{12} \sum_{t'=0}^{11} r_{t-t'}$ . For instance, in the in-sample data, an asset whose last year's average returns has been among the lowest 10% has a rank of 1. Although the focus of my analysis is only on value-weighted portfolios, I report equally-weighted counterpart portfolios in the appendix Table A2. The mean columns show the average of monthly excess realized returns of decile portfolios in the period 1980-2021 in percentage, while  $t$ -stat tests if the realized mean returns are significantly different from zero. The columns SR report annualized Sharpe ratios. Panel (a) shows the results when the universe of stock includes all but the 5% tiniest stocks, whereas panels (b) and (c) show the results when considering all but tiny and large stocks, respectively. The "LS" shows the performance of a long-short portfolio which goes long in decile 10 and short in decile 1. On the left side, in panel (a), the average returns vary between -0.46% ( $t = -1.14$ ) and 1.25% ( $t = 3.14$ ) in the lowest and highest decile portfolio, with a 1.71% ( $t = 8.66$ ) of average returns for the long-short portfolio. The annualized Sharpe ratio in this case is the highest, 1.34. An equally weighted portfolio even has a higher Sharpe ratio of 1.87 as shown in Table A2. In the right side of Table 1, the decile 1 average is -0.91%, while decile 10 has a mean average of 1.44%, yielding to a performance of 2.36% for a long-short portfolio with  $t = 7.42$  with a Sharpe ratio of 1.15. The dispersion for panel (b) is lower, 0.93 ( $t = 6.18$ ) on the left side, and 1.33% ( $t = 5.18$ ) on the right side, significantly different than zero. The Sharpe ratio on the left side stands at 0.95. Not surprisingly, when I consider only large stocks this dispersion drops to 0.63 ( $t = 4.12$ ) and 0.88% ( $t = 3.39$ ) in the left and right side of the Table 1 respectively. Apparently, decile 1 (10) has the lowest (highest) average realized returns and all decile portfolios fairly line up monotonically.

Next, I evaluate my long-short portfolios with common factor models. I consider six famous factor models, namely CAPM, Fama French three-factor model (FF3), Carhart four-factor model (Carhart), Fama French five-factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). These factor models are based on a small set of, yet important, characteristics and have been shown to explain many anomalies. I show alphas of a value-weighted long-short portfolio, their  $t$  statistics, as well as adjusted  $R^2$  for all six-factor models in the panel (a) of Table 2 in time-series regressions. Again the first three columns are related to the case when  $r_t$  is used as the target variable and the last three

	Target variable: $r_t$				Target variable: $E(r_t)$			
	mean	std	$t$ -stat	SR	mean	std	$t$ -stat	SR
<b>Panel (a): All data</b>								
<b>1</b>	-0.46	9.07	-1.14	-0.18	-0.91	10.19	-2.01	-0.31
<b>2</b>	0.21	7.63	0.61	0.09	0.23	7.87	0.65	0.10
<b>3</b>	0.45	6.15	1.65	0.26	0.59	5.88	2.26	0.35
<b>4</b>	0.68	4.77	3.20	0.49	0.60	5.15	2.62	0.40
<b>5</b>	0.72	3.99	4.03	0.62	0.61	4.33	3.16	0.49
<b>6</b>	0.69	3.94	3.90	0.60	0.66	4.38	3.38	0.52
<b>7</b>	0.83	4.34	4.28	0.66	0.78	4.40	3.99	0.62
<b>8</b>	0.80	5.15	3.48	0.54	0.96	5.18	4.16	0.64
<b>9</b>	0.94	7.12	2.98	0.46	0.91	6.62	3.09	0.48
<b>10</b>	1.25	8.90	3.14	0.48	1.44	9.11	3.55	0.55
<b>LS</b>	1.71	4.43	8.66	1.34	2.36	7.13	7.42	1.15
<b>Panel (b): All but tiny stocks</b>								
<b>1</b>	0.22	8.35	0.58	0.09	-0.11	9.18	-0.26	-0.04
<b>2</b>	0.52	6.44	1.81	0.28	0.52	6.47	1.81	0.28
<b>3</b>	0.70	5.44	2.90	0.45	0.64	5.41	2.63	0.41
<b>4</b>	0.77	4.47	3.87	0.60	0.70	4.41	3.55	0.55
<b>5</b>	0.79	3.98	4.44	0.69	0.73	4.21	3.90	0.60
<b>6</b>	0.67	4.13	3.62	0.56	0.67	4.17	3.58	0.55
<b>7</b>	0.79	4.46	3.96	0.61	0.73	4.47	3.69	0.57
<b>8</b>	0.74	5.08	3.26	0.50	0.89	4.71	4.26	0.66
<b>9</b>	0.85	6.18	3.10	0.48	0.83	5.62	3.32	0.51
<b>10</b>	1.15	7.92	3.26	0.50	1.22	8.09	3.40	0.52
<b>LS</b>	0.93	3.39	6.18	0.95	1.33	5.77	5.18	0.80
<b>Panel (c): Large stocks</b>								
<b>1</b>	0.38	7.92	1.07	0.16	0.22	8.54	0.57	0.09
<b>2</b>	0.43	6.33	1.52	0.24	0.49	6.13	1.80	0.28
<b>3</b>	0.74	5.16	3.24	0.50	0.62	5.10	2.73	0.42
<b>4</b>	0.62	4.34	3.22	0.50	0.74	4.47	3.72	0.57
<b>5</b>	0.69	3.95	3.94	0.61	0.79	4.19	4.26	0.66
<b>6</b>	0.84	4.25	4.43	0.68	0.60	4.31	3.13	0.48
<b>7</b>	0.75	4.45	3.79	0.58	0.77	4.37	3.93	0.61
<b>8</b>	0.83	5.02	3.69	0.57	0.82	4.78	3.83	0.59
<b>9</b>	0.91	6.11	3.34	0.52	0.83	5.31	3.50	0.54
<b>10</b>	1.01	7.45	3.04	0.47	1.09	7.50	3.26	0.50
<b>LS</b>	0.63	3.45	4.12	0.64	0.88	5.81	3.39	0.52

**Table 1-** The performance of portfolios based on 94 characteristics in 1980-2021

This table reports the average monthly out-of-sample performance of value-weighted portfolios based on the all 94 characteristics listed in table A1. The four left columns show the results when the in-sample labels are ranked based on the realized returns (according to equation 4), while the four right columns show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). Columns titled "mean" show the average monthly excess returns of created portfolios from Jan 1980 to Dec 2021 in percentage. std is the monthly standard deviation of portfolios.  $t$  stat shows if the risk premiums are significantly different from zero and SR demonstrates the annualized Sharpe ratio. Panel (a) considers all data except the 5% tiniest for creating portfolios, while panel (b) and (c) includes all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median), respectively. LS shows the performance of a long-short portfolio which goes long (short) in decile 10 (1). The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The equally-weighted counterparts are shown in table A2.

columns show the results when the target variable is  $\frac{1}{12} \hat{\alpha}_{t^0=0}^{11} r_{t-t^0}$ . The equally-weighted portfolio alphas are shown in Table A3. For the case of all data in the right side of Table 2 panel (a), a FF3 model generates the highest monthly alpha of 2.72% with  $t = 8.86$ . The left side FF3 is equally significant: FF3 monthly alpha 1.77% with  $t = 8.89$ , which is the highest  $t$  value. Still, the equally-weighted portfolio has a higher significant alpha: 1.77% with  $t = 12.30$ , as shown in Table A3. In the right side of Table 2, as the target variables are correlated with the momentum, a Carhart model adjusted  $R^2$  varies between 0.57 and 0.63. Still, in all cases, alphas from a Carhart model are significantly different than zero: for all data with 1.81% ( $t = 8.39$ ). On the left side, the  $t$  values are almost equal to the right side in all cases.

These portfolios are very well diversified. Panel (b) of Table 2 shows the average number of assets in each portfolio in the out-of-sample set. Of course, portfolios in the in-sample have an equal number of assets.

Now I consider two simpler models, one with three well-known characteristics including size, book-to-market and 12-month momentum and the second model which adds 9 more characteristics to the previous ones. These 9 characteristics are documented to be among the most important features by Gu et al. (2020) among others. The second model, therefore, contains 12 characteristics including 1-month momentum (Jegadeesh and Titman (1993)), change in momentum (Gettleman and Marks (2006)), maximum daily return (Bali et al. (2011)), industry momentum (Moskowitz and Grinblatt (1999)), return volatility (Ang et al. (2006)), dollar trading volume (Chordia et al. (2001)), sales to price (Barbee Jr et al. (1996)), share turnover (Datar et al. (1998)), and asset growth (Datar et al. (1998)) as well as the 3 characteristics in the first model. The neighbouring assets will be found only using these characteristics and the portfolios are updated monthly. The results for value-weighted portfolios are shown in Table 3. Panel (a) shows the case when the target variable is  $r_t$ , while panel (b) shows the counterpart results for the case when the target variable is  $\frac{1}{12} \hat{\alpha}_{t^0=0}^{11} r_{t-t^0}$ . Mean (1), (10) and (LS) show the average returns for the lowest and highest deciles, and a long-short portfolio, respectively. SR shows the annualized Sharpe ratio of the long-short portfolio. In both panels in all data sets the Sharpe ratios and  $t$ -values increase by increasing the number of characteristics, suggesting that the predictive power of neighbouring assets increases when more characteristics are taken into account for finding the neighbours. Even for large stocks, in most cases alphas are significant. For all data, the Sharpe ratio almost doubles once we move from 3 characteristics to 94 characteristics. The equally-weighted portfolios are shown in Table A4.

### C.1. Stochastically shrinking the in-sample set

The in-sample set includes 120 months of data, and in each month there are several thousands of assets. Therefore, in-sample data contains several hundred thousand assets. Finding  $k = 1000$  nearest neighbours through several hundred thousand is computationally time-consuming. When I use the average returns to define the target variables, it is likely that most of the observations in the in-sample data do not affect the results. In other words, if one considers a smaller set of data the results should not change. However, as I assume

	Target variable: $r_t$			Target variable: $E(r_t)$		
	All data	All but tiny	Large stocks	All data	All but tiny	Large stocks
<b>Panel (a): Performance of long-short portfolios</b>						
$a_{\text{CAPM}}$	1.70	0.99	0.72	2.51	1.50	1.03
$t$ -stat	8.51	6.52	4.71	7.85	5.86	4.00
adj $R^2$	0.00	0.01	0.02	0.01	0.03	0.03
$a_{\text{FF3}}$	1.77	1.02	0.77	2.72	1.68	1.26
$t$ -stat	8.89	6.68	5.02	8.86	6.88	5.21
adj $R^2$	0.02	0.01	0.04	0.10	0.13	0.16
$a_{\text{Carhart}}$	1.34	0.65	0.41	1.81	0.92	0.52
$t$ -stat	7.78	5.18	3.20	8.39	5.70	3.17
adj $R^2$	0.29	0.36	0.36	0.57	0.63	0.62
$a_{\text{FF5}}$	1.56	0.88	0.64	2.52	1.50	1.08
$t$ -stat	7.63	5.59	4.02	7.90	5.90	4.33
adj $R^2$	0.04	0.03	0.06	0.11	0.14	0.17
$a_q$	1.33	0.70	0.47	1.95	1.02	0.67
$t$ -stat	6.75	4.62	3.09	6.40	4.28	2.77
adj $R^2$	0.13	0.13	0.15	0.20	0.25	0.24
$a_{q^5}$	1.18	0.56	0.29	1.73	0.73	0.27
$t$ -stat	5.61	3.52	1.78	5.32	2.88	1.05
adj $R^2$	0.14	0.14	0.17	0.20	0.27	0.27
<b>Panel (b): The average number of assets in each portfolio</b>						
<b>1</b>	1001	489	188	620	271	103
<b>2</b>	412	115	49	565	258	96
<b>3</b>	322	147	83	476	233	125
<b>4</b>	498	223	148	551	303	184
<b>5</b>	509	415	239	692	352	217
<b>6</b>	1052	526	274	697	338	185
<b>7</b>	724	378	202	599	309	157
<b>8</b>	555	244	108	512	256	117
<b>9</b>	585	213	87	551	274	119
<b>10</b>	446	371	184	365	238	115

**Table 2-** Risk adjusted returns for value-weighted long-short portfolios

Panel (a) of this table reports monthly alphas,  $t$  values and adjusted  $R^2$  for out-of-sample performance of value-weighted long-short portfolios. The three left columns show the results when the in-sample stocks are ranked based on the realized returns (according to equation 4), while the three right columns show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). All data includes the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The counterpart results for equally-weighted portfolios are shown in table A3 in the appendix. Panel (b) shows the average number of assets in each portfolio.

	All data		All but tiny		Large stocks	
	3 char	12 char	3 char	12 char	3 char	12 char
<b>Panel (a): Predicting <math>r_t</math></b>						
<b>mean (1)</b>	0.23	-0.23	0.36	0.10	0.47	0.25
<i>t</i> -stat	0.61	-0.57	1.11	0.26	1.58	0.74
<b>mean (10)</b>	1.25	1.37	1.17	1.28	1.16	1.17
<i>t</i> -stat	3.63	3.72	3.83	3.85	4.00	3.69
<b>mean (LS)</b>	1.02	1.60	0.82	1.18	0.68	0.92
<i>t</i> -stat	4.29	8.22	3.92	6.79	3.30	5.68
<b>SR</b>	0.66	1.27	0.61	1.05	0.51	0.88
$\partial\text{CAPM}$	1.16	1.69	0.91	1.30	0.76	1.00
<i>t</i> -stat	4.88	8.64	4.34	7.52	3.66	6.14
$\partial\text{FF3}$	1.28	1.75	1.03	1.33	0.91	1.04
<i>t</i> -stat	5.52	8.93	5.09	7.67	4.59	6.44
$\partial\text{Carhart}$	0.60	1.29	0.36	0.87	0.24	0.62
<i>t</i> -stat	3.64	7.87	2.97	6.47	2.10	4.88
$\partial\text{FF5}$	1.03	1.41	0.84	1.10	0.80	0.85
<i>t</i> -stat	4.30	7.17	4.00	6.19	3.89	5.11
$a_q$	0.60	1.19	0.47	0.88	0.44	0.64
<i>t</i> -stat	2.73	6.37	2.39	5.20	2.24	4.07
$a_{q^5}$	0.46	1.05	0.28	0.69	0.29	0.44
<i>t</i> -stat	1.96	5.30	1.35	3.82	1.36	2.66
<b>Panel (b): Predicting <math>E(r_t)</math></b>						
<b>mean (1)</b>	-0.32	-0.67	-0.02	-0.17	0.17	0.00
<i>t</i> -stat	-0.72	-1.44	-0.04	-0.42	0.47	0.00
<b>mean (10)</b>	1.32	1.29	1.30	1.15	1.15	1.11
<i>t</i> -stat	3.89	3.54	4.19	3.46	3.98	3.54
<b>mean (LS)</b>	1.64	1.96	1.31	1.32	0.99	1.11
<i>t</i> -stat	4.49	5.57	4.05	4.46	3.11	3.92
<b>SR</b>	0.69	0.86	0.63	0.69	0.48	0.60
$\partial\text{CAPM}$	1.88	2.22	1.53	1.57	1.18	1.32
<i>t</i> -stat	5.16	6.34	4.75	5.36	3.73	4.71
$\partial\text{FF3}$	2.13	2.45	1.77	1.78	1.43	1.52
<i>t</i> -stat	6.08	7.30	5.74	6.42	4.79	5.64
$\partial\text{Carhart}$	0.91	1.37	0.64	0.83	0.33	0.61
<i>t</i> -stat	4.84	6.47	4.56	5.33	2.47	3.92
$\partial\text{FF5}$	1.82	2.19	1.44	1.59	1.16	1.34
<i>t</i> -stat	5.02	6.29	4.53	5.51	3.75	4.79
$a_q$	1.05	1.50	0.77	1.01	0.56	0.81
<i>t</i> -stat	3.16	4.64	2.61	3.70	1.89	3.04
$a_{q^5}$	0.69	1.00	0.41	0.57	0.22	0.39
<i>t</i> -stat	1.94	2.93	1.31	1.97	0.68	1.38

**Table 3-** The performance of value-weighted portfolios based on 3 and 12 characteristics in 1980-2021

This table reports the average monthly out-of-sample performance of value-weighted portfolios based on 3 and 12 characteristics. Panel (a) show the results when the in-sample labels are ranked based on the realized returns (according to equation 4), while panel (b) show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). Rows titled "mean (1), (10) and (LS)" show the average monthly excess returns of portfolios 1, 10 and long-short portfolio from Jan 1980 to Dec 2021 in percentage and SR demonstrates the annualized Sharpe ratio. The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The equally-weighted counterpart results are presented in table A4.

that the past data up to 10 years contain predictive information, instead of shortening the length of the in-sample rolling window, I propose a stochastic selection of assets that yields shrinking the in-sample set. In this case, I uniformly draw a sample from the training set and make the out-of-sample prediction based on the selected observations in the in-sample. By doing so, outliers are very likely to be removed from the in-sample set, helping to increase the model performance. [Kusner, Tyree, Weinberger, and Agrawal \(2014\)](#) point out that using the stochastic selection in the in-sample data makes the  $k$ NN classifier substantially more robust in noisy environments.<sup>12</sup> With a stochastic selection of assets in the in-sample data, it turns out that the performance of the model does not come from a specific group of stocks. This suggests that the factor structure embedded in cross-sections of returns exists even in a smaller group of randomly selected stocks. Here I show that the dispersion does not disappear, if not getting stronger, when the size of in-sample data is stochastically shrunk.

I consider a parameter  $\rho$  which affects the size of in-sample data  $S_t$ .  $\rho = 1$  means that the whole universe of stocks is included for finding the neighbours (no stochastic selection). A  $\rho < 1$  samples from the universe of data and finds the neighbouring assets among them. A small  $\rho$  reduces the sample size with the order of  $1 - \rho$ . A combination of  $\rho$  and  $t$  provides a wealth of information from the time-series and the cross-sections with reducing the noise effect. While a large  $t$  (in this 120 months) assures that the past information from long enough is included in the in-sample,  $\rho$  prunes this sample by removing the outliers and noisy data. With a stochastic selection, in order to keep the in-sample size relatively small for the three data sets, I repeat the analysis with  $\rho = 0.05$  for three different data sets, meaning that only 5% of assets in the in-sample are used for out-of-sample predictions. In section [II.G.4](#), I show that the results are robust for different amounts of  $\rho$ . The average returns of long-short portfolios for all panels increase by increasing the number of characteristics, implying, again, that many characteristics jointly contribute to the cross-sectional variation in expected returns. For the case of 94 characteristics, the mean realized returns for all, all but tiny, and large stocks value-weighted long-short portfolios are 2.41% ( $t = 7.03$ ), 1.75% ( $t = 5.96$ ) and 1.30% ( $t = 3.65$ ), respectively. The realized returns line up fairly monotonically, with decile 1 (10) having the lowest (highest) average realized returns.

Considering 3, 12 and 94 characteristics, [Table 4](#) shows the performance of value-weighted long-short portfolios when there is a stochastic selection with  $\rho = 0.05$ . In most cases, alphas and  $t$ -statistics increase by increasing the number of characteristics. When 94 characteristics are considered in the third column, a FF3 model generates the highest monthly alpha of 2.70% with  $t = 8.15$ . With 94 characteristics, for all but tiny and large stocks FF3 alphas are 2.11% ( $t = 7.53$ ) and 1.71% ( $t = 4.94$ ), respectively. Almost all alphas are economically and statistically significant. For all data, all monthly alphas are above 1.65%. The counterpart results for equally-weighted portfolios are shown in [Table A5](#).

---

<sup>12</sup>[Hinton and Roweis \(2002\)](#) and [Tarlow, Swersky, Charlin, Sutskever, and Zemel \(2013\)](#) propose stochastic selections for a  $k$ -nearest neighbours classifier in a machine learning setting.

	All data			All but tiny			Large stocks		
	3 char	12 char	94 char	3 char	12 char	94 char	3 char	12 char	94 char
$\alpha_{\text{CAPM}}$	1.92	1.99	2.46	1.59	1.61	1.88	1.28	1.44	1.50
$t$ -stat	5.41	5.81	7.08	4.89	5.87	6.35	3.97	4.94	4.20
adj $R^2$	0.03	0.01	0.00	0.03	0.02	0.01	0.02	0.03	0.02
$\alpha_{\text{FF3}}$	2.14	2.23	2.70	1.83	1.81	2.11	1.52	1.62	1.71
$t$ -stat	6.21	6.79	8.15	5.92	6.86	7.53	4.95	5.77	4.94
adj $R^2$	0.10	0.10	0.10	0.13	0.11	0.13	0.13	0.10	0.09
$\alpha_{\text{Carhart}}$	0.94	1.23	1.85	0.70	0.96	1.36	0.42	0.75	0.93
$t$ -stat	5.05	5.48	7.04	4.92	5.77	6.29	2.76	4.00	3.16
adj $R^2$	0.74	0.59	0.45	0.82	0.66	0.50	0.79	0.61	0.36
$\alpha_{\text{FF5}}$	1.78	2.06	2.56	1.51	1.68	1.96	1.26	1.47	1.69
$t$ -stat	5.03	6.04	7.41	4.73	6.15	6.72	3.98	5.02	4.67
adj $R^2$	0.12	0.10	0.10	0.15	0.12	0.13	0.14	0.11	0.09
$\alpha_q$	1.03	1.47	1.93	0.83	1.19	1.49	0.64	0.94	1.12
$t$ -stat	3.20	4.47	5.78	2.77	4.44	5.20	2.09	3.31	3.16
adj $R^2$	0.29	0.19	0.18	0.28	0.17	0.17	0.22	0.18	0.14
$\alpha_{q^5}$	0.72	1.00	1.65	0.49	0.78	1.16	0.31	0.56	0.58
$t$ -stat	2.11	2.88	4.63	1.55	2.77	3.81	0.96	1.86	1.54
adj $R^2$	0.30	0.21	0.18	0.29	0.20	0.19	0.23	0.20	0.17

**Table 4-** Risk adjusted returns for a value-weighted long-short portfolio with a stochastic selection

This table reports monthly alphas,  $t$  values and adjusted  $R^2$  for out-of-sample performance of value-weighted long-short portfolios where there is a stochastic selection in the training sample with  $\rho = 0.05$ . The results are shown for the case three cases when 3 characteristics, 12 characteristics and 94 characteristics are used to find neighbouring assets. The target variables are defined based on the average of last year returns ( $T = 12$  in equation 3). All data includes the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The number of neighbours is considered  $k = 1000$  with  $t = 120$  of months included in the in-sample data. The counterpart results for equally-weighted portfolios are shown in table A5 in the appendix.

## C.2. Portfolios of Neighbouring Assets

In this section, I create portfolios of neighbouring assets according to section I.E. I use all but the 5% tiniest stocks for creating portfolios. For each asset at each month between Jan 1980 and Dec 2021, I create  $t = 120$  months past portfolios, which contain  $k = 50$  nearest neighbours at each month.<sup>13</sup> By monthly updating portfolios, I show the out-of-sample performance of decile portfolios in Table 5 panel (a). On average there are 643 assets in each portfolio. The left side shows the results for a value-weighted portfolio while the right side presents the equally-weighted counterparts. Again, the average realized returns line up monotonically, and a value-weighted long-short portfolio generates a monthly average return of 1.66% ( $t = 6.17$ ), with an annualized Sharpe ratio of 0.95. The equally-weighted long-short portfolio has a higher average monthly return of 2.79% ( $t = 13.74$ ) with a Sharpe

<sup>13</sup>I find that the results are robust to different values of  $k$ .

	value-weighted			equally-weighted		
	mean	<i>t</i> stat	SR	mean	<i>t</i> -stat	SR
<b>Panel (a): Decile portfolios</b>						
<b>1</b>	-0.37	-1.08	-0.17	-0.61	-1.68	-0.26
<b>2</b>	0.35	1.36	0.21	0.45	1.75	0.27
<b>3</b>	0.47	2.21	0.34	0.57	2.60	0.40
<b>4</b>	0.61	2.84	0.44	0.60	2.73	0.42
<b>5</b>	0.64	3.02	0.47	0.75	3.42	0.53
<b>6</b>	0.72	3.31	0.51	0.85	3.81	0.59
<b>7</b>	0.77	3.56	0.55	0.93	4.05	0.63
<b>8</b>	0.77	3.44	0.53	1.05	4.42	0.68
<b>9</b>	0.92	3.64	0.56	1.26	4.89	0.75
<b>10</b>	1.29	4.20	0.65	2.18	6.44	0.99
<b>LS</b>	1.66	6.17	0.95	2.79	13.74	2.12
<b>Panel (b): Risk adjusted returns of long-short portfolios</b>						
	alpha	<i>t</i> stat	adj $R^2$	alpha	<i>t</i> stat	adj $R^2$
<b>CAPM</b>	1.70	6.23	0.00	2.83	13.72	0.00
<b>FF3</b>	1.78	6.56	0.02	2.87	13.98	0.02
<b>Carhart</b>	1.05	5.04	0.45	2.38	14.03	0.35
<b>FF5</b>	1.38	5.00	0.06	2.61	12.44	0.05
<i>q</i>	1.03	3.91	0.17	2.36	11.64	0.13
<i>q</i> <sup>5</sup>	0.82	2.91	0.18	2.19	10.12	0.14

**Table 5-** Out-of-sample performance of decile portfolios formed by the portfolios of neighbouring assets

This table reports the out-of-sample performance of decile portfolios predicted based on the portfolios of neighbouring assets. For each asset, I create a portfolio containing 50 assets of its nearest neighbours for each month in all past 120 months. Then I predict future return of this asset based on the monthly average of its 120 past neighbouring portfolios. Finally, I sort assets into decile portfolios based on the predicted returns. I update decile portfolios monthly from Jan 1980 till Dec 2021. The average number of assets in each decile portfolio is 643 assets. Panel (a) shows the out-of-sample performance of decile portfolios. Column "mean" shows the monthly average returns of each portfolio, and SR shows the annualized Sharpe ratio. LS is a long-short portfolio. In panel (b), I show monthly risk-adjusted returns of long-short portfolios with respect to six factor models. adj  $R^2$  shows the adjusted  $R^2$  of time-series regressions. The left (right) side presents value-weighted (equally-weighted) portfolios.

ratio of 2.12. Panel (b) shows the risk-adjusted returns for long-short portfolios. The CAPM monthly alpha for a value-weighted long-short portfolio is 1.70% ( $t = 6.23$ ), while for an equally-weighted counterpart, it increases to 2.83% ( $t = 13.72$ ).

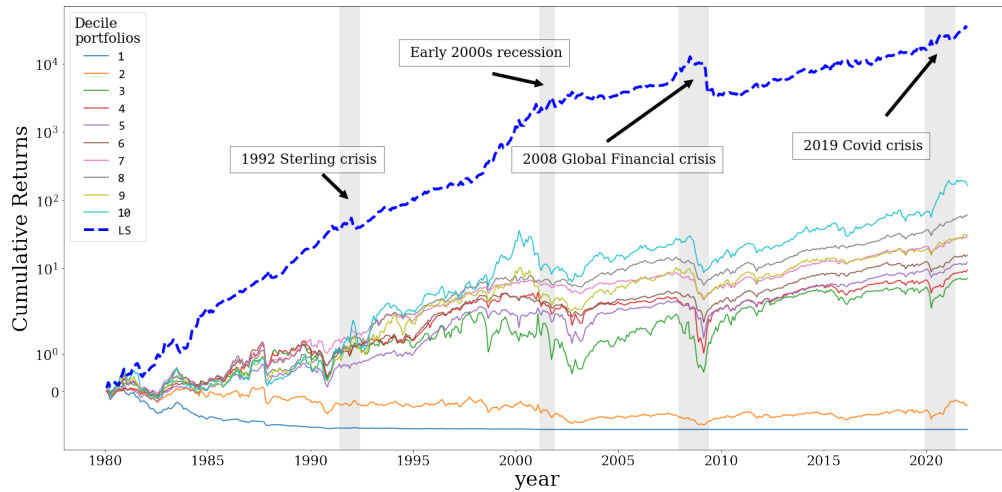
The dispersion increases when I increase the number of portfolios. Sorting assets into 25 portfolios leads to a Sharpe ratio of 1.29 (2.81) for a value-weighted (equally-weighted) long-short portfolio with a monthly average return of 3.14% with  $t = 8.37$  (4.52%,  $t = 18.21$ ). The results stay the same when I create a distance-weighting portfolio of neighbours, that is, weighting each asset based on the inverse of their distance according to equation 7. The results also are robust to changing the number of neighbours in each portfolio in the in-sample data. The results are weaker when I remove tiny stocks. For example, a value-weighted long-short portfolio from all but tiny stocks generates a monthly FF3 alpha of 0.89% ( $t = 3.84$ ). This alpha drops to 0.46% ( $t = 2.21$ ) when considering only large stocks.

#### D. *The Dispersion over Time*

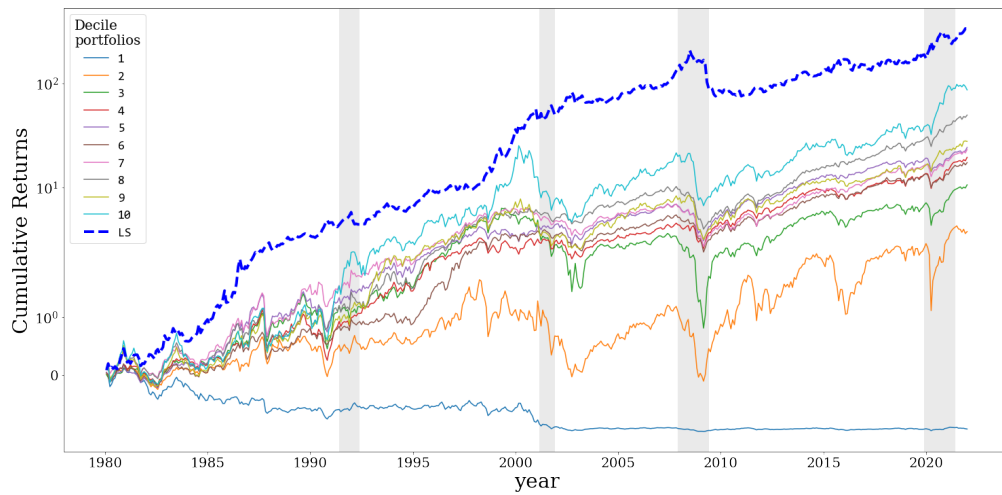
In the long run, my long-short portfolios generate huge cumulative returns. The cumulative returns of a long-short portfolio are almost ascending in over 40 years. Figure 3 shows the cumulative returns of all decile portfolios in the period 1980-2021 where the target variable is the average of the past 12 months. Panel (a), (b) and (c) illustrate cumulative returns when all, all but tiny, and large stocks are considered for creating the portfolios. The cumulative returns are shown in the vertical axis on a logarithmic scale. In panel (a), the decile portfolio 1 and 2 generate extremely negative cumulative returns. For other panels, decile portfolio 1 produces negative cumulative returns constantly. In all cases, the cumulative returns fairly line up monotonically with the number of decile portfolios. There are four periods that the behaviour of long-short portfolios turn to descending and they are known as the financial crisis shown in grey colour: The first period is the 1992 Sterling crisis, the second is the early 2000s recession, the third is the 2008 global financial crisis, and the last one is 2019 Covid crisis. All of these crisis affects the performance of a long-short portfolio but still the return trend is recovered after the crisis period ends. The cumulative returns for the case where there is a stochastic selection are shown in Figure A2. When the target variable is the current realized returns, the cumulative returns are shown in Figure A1.

The dispersion generated by decile portfolios persists over time. In Figure 4, I plot the rolling average of realized returns for deciles 1 and 10 with a stochastic selection. The left column shows the 240 months (20 years) of rolling averages, the middle column shows a 10-year rolling average, and the right column shows 5 years moving average of realized returns. Panel (a), (b) and (c) show portfolios consisting of all, all but tiny, and large stocks, respectively. For 20 years moving averages, deciles 1 and 10 always produce the moving average spread of more than 1.26% for all data, more than 0.84% for all but tiny, and more than 0.51% for large stocks. The averages of these rolling spreads for a 20-year rolling window are 2.56% (with std = 0.64), 1.45% (std = 0.25) and 1.04% (std = 0.18) for all, all but tiny and large stocks, respectively. The spread, naturally, decreases in shorter periods. The average of rolling spread when considering only 5 years reaches 2.35% (std = 1.34), 1.28% (std = 0.83) and 0.89% (std = 0.67) for all, all but tiny and large stocks. All of the means are significantly positive. As Figure 4 shows for the rolling windows of 120 and 60, the mean returns of two deciles approach each other when the 2008 financial crisis occurs. However, after the crisis ends again the spread starts getting larger. In the 2019 Covid crisis, these two deciles in the 60 months rolling window, do not touch each other. For the period 120 and 240 months, the long-short portfolio average returns are always positive even in a crisis time. I show the counterpart graphs with a stochastic selection in Figure A3 and when the target variable is the realized returns in Figure A5.

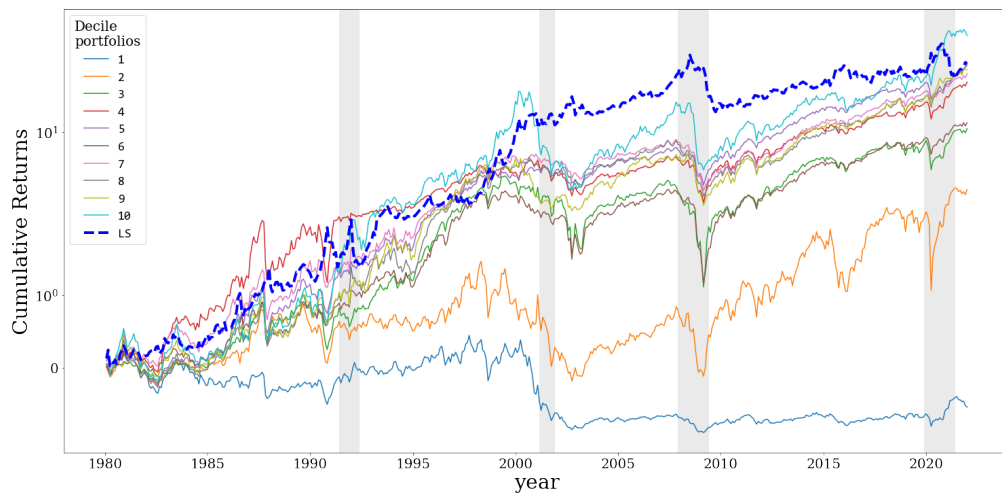
I also show the behaviour of annualized Sharpe ratio for deciles 1 and 10 as well as a long-short portfolio, all value-weighted, in Figure 5. I show the crisis periods with grey. It is clear that all Sharpe ratios react to the crisis period, especially since there is a decreasing trend in the 2008 financial crisis. However, for shorter rolling periods such as 5 years rolling windows as shown in the right columns, the Sharpe ratio for a long-short portfolio



(a) All data



(b) All but tiny stocks



(c) Large stocks

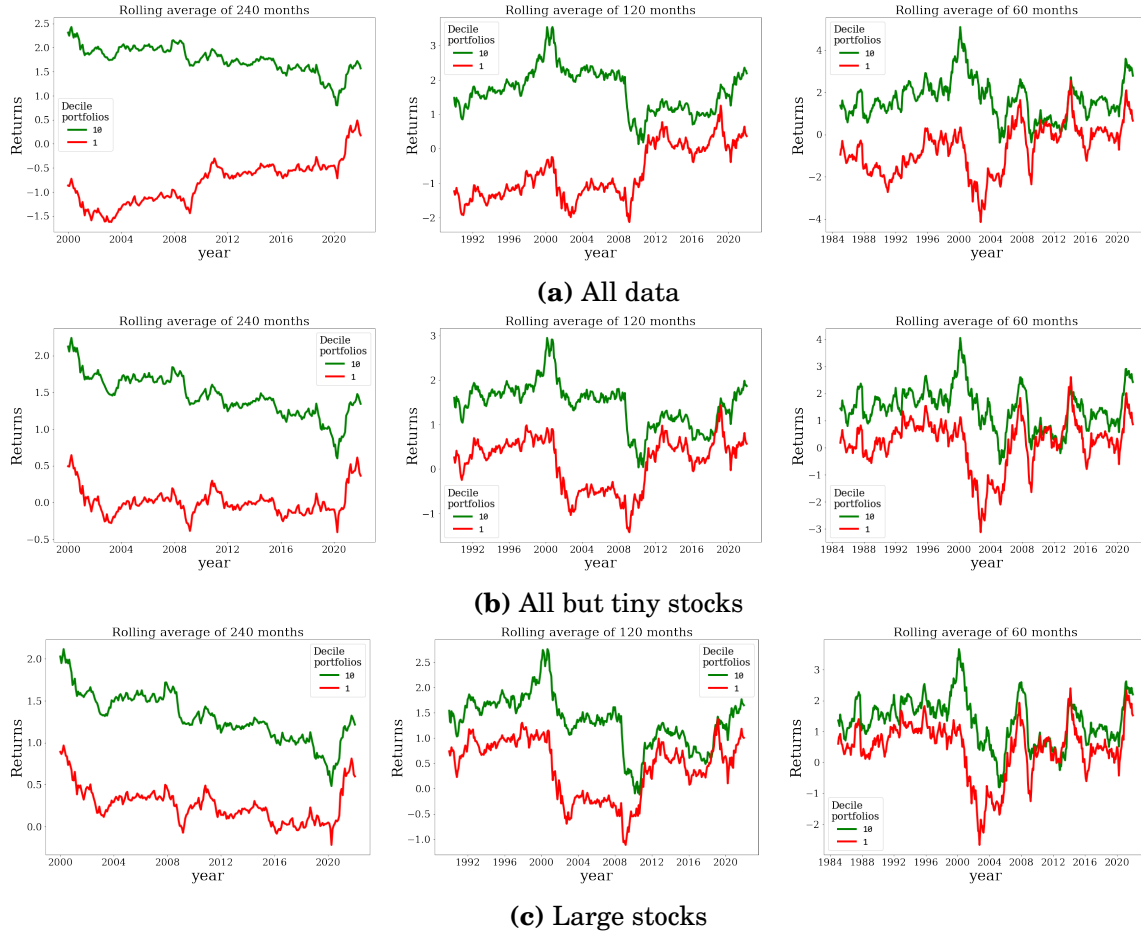
**Figure 3.** Cumulative returns of decile portfolios for the period 1980-2021

This figure shows the cumulative returns of all portfolios from 1980-2020 in the main analysis when  $k = 1000$  and the target variables are defined based on the average of last year's returns ( $T = 12$  in equation 3). Panel (a) shows portfolios consisting of all but 5% tiniest ones. Panel (b) and (c) show portfolios containing all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median). Four crisis periods are shown in grey and they include the 1992 Sterling crisis, the Early 2000s recession, 2008 global financial crisis and 2019 Covid crisis. The long-short portfolio strategy reacts to crisis periods. The y-axis is in a logarithmic scale.

recovers to 1.11 in January 2015 in panel (a) for all data. The increasing pattern also can be seen in all but tiny and large stocks. For all but tiny stocks, the average of a long-short portfolio Sharpe ratio is 0.83 for all rolling windows. Results for the case with a stochastic selection and with predicting realized returns are shown in figures A4 and A6.

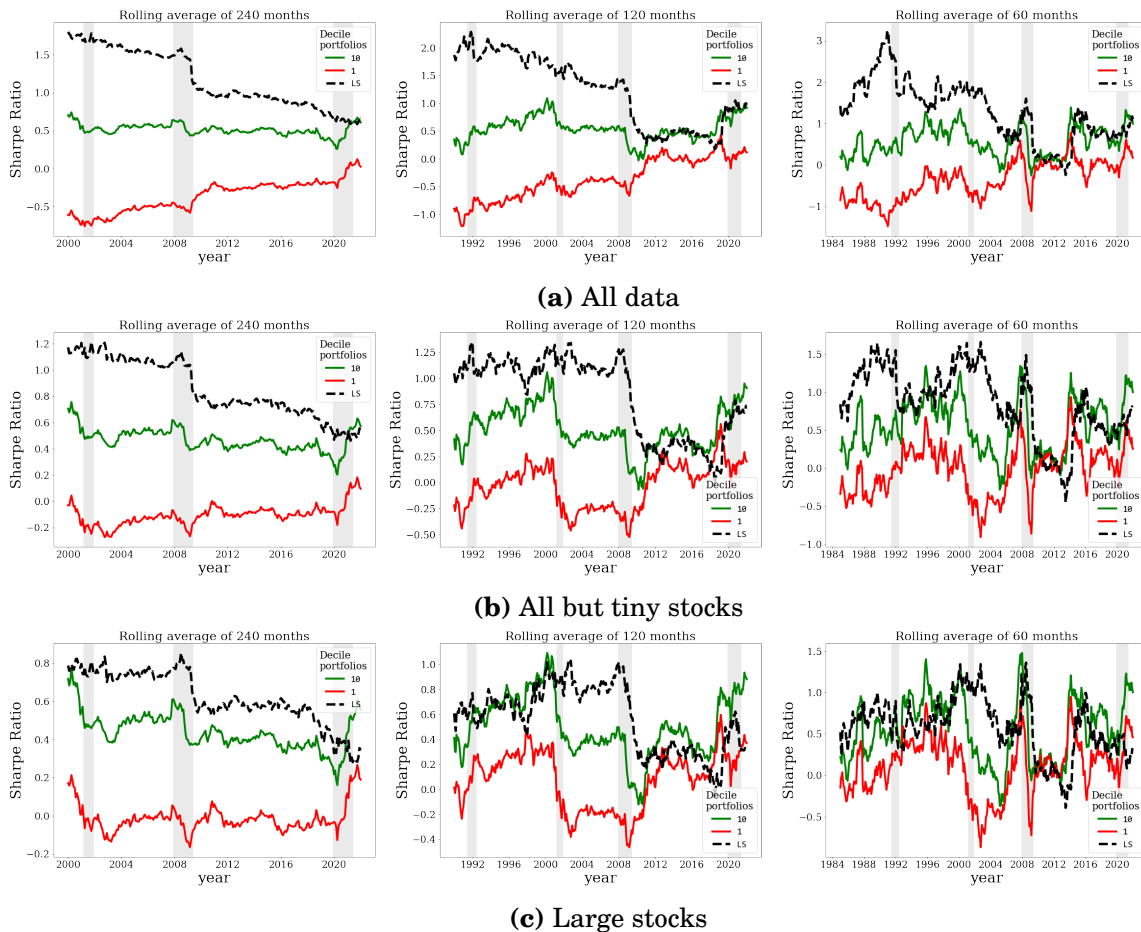
### D.1. Subsample analysis

Now I consider three different time horizons and study the performance of value-weighted long-short portfolios in these horizons. First, I consider the first 20 years, i.e. from the beginning of 1980 until the end of 1999 (1980-2000). Second I consider the time after 2000 till the end of 2021 (2000-2021). Lastly, I focus on the performances of the last 12 years (2010-2021). Table 6 shows the results. The first three left columns show the results for the case that the realized returns, while the second three columns show the counterpart results for the case where the decile portfolios show the case where the target variable is a decile portfolio based on  $\frac{1}{12} \hat{\alpha}_{t^*=0}^{11} r_{t-t^*}$ . Panel (a) shows the results for the all data. In the first three columns in panel (a), all of the alphas are economically meaningful and statistically significant. The Sharpe ratio for the last 12 years is 0.88. For the counterpart case in the second three columns, the Sharpe ratio is 0.98 in the last 12 years. Panels (b) and (c) show the results for all but tiny and large stocks. In most cases, the mean returns and alphas are statistically significant and economically large. Overall the results seem robust in different time horizons.



**Figure 4.** Rolling average of spreads between decile 1 and 10 in the period 1980-2021

This figure shows the rolling average of the spread generated from decile portfolios 1 and 10. The left column shows the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph, the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the average of last year's returns ( $T = 12$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ .



**Figure 5.** Rolling average of Sharpe Ratio for decile 1 and 10 and a long-short portfolio in the period 1980-2021

This figure shows the rolling average of the Sharpe Ratios generated from decile portfolios 1 and 10 and also a long-short portfolio. The left column shows the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph, the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the average of last year's returns ( $T = 12$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ . The crisis periods are shown in grey.

	Predicting $r_t$			Predicting $E(r_t)$		
	1980-2000	2000-2021	2010-2021	1980-2000	2000-2021	2010-2021
<b>Panel (a): All data</b>						
<b>mean</b>	2.40	1.08	0.90	3.18	1.61	1.77
<i>t</i> -stat	8.74	3.90	3.06	8.06	3.32	3.40
<b>SR</b>	1.95	0.83	0.88	1.80	0.71	0.98
<i>a</i> <sub>CAPM</sub>	2.17	1.22	0.95	3.01	1.95	2.25
<i>t</i> -stat	8.04	4.46	3.08	7.54	4.20	4.29
<i>a</i> <sub>FF3</sub>	2.39	1.21	0.84	3.34	2.02	1.85
<i>t</i> -stat	8.90	4.41	2.72	8.77	4.46	3.81
<i>a</i> <sub>Carhart</sub>	2.00	0.95	0.70	2.39	1.46	1.50
<i>t</i> -stat	7.53	4.33	2.43	7.20	5.15	3.81
<i>a</i> <sub>FF5</sub>	2.66	0.76	0.73	3.67	1.52	1.72
<i>t</i> -stat	9.71	2.74	2.35	9.35	3.25	3.49
<i>a</i> <sub>q</sub>	2.40	0.70	0.67	2.72	1.28	1.84
<i>t</i> -stat	7.91	2.84	2.27	6.45	3.01	3.67
<i>a</i> <sub>q<sup>5</sup></sub>	2.32	0.57	0.63	2.34	1.19	1.74
<i>t</i> -stat	6.93	2.21	2.08	5.07	2.67	3.38
<b>Panel (b): All but tiny</b>						
<b>mean</b>	1.09	0.79	0.85	1.63	1.06	1.13
<i>t</i> -stat	5.74	3.43	3.19	5.24	2.65	2.33
<b>SR</b>	1.28	0.73	0.92	1.17	0.56	0.67
<i>a</i> <sub>CAPM</sub>	1.09	0.88	0.78	1.58	1.35	1.58
<i>t</i> -stat	5.63	3.85	2.82	4.99	3.55	3.23
<i>a</i> <sub>FF3</sub>	1.13	0.89	0.58	1.90	1.42	1.16
<i>t</i> -stat	5.69	3.88	2.23	6.25	3.85	2.66
<i>a</i> <sub>Carhart</sub>	0.67	0.66	0.45	0.94	0.97	0.83
<i>t</i> -stat	3.75	3.71	1.89	4.13	4.20	2.42
<i>a</i> <sub>FF5</sub>	1.30	0.52	0.53	2.07	1.01	1.14
<i>t</i> -stat	6.41	2.24	2.03	6.55	2.63	2.56
<i>a</i> <sub>q</sub>	0.97	0.48	0.56	1.17	0.76	1.15
<i>t</i> -stat	4.41	2.31	2.18	3.45	2.29	2.61
<i>a</i> <sub>q<sup>5</sup></sub>	0.92	0.34	0.51	0.80	0.56	0.98
<i>t</i> -stat	3.77	1.61	1.94	2.17	1.63	2.19
<b>Panel (c): Large stocks</b>						
<b>mean</b>	0.76	0.52	0.33	1.14	0.64	0.66
<i>t</i> -stat	3.95	2.20	1.30	3.52	1.61	1.30
<b>SR</b>	0.88	0.47	0.38	0.79	0.34	0.38
<i>a</i> <sub>CAPM</sub>	0.77	0.66	0.39	1.07	0.93	1.23
<i>t</i> -stat	3.91	2.86	1.48	3.25	2.45	2.44
<i>a</i> <sub>FF3</sub>	0.82	0.65	0.20	1.35	1.02	0.72
<i>t</i> -stat	4.10	2.87	0.81	4.19	2.92	1.71
<i>a</i> <sub>Carhart</sub>	0.40	0.43	0.09	0.32	0.61	0.36
<i>t</i> -stat	2.16	2.40	0.41	1.34	2.62	1.19
<i>a</i> <sub>FF5</sub>	0.98	0.30	0.13	1.48	0.62	0.67
<i>t</i> -stat	4.76	1.30	0.51	4.39	1.71	1.57
<i>a</i> <sub>q</sub>	0.66	0.28	0.23	0.62	0.46	0.76
<i>t</i> -stat	2.99	1.35	0.92	1.74	1.40	1.71
<i>a</i> <sub>q<sup>5</sup></sub>	0.55	0.10	0.12	0.18	0.16	0.47
<i>t</i> -stat	2.23	0.49	0.45	0.47	0.48	1.07

**Table 6-** Subsample Analysis

**Continued from previous page** - This table reports the performance of value-weighted long-short portfolios in different subsamples. The first three left columns show the results when the in-sample stocks are ranked based on the average of last year's returns ( $T = 12$  in equation 3), while the second three right columns show the counterpart results when the labels are based on the realized returns (according to equation 4). The results show different time horizons: Jan 1980 till Dec 1999, Jan 2000 till Dec 2021, and Jan 2010 till Dec 2021. Panel (a), (b) and (c) show the results for all, all but tiny and large stocks, respectively.

### *E. Momentum Anomalies*

The momentum factor is made up of a long-short portfolio that goes long (short) in portfolios with high (low) momentum. My long-short portfolios also go long (short) in high (low) decile portfolios. Especially, when the target variable is based on the average of the last 12 months, the portfolios might show a high correlation with the momentum factor. It is reasonable, then, to see if it can span the momentum factor or if they are just a proxy for the momentum factor. Also, I test if my long-short portfolios as a factor can explain momentum anomalies.

To begin, I consider a list of 41 momentum anomalies provided by [Hou, Xue, and Zhang \(2020\)](#). Table A6 in the Appendix describes these anomalies. I make long-short value-weighted portfolios from these momentum anomalies and regress them on my value-weighted long-short portfolio from all but tiny stocks when the target variable is the rank based on the realized returns. I treat this long-short portfolio as a Neighbouring factor (Neighb). This long-short portfolio, of course, is highly correlated with other long-short portfolios that I make (for example, it shows a correlation of 0.73 with another long-short portfolio when the target variable is the average returns for the last 12 months). With a critical value of  $|t| = 3.09$ , only 1 anomaly portfolio can generate a significant alpha when exposed to my factor. As shown in Table 7 panel (a), other methods generate more significant alphas. For  $|t| = 1.96$  only 7 alphas stand significant with my factor, while even a Carhart model cannot explain 14 of these momentum anomalies. Only the  $q^5$  model competes closely with my factor; still, the Neighb factor solely is better than the rest, including  $q$  and  $q^5$  which contain 4 and 5 factors. The average of absolute values of alphas for 41 momentum anomaly is 0.20 with an average absolute value of  $t = 1.03$ . Neighb factor has the lowest average absolute  $t$  values among all other factor models in explaining the anomalies similar to the five-factor model  $q^5$ . The next best is the four-factor  $q$  model. A Carhart model generates an average absolute  $t$  value of 1.53 and other methods (CAPM, FF3, FF5) produce average absolute values of  $t$  greater than 2.62.

Next, I regress my Neighb factor on the momentum factor. I consider three long-short portfolios as a proxy for the momentum factor. First, I collect the up-minus-down factor from Fama and French library which is the momentum factor. Second, I sort assets into 7 and 10 decile portfolios based on their momentum and create long-short portfolios from extreme deciles. Table 7 panel (b) shows the regression result. In fact, the momentum factor, as a tradable portfolio, no longer generates a profit when exposed to the Neighb portfolio. The intercept simply is -0.20% with  $t = 1.18$ , not distinguishable from zero and

with a negative sign. The Neighb factor coefficient, in this case, is 0.79 ( $t = 16.57$ ), meaning that the Neighb factor strongly explains the momentum factor. When using a 7 long-short momentum portfolio as a proxy for momentum factor, the alpha is still negative, namely -0.05 with  $t = -0.19$ . For a 10 long-short momentum portfolio, the alpha magnitude is even higher: -0.86 ( $t = -3.05$ ). To make sure that Neighb factor is not simply a proxy for the momentum factor, next I regress momentum factors on my long-short portfolio. Panel (c) shows the results. With momentum factors, 7 and 10 long-short momentum portfolios, Neighb portfolio generates a statistically significant positive alpha of 0.69, 0.64, and 0.81 with  $t$ -statistics of 5.65, 5.10, and 6.81, respectively. The result suggests that only one factor, which is a reflection of all characteristics at once, strongly explains a large number of anomalies as well as the momentum factor, while a momentum factor cannot explain it. In other words, my long-short portfolio spans the momentum factor. The results are similar for long-short portfolios constructed but all but 5% tiniest and large stocks (above 50% NYSE percentile).

#### *F. Characteristics Patterns*

The created portfolios through neighbouring assets are characteristic-managed portfolios because they are formed such that assets with similar characteristics are grouped together. Each portfolio has specific properties with respect to each characteristic. As decile portfolios fairly line up monotonically, we can attribute the set of obtained characteristics to the mean realized returns. Naturally, these portfolios are the right tools for studying the relationship between characteristics-mean return.

To begin with, considering all 94 characteristics, I find the average of characteristics in each portfolio. Then I find the time-series average of all characteristics from 1980-2021 for each decile portfolio. Then I rank the average of each characteristic from 1 to 10. Doing so gives a  $94 \times 10$  matrix of ranked characteristics for each decile portfolio. I group all characteristics into a few categories. First are those characteristics which increase with mean returns. For example, momentum is among these features. These features are almost linearly related to the expected returns. The second category is for those characteristics which first decrease and then increase, having a U-shape relationship with expected returns. Among 94 characteristics, I categorize 29 in this group. Another category which has an opposite behaviour is for those that first decrease on mean returns and then increase. These characteristics have an inverse-U shape relationship with mean returns. In my sample 22 out of 94 characteristics belong to this group. These two groups of characteristics seem to be related to the mean returns non-linearly and in the order of 2. The next category is the group of characteristics which first decrease and then increase, and again repeat decreasing and increasing. For example, for Industry sales concentration ([href](#)) the decile portfolio 1 has a mean standardized value of 0.01. For decile 2, it increases to 0.11, then decreases to -0.12 for decile 5. For decile 9 it reaches 0.14 and then decreases to 0.05 in decile 10. These characteristics have an M-shape and their relationship is from the order of 4. a small number of characteristics have an opposite shape which is a W-shape. Also, some characteristics have the highest values in decile 1 but decrease weakly for other

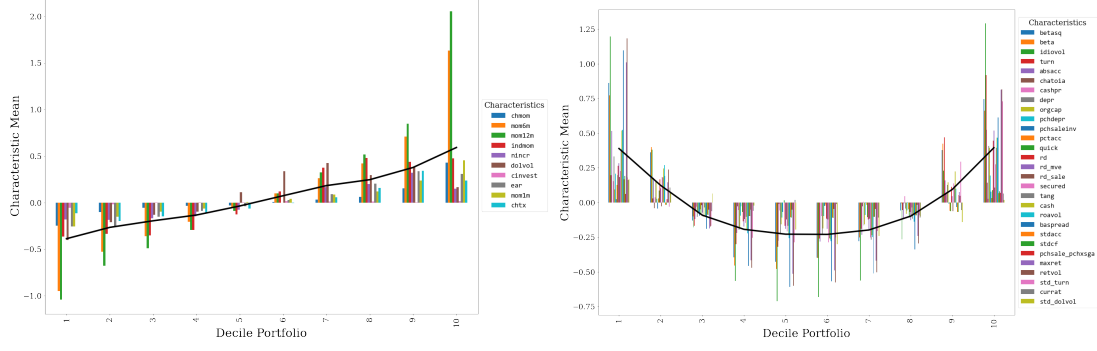
	avg $jaj$	avg $jtj$	num $jtj$	num $jtj$	num $jtj$	avg $jb_{\text{Neighb.}j}$	avg $jt(b_{\text{Neighb.}j})$
			1.94	2.58	3.09		
<b>Panel (a): Regressing momentum anomalies on different models</b>							
<b>Neighb.</b>	0.20	1.03	7	1	1	0.44	9.28
<b>Neighb. + Market</b>	0.21	1.04	6	2	1	0.29	7.82
<b>CAPM</b>	0.49	2.78	33	23	15		
<b>FF3</b>	0.57	3.35	39	30	25		
<b>Carhart</b>	0.22	1.53	14	7	4		
<b>FF5</b>	0.46	2.62	32	20	11		
<b><math>q</math> model</b>	0.19	1.17	8	5	3		
<b><math>q^5</math> model</b>	0.18	1.03	6	2	1		
<b>Panel (b): Regressing momentum factors on Neighbouring Assets long-short factor</b>							
Regressand	alpha	$t$ stat				$b_{\text{Neighb.}}$	$t(b_{\text{Neighb.}})$
Momentum factor	-0.20	-1.18				0.79	16.57
7 LS Mom Portfolio	-0.05	-0.19				1.17	15.83
10 LS Mom Portfolio	-0.86	-3.05				1.39	17.39
<b>Panel (c): Regressing Neighbouring Assets long-short factor on momentum factors</b>							
Regressor	alpha	$t$ stat				$b_{\text{MOM}}$	$t(b_{\text{MOM}})$
Momentum factor	0.69	5.65				0.45	16.57
7 LS Mom Portfolio	0.64	5.10				0.28	15.83
10 LS Mom Portfolio	0.81	6.81				0.27	17.39

**Table 7-** Performance of neighbouring assets long-short portfolio as a factor on momentum anomalies

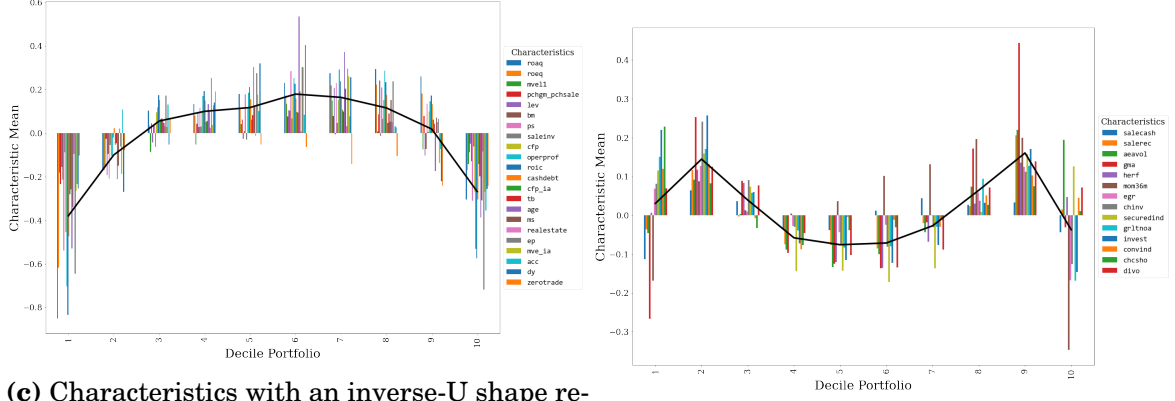
This table shows the regression results when a long-short portfolio formed from neighbouring assets is used as a tradeable factor (Neighb factor). I collect 41 momentum anomalies provided by Hou et al. (2020) and check if my Neighb factor can explain a long-short strategy from those anomalies. Panel (a) shows the results of time-series regression of Neighb factor as well as other factor models on momentum anomalies. In panel (b), I regress momentum factors on my Neighb factor. I consider three different portfolios as a proxy for momentum anomalies: (i) momentum factor from Fama and French data library, (ii) a long-short portfolio from 7 portfolios sorted on the momentum (7 LS Mom Portfolio) and (iii) a long-short portfolio from decile portfolios sorted on the momentum (10 LS Mom Portfolio). Panel (c) shows the regression results when these momentum factors are regressed on the Neighb Portfolio. I use a value-weighted long-short portfolio from all but tiny stocks with a stochastic selection with  $\rho_{\text{all but tiny}} = 0.02$  with a target variable based on the average of last year returns ( $T = 12$  in equation 3) as Neighb factor.

portfolios. I group these characteristics, though weakly, as a descending group. Lastly, for some characteristics, their patterns are not as strong as previous ones, and I categorise them in others.

I show the first four groups of characteristics in Figure 6. Panel (a) shows the characteristics which are linearly and monotonically related to the expected returns. Panel (b) and (c) plot those characteristics which have a U-shape or an inverse-U-shape relationship with mean returns. Panel (d) contains characteristics with an M-shape. The black line demonstrates the average of these characteristics. Figure 6 confirms the non-linearity of characteristics-expected returns relationship. This figure shows the average of characteristics and portfolios when all stocks except the smallest 5% are used for creating the



(a) Characteristics with an ascending relationship (b) Characteristics with a U-shape relationship



(c) Characteristics with an inverse-U shape relationship (d) Characteristics with a M-shape relationship

**Figure 6.** Relationship between characteristics and mean realized returns in 1980-2021

This figure shows the pattern of characteristics and realized returns. Panel (a) shows characteristics which monotonically and linearly increase with the decile portfolio. Panel (b) shows characteristics which first decrease and then increase with the decile portfolio, yielding to have a U-shape relationship with realized returns. Panel (c) illustrates characteristics with an inverse-U shape with mean returns. The relationship between characteristics and mean returns are in the order of 2 in panels (b) and (c). Panel (d) shows the characteristics which first increase, then decrease, and then again increase and decrease, having an M-shape pattern. In this case, the relationship of the characteristics-mean return is from the order of 4. The black line shows the average of bars in a decile portfolio. The x-axis shows the decile portfolios, while the y-axis shows the mean of characteristics. The characteristics are from portfolios containing all but the tiniest 5% stocks. The list of characteristics is defined in Table A1. Before creating portfolios, all characteristics are cross-sectionally standardized and winsorised at 0.01 level. I use value-weighted long-short portfolios from all stocks and target variables based on the average of last year's returns ( $T = 12$  in equation 3) to study the characteristics patterns.

portfolios. The patterns also persist strongly among all but tiny and large stocks. However, these relationships are stronger when the universe of data includes all.

Neither size ( $mvel$  1) nor book-to-market ( $bm$ ) seem to have a linear relationship with mean realized returns, and this is why a model containing factors based on these characteristics fails to explain the portfolios created by these characteristics in a non-linear setting. I confirm Kozak (2020)'s argument that while Fama and French (2016) link the market-to-book  $\frac{M}{B}$  to expected earnings  $E(Y)$ , investment  $DB$ , and discount rate in a linear way as

$$\frac{M_t}{B_t} = \frac{1}{B_t} \sum_{s=1}^{\infty} \frac{E(Y_{t+s} B_{t+s})}{(1+r)^s}, \quad (18)$$

the discount rate itself is a function of other characteristics and book-to-market itself. It

is not, therefore, determined independently, and hence the equation above links book-to-market in a non-linear way to the expected returns. Moreover, based on the equation 18 these characteristics have interdependencies. Therefore, including them at once provides a wealth of information while creating portfolios.

### G. Robustness Checks

In this section, I show that the results are not sensitive to the neighbourhood definition, nor to the distance measurement. I show that the results are robust to different rolling years embedded in the in-sample set. I also investigate the role of  $\rho$  which determines the in-sample set size. I create portfolios for the case when the target variable is  $\frac{1}{12} \hat{\alpha}_{t=0}^{11} r_t$ , but the results are similar for the case that the target variable is  $r_t$ .

#### G.1. Different Number of Neighbours

I start with checking how the results are sensitive to the number of neighbours considered for classifying assets into portfolios. For robustness checks, I hold everything else constant and change  $k$ , and report the dispersion produced by value-weighted decile portfolios 1 and 10. I repeat the analysis for three sets of data in Table 8. I change  $k$  from 100 to 5000 and report the performance of a value-weighted long-short portfolio. In order to just focus on  $k$ , I fix everything else, including  $\rho$  for all three datasets on 10% to make sure that the number of neighbours  $k$  is less than the size of the in-sample set,  $|S_{tj}|$ , for all  $k$  in all data set and for each cross-section. When considering all but 5% tiniest stocks, the average realized return of a long-short portfolio does not go below 2.06%. The performance of the model starts increasing by increasing  $k$ , and it reaches its maximum around between  $k = 900$  and  $k = 1200$ . The annualized Sharpe ratio is 1.19 for  $k = 900$ . In all  $k$  from 100 to 5000, all the risk-adjusted alphas in six different factor models are both statistically and economically different from zero. Needless to say, the closer assets have the highest weights and they are most likely to determine the label class. It is not surprising that the performance does not significantly change when I still use a distance weighting scheme. However, the further assets also contain information relevant to the classification problem. This can be confirmed from panel (b) and (c) when the performance of long-short portfolios from all but tiny and large stocks increase by increasing the  $k$ . However, for  $k \geq 2000$  the Sharpe ratios drop for all three datasets, and the extra profitability comes from the short legs. In fact for panel (a), the average return of the long portfolio is 1.39% (std = 0.11) with an average  $t$  statistics of 3.16. The average long-portfolios for panel (b) is 1.22% (std = 0.05) with an average of  $t = 3.05$ . For large stocks in panel (c), the long-legs generate on average 1.28% (std = 0.09) with an average  $t$  of 3.30. Overall the results are suggestive that the clustering effects are not sensitive to the neighbourhood definition and persist strongly over all three data sets.

$k$	100	300	700	900	1200	1500	2000	2500	5000
<b>Panel (a): All data</b>									
<b>mean</b>	2.06	2.31	2.40	2.48	2.48	2.30	2.35	2.33	2.29
<i>t</i> -stat	6.55	7.02	7.25	7.69	7.55	7.25	7.21	6.79	6.02
<b>SR</b>	1.01	1.08	1.12	1.19	1.17	1.12	1.11	1.05	0.93
$\alpha_{\text{CAPM}}$	2.20	2.40	2.47	2.54	2.52	2.36	2.38	2.34	2.32
<i>t</i> -stat	6.96	7.22	7.34	7.78	7.59	7.33	7.20	6.75	6.01
$\alpha_{\text{FF3}}$	2.41	2.61	2.67	2.75	2.73	2.55	2.58	2.56	2.57
<i>t</i> -stat	7.89	8.12	8.22	8.72	8.51	8.19	8.04	7.66	6.96
$\alpha_{\text{Carhart}}$	1.50	1.68	1.76	1.88	1.85	1.73	1.76	1.73	1.74
<i>t</i> -stat	7.03	7.28	7.30	7.91	7.63	7.14	6.89	6.39	5.54
$\alpha_{\text{FF5}}$	2.23	2.40	2.49	2.58	2.56	2.35	2.42	2.33	2.36
<i>t</i> -stat	7.03	7.19	7.35	7.85	7.66	7.27	7.24	6.70	6.14
$\alpha_q$	1.67	1.86	1.93	2.06	2.00	1.78	1.82	1.74	1.82
<i>t</i> -stat	5.51	5.80	5.94	6.48	6.25	5.74	5.64	5.12	4.75
$\alpha_{q^c}$	1.41	1.57	1.72	1.81	1.77	1.47	1.48	1.42	1.48
<i>t</i> -stat	4.38	4.59	4.96	5.35	5.19	4.46	4.30	3.93	3.63
<b>Panel (b): All but tiny stocks</b>									
<b>mean</b>	1.04	1.35	1.59	1.63	1.70	1.73	1.73	1.77	1.91
<i>t</i> -stat	4.09	5.39	6.21	6.19	6.20	6.24	5.97	5.82	5.90
<b>SR</b>	0.63	0.83	0.96	0.96	0.96	0.96	0.92	0.90	0.91
$\alpha_{\text{CAPM}}$	1.20	1.49	1.73	1.77	1.84	1.86	1.87	1.92	2.05
<i>t</i> -stat	4.74	5.93	6.74	6.71	6.69	6.67	6.42	6.28	6.26
$\alpha_{\text{FF3}}$	1.37	1.66	1.89	1.94	2.03	2.05	2.09	2.14	2.28
<i>t</i> -stat	5.65	6.91	7.64	7.65	7.71	7.66	7.51	7.31	7.34
$\alpha_{\text{Carhart}}$	0.62	0.96	1.17	1.23	1.30	1.32	1.34	1.37	1.57
<i>t</i> -stat	3.83	5.59	6.60	6.54	6.60	6.51	6.29	6.01	5.99
$\alpha_{\text{FF5}}$	1.14	1.50	1.77	1.85	1.91	1.96	1.97	2.02	2.18
<i>t</i> -stat	4.55	6.02	6.86	6.98	6.97	7.01	6.80	6.63	6.73
$\alpha_q$	0.71	1.04	1.31	1.41	1.48	1.51	1.51	1.54	1.75
<i>t</i> -stat	3.03	4.45	5.39	5.50	5.54	5.56	5.33	5.20	5.41
$\alpha_{q^c}$	0.44	0.75	1.06	1.09	1.14	1.19	1.17	1.20	1.42
<i>t</i> -stat	1.75	3.00	4.10	4.01	4.05	4.14	3.90	3.81	4.14
<b>Panel (c): Large stocks</b>									
<b>mean</b>	0.97	1.06	1.24	1.29	1.36	1.51	1.63	1.60	1.99
<i>t</i> -stat	3.89	4.04	4.30	4.35	4.44	4.69	4.86	4.48	4.37
<b>SR</b>	0.60	0.62	0.66	0.67	0.68	0.72	0.75	0.69	0.67
$\alpha_{\text{CAPM}}$	1.13	1.21	1.36	1.44	1.48	1.67	1.79	1.79	2.05
<i>t</i> -stat	4.52	4.59	4.71	4.84	4.76	5.15	5.31	4.99	4.43
$\alpha_{\text{FF3}}$	1.34	1.42	1.57	1.64	1.68	1.86	2.02	2.02	2.29
<i>t</i> -stat	5.71	5.66	5.66	5.73	5.63	5.96	6.24	5.82	5.07
$\alpha_{\text{Carhart}}$	0.65	0.71	0.88	0.96	0.96	1.13	1.27	1.33	1.58
<i>t</i> -stat	3.92	3.85	3.92	4.06	3.92	4.33	4.68	4.30	3.72
$\alpha_{\text{FF5}}$	1.21	1.31	1.50	1.56	1.56	1.77	1.92	1.97	2.10
<i>t</i> -stat	4.96	5.03	5.19	5.25	5.02	5.43	5.70	5.43	4.46
$\alpha_q$	0.77	0.84	1.04	1.09	1.07	1.25	1.38	1.55	1.63
<i>t</i> -stat	3.24	3.37	3.71	3.76	3.54	3.92	4.18	4.33	3.43
$\alpha_{q^c}$	0.45	0.48	0.73	0.75	0.77	0.95	0.99	0.98	1.17
<i>t</i> -stat	1.81	1.82	2.45	2.44	2.38	2.79	2.81	2.61	2.32

**Table 8-** Performance of a value-weighted long-short portfolio for different values of  $k$

**Continued from previous page** - This table reports performance of a value-weighted long-short portfolio formed based on different number of neighbours  $k$  with a target variable based on the average of last year returns ( $T = 12$  in equation 3). All 94 characteristics are taken into account while creating portfolios. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The regression periods captures realized returns in percentage from 1980 to 2021. In order to keep everything else constant rather than  $k$ , I fix  $\rho$  large enough so that  $k$  is always less than size of training sample,  $|S_{tj}|$ . So  $\rho_{\text{large}} = \rho_{\text{all}}$  but  $\rho_{\text{tiny}} = \rho_{\text{all}} = 0.1$ . The number of months included in the in-sample data is considered  $t = 120$ . Panel (a) reports the results for all but 5% tiniest stocks, panel (b) shows the counterpart results for all but tiny stocks, and panel (c) represents the results for large stocks.

## G.2. Different Rolling Windows in the In-sample Set

Naturally, the choice of  $t$ , the number of months included in the training sample, could potentially affect the results because of the long-term dependency structure of the time-series data. In my framework, I assume that not merely last month, but observations from several years ago are also relevant for classifying a new asset. Including previous years in the in-sample data helps capture this interdependency structure of stock returns. The main analysis is based on 10 years ( $t = 120$  months) rolling window. In this section, I repeat the calculations from 2 to 14 years of rolling windows and show that the results do not change significantly.

Table 9 shows the performance of value-weighted long-short portfolios for different  $t$  from 1980 to 2021. To keep everything else constant, again, I fix  $\rho = 0.1$  for all three datasets to make sure that there are enough neighbours in each cross-section. I set  $k = 1000$  as the main analysis. In panel (a) which contains all but 5% tiniest stocks, for all columns, the mean is over 2.06%, and the Sharpe ratio goes over 1 when I consider  $t$  greater than 4 years (48 months). All models produce a significant alpha. Panel (b) shows the counterpart results for all but tiny stocks. Again all risk-adjusted alphas are significantly greater than zero. Large stocks are shown in panel (c) that produce significant alphas as well. The results from the whole table suggest that the model performance is not sensitive to the rolling window considered in the in-sample set, and the clustering effect persists even in shorter horizons. The portfolios are well diversified.

## G.3. Different Distance Measures

In this section, I show that the results are not sensitive to distance measurement. Basically, different distance measurements must lead to almost the same neighbours. The distance measurement I use for the whole analysis is based on a Euclidean distance, which is  $l^2$ -norm, and then I use a distance weighting scheme. For a robustness check, I consider a uniform distance weighting such that all neighbouring assets receive equal weights. In other words,  $w(a_{s,t} \neq j|a_{j,t})$  in equation 7 will be equal across all assets (for  $s = 1, \dots, k$ ) in the neighbourhood of asset  $a_{j,t}$ . Next, I deviate from a  $l^2$ -norm measurement to a  $l$ -norm. I show that the patterns persist among different definitions of distance and uniformly weighting. Table 10 shows the performance of value-weighted long-short portfolios based

$t$	24	36	48	96	132	144	156	168
<b>Panel (a): All data</b>								
<b>mean</b>	2.31	2.06	2.30	2.21	2.25	2.16	2.29	2.15
$t$ -stat	5.90	5.47	6.46	6.63	6.97	6.57	7.24	6.67
<b>SR</b>	0.91	0.84	1.00	1.02	1.08	1.01	1.12	1.03
$a_{\text{CAPM}}$	2.28	2.13	2.34	2.28	2.30	2.23	2.34	2.23
$t$ -stat	5.74	5.58	6.49	6.76	7.02	6.69	7.29	6.83
$a_{\text{FF3}}$	2.56	2.39	2.60	2.51	2.49	2.44	2.53	2.43
$t$ -stat	6.76	6.56	7.56	7.78	7.80	7.61	8.15	7.67
$a_{\text{Carhart}}$	1.72	1.49	1.74	1.65	1.64	1.56	1.71	1.55
$t$ -stat	5.31	5.04	6.28	6.63	6.64	6.44	7.09	6.54
$a_{\text{FF5}}$	2.38	2.26	2.42	2.40	2.32	2.30	2.35	2.24
$t$ -stat	6.05	5.95	6.75	7.13	6.99	6.89	7.30	6.80
$a_q$	1.74	1.59	1.79	1.82	1.82	1.72	1.84	1.75
$t$ -stat	4.49	4.37	5.12	5.61	5.64	5.34	5.94	5.46
$a_{q^5}$	1.42	1.23	1.53	1.55	1.60	1.48	1.69	1.57
$t$ -stat	3.44	3.18	4.10	4.49	4.67	4.28	5.09	4.57
<b>Panel (b): All but tiny stocks</b>								
<b>mean</b>	2.04	1.69	1.89	1.66	1.50	1.48	1.45	1.50
$t$ -stat	5.31	4.98	6.23	5.88	5.52	5.49	5.36	5.83
<b>SR</b>	0.82	0.77	0.96	0.91	0.85	0.85	0.83	0.90
$a_{\text{CAPM}}$	2.13	1.82	2.08	1.81	1.63	1.66	1.59	1.65
$t$ -stat	5.50	5.31	6.85	6.39	5.96	6.15	5.85	6.40
$a_{\text{FF3}}$	2.40	2.12	2.32	2.01	1.81	1.84	1.77	1.82
$t$ -stat	6.49	6.73	8.10	7.40	6.90	7.10	6.79	7.30
$a_{\text{Carhart}}$	1.63	1.41	1.57	1.24	1.06	1.10	1.02	1.12
$t$ -stat	5.03	5.27	7.02	6.20	5.54	5.81	5.41	6.06
$a_{\text{FF5}}$	2.23	2.03	2.22	1.87	1.70	1.77	1.68	1.72
$t$ -stat	5.80	6.17	7.44	6.63	6.22	6.54	6.18	6.62
$a_q$	1.74	1.58	1.71	1.40	1.23	1.29	1.16	1.26
$t$ -stat	4.57	4.78	5.85	5.08	4.66	4.93	4.52	5.03
$a_{q^5}$	1.34	1.08	1.31	1.08	0.92	1.03	0.88	0.93
$t$ -stat	3.32	3.09	4.25	3.70	3.31	3.72	3.23	3.49
<b>Panel (c): Large stocks</b>								
<b>mean</b>	2.14	1.43	1.32	1.31	1.42	1.26	1.36	1.25
$t$ -stat	4.29	3.51	3.58	4.32	5.03	4.44	4.78	4.56
<b>SR</b>	0.66	0.54	0.55	0.67	0.78	0.68	0.74	0.70
$a_{\text{CAPM}}$	2.24	1.52	1.49	1.46	1.59	1.40	1.53	1.41
$t$ -stat	4.44	3.69	3.99	4.79	5.62	4.88	5.36	5.14
$a_{\text{FF3}}$	2.50	1.76	1.70	1.69	1.78	1.60	1.71	1.59
$t$ -stat	5.05	4.38	4.70	5.83	6.58	5.83	6.26	5.99
$a_{\text{Carhart}}$	1.71	1.04	0.97	0.95	1.08	0.89	1.02	0.93
$t$ -stat	3.68	2.83	3.03	4.12	5.06	4.10	4.66	4.34
$a_{\text{FF5}}$	2.54	1.53	1.55	1.59	1.69	1.46	1.59	1.49
$t$ -stat	4.93	3.67	4.12	5.25	5.99	5.09	5.58	5.37
$a_q$	2.01	1.12	1.18	1.13	1.25	0.99	1.14	1.09
$t$ -stat	3.93	2.71	3.11	3.83	4.57	3.57	4.13	3.97
$a_{q^5}$	1.27	0.90	0.60	0.71	0.94	0.69	0.80	0.76
$t$ -stat	2.35	2.02	1.51	2.28	3.23	2.33	2.73	2.61

**Table 9-** Performance of a value-weighted long-short portfolio for different values of  $t$

**Continued from previous page** - This table reports performance of a value-weighted long-short portfolio formed based on different length of rolling window,  $t$ , with a target variable based on the average of last year returns ( $T = 12$  in equation 3). All 94 characteristics are taken into account while creating portfolios. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The regression periods captures realized returns in percentage from 1980 to 2021. In order to keep everything else constant rather than  $k$ , I fix  $\rho$  large enough so that  $k$  is always less than size of training sample,  $|S_t|$ . So  $\rho_{\text{large}} = \rho_{\text{all but tiny}} = \rho_{\text{all}} = 0.1$ . The number of months included in the in-sample data is considered  $t = 120$ . Panel (a) reports the results for all but 5% tiniest stocks, panel (b) shows the counterpart results for all but tiny stocks, and panel (c) represents the results for large stocks.

on both uniform weighting and  $l$ -norm distance. In all of the portfolios, I employ the entire past data ( $\rho = 1$ ) for the in-sample set. Panel (a) shows the case where the target variable is based on the realized returns at each month. For all data, when uniform weighting the average return of a long-short portfolio is 1.73% ( $t = 9.08$ ) with a Sharpe ratio of 1.40. The FF3 alpha is 1.78% with  $t = 9.21$ . Similarly, in panel (b) the FF3 alpha is 2.69% ( $t = 8.81$ ). The results are suggestive that the model performance is not sensitive to changing the distance metrics.

#### G.4. Different size for the training sample, $\rho$

In my framework,  $\rho$  plays a role in defining the size of in-sample data. Maybe one drawback of a  $k$ NN is the high computational cost which requires the calculation of pairwise distances. By shrinking the training sample, the computational cost significantly decreases. More importantly, the outliers and noisier data are more likely to be removed. In this section, I show that how changing  $\rho$  can affect the results. All in all, I show that the results are fairly robust with respect to different choices of  $\rho$ .

For studying the effect of stochastic selection, I repeat the main analysis with  $\rho = 0.5, 0.25$  and  $0.1$ , that is, to stochastically select assets from in-sample data such that each asset has a probability of  $\rho$  to be selected. I keep  $k = 1000$  and  $t = 120$  months and focus on a value-weighted long-short portfolio which contains all 94 characteristics. The performance of this portfolio is presented in Table 11. For all data, the performance of a long-short portfolio roughly remains the same, with a Sharpe ratio changing from 1.10 to 1.12. For all but tiny and large stocks, the performance of the model increases by decreasing  $\rho$ , suggesting that the model becomes more robust to label the noise after pruning. For large stock the average monthly returns increases to 1.29% ( $t = 4.45$ ) when  $\rho = 0.1$ . In this case, a long-short portfolio of large stocks generates significant alphas with respect to all considered factor models.

	All data		All but tiny stocks		Large stocks	
	uniform	$l$ -norm	uniform	$l$ -norm	uniform	$l$ -norm
<b>Panel (a): Predicting <math>r_z</math></b>						
<b>mean</b>	1.73	1.56	0.93	0.87	0.63	0.73
$t$ -stat	9.08	7.84	6.38	5.41	4.28	4.30
<b>SR</b>	1.40	1.21	0.98	0.84	0.66	0.66
$a_{\text{CAPM}}$	1.72	1.58	0.98	0.96	0.72	0.82
$t$ -stat	8.89	7.83	6.63	5.93	4.85	4.81
$a_{\text{FF3}}$	1.78	1.64	1.00	1.00	0.76	0.90
$t$ -stat	9.21	8.15	6.74	6.19	5.15	5.35
$a_{\text{Carhart}}$	1.37	1.22	0.66	0.58	0.41	0.48
$t$ -stat	8.13	6.92	5.27	4.56	3.36	3.54
$a_{\text{FF5}}$	1.56	1.43	0.86	0.80	0.63	0.77
$t$ -stat	7.89	6.89	5.61	4.84	4.14	4.42
$a_q$	1.34	1.23	0.70	0.60	0.48	0.59
$t$ -stat	7.05	6.12	4.74	3.78	3.27	3.50
$a_{q^5}$	1.21	1.19	0.57	0.43	0.32	0.30
$t$ -stat	5.94	5.52	3.60	2.54	2.04	1.71
<b>Panel (b): Predicting <math>E(r)</math></b>						
<b>mean</b>	2.34	2.24	1.33	1.29	0.90	0.92
$t$ -stat	7.42	6.84	5.23	4.71	3.57	3.38
<b>SR</b>	1.14	1.05	0.81	0.73	0.55	0.52
$a_{\text{CAPM}}$	2.47	2.39	1.50	1.46	1.06	1.10
$t$ -stat	7.79	7.26	5.91	5.35	4.18	4.06
$a_{\text{FF3}}$	2.69	2.61	1.69	1.65	1.27	1.34
$t$ -stat	8.81	8.27	6.97	6.29	5.39	5.30
$a_{\text{Carhart}}$	1.80	1.67	0.94	0.82	0.58	0.57
$t$ -stat	8.22	7.56	5.81	4.82	3.45	3.30
$a_{\text{FF5}}$	2.50	2.39	1.52	1.46	1.11	1.16
$t$ -stat	7.89	7.30	6.07	5.39	4.54	4.45
$a_q$	1.95	1.81	1.06	0.91	0.72	0.69
$t$ -stat	6.41	5.84	4.48	3.67	3.03	2.77
$a_{q^5}$	1.69	1.55	0.78	0.61	0.32	0.24
$t$ -stat	5.22	4.70	3.11	2.32	1.30	0.91

**Table 10-** Performance of a value-weighted long-short portfolio for different distance measures

This table reports the performance of value-weighted long-short portfolios created with uniform weighting scheme and distance weighting with  $l$ -norm distance metrics. The regression period is from Jan 1980 till Dec 2021. The uniform column shows the case when I use equal weights for neighbouring assets when making the prediction. (All the portfolios are still value-weighted with market-cap).  $l$  norm shows the case when I use absolute-value norm (instead of Euclidean norm) to find  $k$  closest assets. Panel (a) shows the case where the target variable is the asset's ranks based on the realized returns, while in panel (b) the target variable is defined based on the average of last year returns ( $T = 12$  in equation 3).

$\rho$	All data			All but tiny stocks			Large stocks		
	0.5	0.25	0.1	0.5	0.25	0.1	0.5	0.25	0.1
<b>mean</b>	2.24	2.31	2.36	1.41	1.35	1.62	1.02	1.10	1.36
<b>std</b>	7.18	7.20	7.33	5.94	5.80	6.13	5.93	5.89	6.57
<i>t</i> -stat	7.00	7.19	7.21	5.32	5.21	5.94	3.88	4.17	4.64
<b>SR</b>	1.08	1.11	1.11	0.82	0.80	0.92	0.60	0.64	0.72
$a_{\text{CAPM}}$	2.35	2.38	2.44	1.57	1.50	1.76	1.20	1.22	1.50
<i>t</i> -stat	7.28	7.32	7.37	5.95	5.78	6.42	4.55	4.61	5.10
$a_{\text{FF3}}$	2.56	2.58	2.64	1.76	1.68	1.95	1.42	1.42	1.72
<i>t</i> -stat	8.22	8.21	8.22	7.01	6.76	7.45	5.76	5.68	6.18
$a_{\text{Carhart}}$	1.66	1.68	1.75	1.00	0.94	1.24	0.70	0.74	1.03
<i>t</i> -stat	7.38	7.33	7.27	5.80	5.40	6.22	3.98	3.88	4.56
$a_{\text{FF5}}$	2.36	2.39	2.46	1.61	1.54	1.85	1.31	1.34	1.66
<i>t</i> -stat	7.29	7.31	7.38	6.14	5.96	6.78	5.11	5.14	5.71
$a_q$	1.84	1.81	1.89	1.10	1.05	1.40	0.82	0.86	1.24
<i>t</i> -stat	5.88	5.80	5.89	4.48	4.31	5.32	3.41	3.46	4.28
$a_{q^s}$	1.60	1.59	1.73	0.81	0.70	1.05	0.49	0.57	0.87
<i>t</i> -stat	4.79	4.78	5.05	3.10	2.72	3.77	1.94	2.15	2.82

**Table 11-** Performance of value-weighted long-short portfolios for different values of  $\rho$

This table reports the performance of value-weighted long-short portfolios, for different  $\rho$ , in the period 1980-2021 for all but tiniest 5%, all tiny stocks which have a market cap below 20% NYSE percentile, and large stocks (above NYSE median). All 94 characteristics are taken into account for creating portfolios.  $\rho = 0.5$  indicates that the in-sample set is shrunk to half stochastically. Similarly,  $\rho = 0.25$  and  $0.1$  are representative of the case when the in-sample is shrunk and only 25% and 10% of in-sample data are used for finding the neighbouring assets. The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. Target variables are defined with a target variable based on the average of last year returns ( $T = 12$  in equation 3).

### III. Summary and Conclusion

If two firms are alike in terms of characteristics, it is likely that they display similar expected rates of returns. The similarity of firms in the characteristics is an indicator of fundamental linkages. By considering a large set of characteristics, for each firm, I recognize the firms that most resemble it. Identifying them as neighbouring firms (assets), I find that neighbouring assets tend to produce similar expected returns over time. This simply is implied by assuming that firm-level characteristics determine expected returns. Hence, because each asset has a fundamental connection with its neighbours in many aspects, each asset should have similar expected returns as its neighbours.

My results, therefore, strongly show that past returns of each asset's neighbours are a strong predictor of its future returns. I borrow the so-called  $k$ -nearest neighbours classifier from machine learning to find assets with the most similarity. Grouping assets to decile portfolios based on the past performance of their neighbours can generate a high dispersion in the cross-sections of returns. These portfolios are indicative of a bunch of characteristics and their expected returns reflect properties of characteristics. Not surprisingly, common factor models cannot explain the behaviour of these portfolios. First, these factors are designed in a linear setting, while the relationship between returns and characteristics can be complex. Second, these factors are based on a few characteristics while neighbouring stocks use a large number of characteristics to predict the future expected returns. The out-of-sample performance of neighbouring stocks is remarkable in a long run and across different types of data.

Time-series and cross-sectional features of stock returns contain massive amounts of information that asset pricing models try to find and explain. In this paper, I map a large set of firm characteristics to their expected returns in a non-linear setting. This sheds light on the relationship between firm characteristics and expected returns. My method also alleviates the curse of dimensionality implied by the zoo of characteristics. I propose a novel way of forming test portfolios by grouping neighbouring assets to decile portfolios. Finally, I develop a method to unify the joint effect of characteristics between connected firms.

## References

- Ahn, Dong-Hyun, Jennifer Conrad, and Robert F Dittmar, 2009, Basis assets, *The Review of Financial Studies* 22, 5133–5174.
- Ali, Usman, and David Hirshleifer, 2020, Shared analyst coverage: Unifying momentum spillover effects, *Journal of Financial Economics* 136, 649–675.
- Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *The Journal of Finance* 61, 259–299.
- Anton, Miguel, and Christopher Polk, 2014, Connected stocks, *The Journal of Finance* 69, 1099–1127.
- Bali, Turan G, Nusret Cakici, and Robert F Whitelaw, 2011, Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of financial economics* 99, 427–446.
- Barbee Jr, William C, Sandip Mukherji, and Gary A Raines, 1996, Do sales–price and debt–equity explain stock returns better than book–market and firm size?, *Financial Analysts Journal* 52, 56–60.
- Barberis, Nicholas, and Andrei Shleifer, 2003, Style investing, *Journal of financial Economics* 68, 161–199.
- Bouwman, Christa HS, 2011, Corporate governance propagation through overlapping directors, *The Review of Financial Studies* 24, 2358–2394.
- Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2022, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *The Journal of Finance* .
- Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, 2020, Forest through the trees: Building cross-sections of stock returns, *Available at SSRN 3493458* .
- Chen, Andrew Y, and Tom Zimmermann, 2021, Open source cross-sectional asset pricing, *Critical Finance Review, Forthcoming* .
- Chen, Luyang, Markus Pelger, and Jason Zhu, 2019, Deep learning in asset pricing, *arXiv preprint arXiv:1904.00745* .

- Chordia, Tarun, Avanidhar Subrahmanyam, and V Ravi Anshuman, 2001, Trading activity and expected stock returns, *Journal of financial Economics* 59, 3–32.
- Cochrane, John H, 2011, Presidential address: Discount rates, *The Journal of finance* 66, 1047–1108.
- Cohen, Lauren, and Andrea Frazzini, 2008, Economic links and predictable returns, *The Journal of Finance* 63, 1977–2011.
- Cover, Thomas, and Peter Hart, 1967, Nearest neighbor pattern classification, *IEEE transactions on information theory* 13, 21–27.
- Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *The Journal of finance* 52, 1035–1058.
- Datar, Vinay T, Narayan Y Naik, and Robert Radcliffe, 1998, Liquidity and stock returns: An alternative test, *Journal of financial markets* 1, 203–219.
- Fama, Eugene F, and Kenneth R French, 1995, Size and book-to-market factors in earnings and returns, *The journal of finance* 50, 131–155.
- Fama, Eugene F, and Kenneth R French, 2008, Dissecting anomalies, *The Journal of Finance* 63, 1653–1678.
- Fama, Eugene F, and Kenneth R French, 2016, Dissecting anomalies with a five-factor model, *The Review of Financial Studies* 29, 69–103.
- Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *The Journal of Finance* 75, 1327–1370.
- Feng, Guanhao, Jingyu He, and Nicholas G Polson, 2018a, Deep learning for predicting asset returns, *arXiv preprint arXiv:1804.09314* .
- Feng, Guanhao, Nicholas G Polson, and Jianeng Xu, 2018b, Deep learning in characteristics-sorted factor models, *arXiv preprint arXiv:1805.01104* .
- Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting characteristics nonparametrically, *The Review of Financial Studies* 33, 2326–2377.

- Gettleman, Eric, and Joseph M Marks, 2006, Acceleration strategies, *SSRN Electronic Journal*.
- Giglio, Stefano, Yuan Liao, and Dacheng Xiu, 2021, Thousands of alpha tests, *The Review of Financial Studies* 34, 3456–3496.
- Goyal, Amit, and Sunil Wahal, 2015, Is momentum an echo?, *Journal of Financial and Quantitative Analysis* 50, 1237–1267.
- Green, Jeremiah, J Hand, and Frank Zhang, 2014, The remarkable multidimensionality in the cross-section of expected us stock returns, *Available at SSRN* 2262374.
- Green, Jeremiah, John RM Hand, and Mark T Soliman, 2011, Going, going, gone? the apparent demise of the accruals anomaly, *Management Science* 57, 797–816.
- Green, Jeremiah, John RM Hand, and X Frank Zhang, 2017, The characteristics that provide independent information about average us monthly stock returns, *The Review of Financial Studies* 30, 4389–4436.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2021, Autoencoder asset pricing models, *Journal of Econometrics* 222, 429–450.
- Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, ... and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.
- Haugen, Robert A, and Nardin L Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of financial economics* 41, 401–439.
- He, Wei, Yuehan Wang, and Jianfeng Yu, 2021, Similar stocks, *Available at SSRN* 3815595 .
- Hinton, Geoffrey E, and Sam Roweis, 2002, Stochastic neighbor embedding, *Advances in neural information processing systems* 15.
- Hirshleifer, David, Kewei Hou, and Siew Hong Teoh, 2012, The accrual anomaly: risk or mispricing?, *Management Science* 58, 320–335.

- Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2019, Which factors?, *Review of Finance* 23, 1–35.
- Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2021, An augmented q-factor model with expected growth, *Review of Finance* 25, 1–41.
- Hou, Kewei, Chen Xue, and Lu Zhang, 2017, A comparison of new factor models, *Fisher college of business working paper* 05.
- Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *The Review of Financial Studies* 33, 2019–2133.
- Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of finance* 48, 65–91.
- Jegadeesh, Narasimhan, and Sheridan Titman, 2001, Profitability of momentum strategies: An evaluation of alternative explanations, *The Journal of finance* 56, 699–720.
- Jensen, Theis Ingerslev, Bryan T Kelly, and Lasse Heje Pedersen, 2021, Is there a replication crisis in finance?, *The Journal of Finance, Forthcoming* .
- Karolyi, G Andrew, and Stijn Van Nieuwerburgh, 2020, New methods for the cross-section of returns, *The Review of Financial Studies* 33, 1879–1890.
- Kaustia, Markku, and Ville Rantala, 2015, Social learning and corporate peer effects, *Journal of Financial Economics* 117, 653–669.
- Kelly, Bryan, Semyon Malamud, and Lasse Heje Pedersen, 2020, Principal portfolios, *The Journal of Finance* .
- Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.
- Kelly, Bryan T, and Dacheng Xiu, 2021, Factor models, machine learning, and asset pricing, *Machine Learning, and Asset Pricing (October 15, 2021)* .
- Kozak, Serhiy, 2020, Kernel trick for the cross-section, *Available at SSRN 3307895* .
- Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

- Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.
- Kusner, Matt, Stephen Tyree, Kilian Weinberger, and Kunal Agrawal, 2014, Stochastic neighbor compression, in *International conference on machine learning*, 622–630, PMLR.
- Leary, Mark T, and Michael R Roberts, 2014, Do peer firms affect corporate financial policy?, *The Journal of Finance* 69, 139–178.
- Lee, Charles, Paul Ma, and Charles CY Wang, 2016, The search for peer firms: When do crowds provide wisdom?, *Harvard Business School Accounting & Management Unit Working Paper* 14–46.
- Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas, 2015, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research* 247, 124–136.
- Lettau, Martin, and Markus Pelger, 2020, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.
- Lewellen, Jonathan, 2015, The cross section of expected stock returns, *Critical Finance Review* 4, 1–44.
- Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial economics* 96, 175–194.
- Li, Bin, and Alberto G Rossi, 2020, Selecting mutual funds from the stocks they hold: A machine learning approach, *Available at SSRN 3737667*.
- Light, Nathaniel, Denys Maslov, and Oleg Rytchkov, 2017, Aggregation of information about the cross section of stock returns: A latent variable approach, *The Review of Financial Studies* 30, 1339–1381.
- Martin, Ian WR, and Stefan Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.
- Menzly, Lior, and Oguzhan Ozbas, 2010, Market segmentation and cross-predictability of returns, *The Journal of Finance* 65, 1555–1580.

- Moritz, Benjamin, and Tom Zimmermann, 2016, Tree-based conditional portfolio sorts: The relation between past and future stock returns, *Available at SSRN 2740751* .
- Moskowitz, Tobias J, and Mark Grinblatt, 1999, Do industries explain momentum?, *The Journal of finance* 54, 1249–1290.
- Müller, Sebastian, 2019, Economic links and cross-predictability of stock returns: Evidence from characteristic-based “styles”, *Review of Finance* 23, 363–395.
- Nagel, Stefan, 2021, Machine learning in asset pricing, in *Machine Learning in Asset Pricing* (Princeton University Press).
- Nagel, Stefan, and Kenneth J Singleton, 2011, Estimation and evaluation of conditional asset pricing models, *The Journal of Finance* 66, 873–909.
- Rapach, David E, Jack K Strauss, and Guofu Zhou, 2013, International stock return predictability: what is the role of the united states?, *The Journal of Finance* 68, 1633–1662.
- Seyfi, Sina, 2023, Essence of the cross section, *Available at SSRN 4466972* .
- Shue, Kelly, 2013, Executive networks and firm policies: Evidence from the random assignment of mba peers, *The Review of Financial Studies* 26, 1401–1442.
- Tarlow, Daniel, Kevin Swersky, Laurent Charlin, Ilya Sutskever, and Rich Zemel, 2013, Stochastic k-neighborhood selection for supervised and unsupervised learning, in *International Conference on Machine Learning*, 199–207, PMLR.

# Appendices

Acronym	Firm characteristic	Acronym	Firm characteristic
absacc	Absolute accruals	mom36m	36-month momentum
acc	Working capital accruals	mom6m	6-month momentum
aeavol	Abnormal earnings announcement	ms	Financial statement score
age	# years since first Compustat coverage	mvel1	Size
agr	Asset growth	mve	ia Industry-adjusted size
baspread	Bid-ask spread	nincr	Number of earnings increases
beta	Beta	operprof	Operating profitability
betasq	Beta squared	orgcap	Organizational capital
bm	Book-to-market	pchcapx_ia	ia Industry adjusted % change in capital expenditures
bm_ia	ia Industry-adjusted book to market	pchcurrat	% change in current ratio
cash	Cash holdings	pchdepr	% change in depreciation
cashdebt	Cash flow to debt	pchgm_pchsale	% change in gross margin - % change in sales
cashpr	Cash productivity	pchquick	% change in quick ratio
cfp	Cash flow to price ratio	pchsale_pchinvt	% change in sales - % change in inventory
cfp_ia	Industry-adjusted cash flow to price ratio Asnes	pchsale_pchrect	% change in sales - % change in A/R
chatoia	Industry-adjusted change in asset turnover	pchsale_pchxsga	% change in sales - % change in SG&A
chesho	Change in shares outstanding	pchsaleinv	% change sales-to-inventory
chempia	Industry-adjusted change in employees	pctacc	Percent accruals
chinv	Change in inventory	pricedelay	Price delay
chmom	Change in 6-month momentum	ps	Financial statements score
chpmia	Industry-adjusted change in profit margin	quick	Quick ratio
chtx	Change in tax expense	rd	R&D increase
cinvest	Corporate investment	rd_mve	R&D to market capitalization
convind	Convertible debt indicator	rd_sale	R&D to sales
currat	Current ratio	realestate	Real estate holdings
depr	Depreciation / PP&E	retvol	Return volatility
divi	Dividend initiation	roaq	Return on assets
divo	Dividend omission	roavol	Earnings volatility
dolvol	Dollar trading volume	roeq	Return on equity
dy	Dividend to price	roic	Return on invested capital
ear	Earnings announcement return	rsup	Revenue surprise
egr	Growth in common shareholder equity	salecash	Sales to cash
ep	Earnings to price	saleinv	Sales to inventory
gma	Gross profitability	salerec	Sales to receivables
grCAPX	Growth in capital expenditures	secured	Secured debt
grltnoa	Growth in long term net operating assets	securedind	Secured debt indicator
herf	Industry sales concentration	sgr	Sales growth
hire	Employee growth rate	sin	Sin stocks
idiovol	Idiosyncratic return volatility	sp	Sales to price
ill	Illiquidity	std_dolvol	Volatility of liquidity (dollar trading volume)
indmom	Industry momentum	std_turn	Volatility of liquidity (share turnover)
invest	Capital expenditures and inventory	stdacc	Accrual volatility
lev	Leverage	stdcf	Cash flow volatility
lgr	Growth in long-term debt	tang	Debt capacity/firm tangibility
maxret	Maximum daily return	tb	Tax income to book income
mom12m	12-month momentum	turn	Share turnover
mom1m	1-month momentum	zerotrade	Zero trading days

**Table A1-** List of acronyms and characteristics provided by [Gu et al. \(2020\)](#)

	Target variable: $r_t$				Target variable: $E(r_t)$			
	mean	std	$t$ -stat	SR	mean	std	$t$ -stat	SR
<b>Panel (a): All data</b>								
<b>1</b>	-0.23	9.25	-0.56	-0.09	-0.20	10.08	-0.44	-0.07
<b>2</b>	0.41	6.91	1.33	0.20	0.41	7.13	1.28	0.20
<b>3</b>	0.66	5.61	2.65	0.41	0.58	5.42	2.42	0.37
<b>4</b>	0.69	4.39	3.53	0.55	0.64	4.41	3.28	0.51
<b>5</b>	0.79	3.38	5.23	0.81	0.74	3.95	4.21	0.65
<b>6</b>	0.76	3.60	4.72	0.73	0.87	3.96	4.95	0.76
<b>7</b>	0.94	4.29	4.91	0.76	0.97	4.31	5.05	0.78
<b>8</b>	1.02	5.24	4.36	0.67	1.06	4.95	4.82	0.74
<b>9</b>	1.10	6.59	3.75	0.58	1.14	6.39	4.01	0.62
<b>10</b>	1.49	8.25	4.05	0.62	1.03	8.52	2.70	0.42
<b>LS</b>	1.72	3.18	12.11	1.87	1.22	5.46	5.02	0.78
<b>Panel (b): All but tiny stocks</b>								
<b>1</b>	0.14	8.59	0.38	0.06	-0.04	9.28	-0.11	-0.02
<b>2</b>	0.63	6.24	2.26	0.35	0.60	6.46	2.10	0.32
<b>3</b>	0.73	5.06	3.24	0.50	0.72	5.17	3.14	0.49
<b>4</b>	0.78	4.16	4.19	0.65	0.77	4.53	3.82	0.59
<b>5</b>	0.78	3.81	4.59	0.71	0.82	4.00	4.60	0.71
<b>6</b>	0.80	4.03	4.47	0.69	0.77	3.98	4.37	0.67
<b>7</b>	0.90	4.36	4.63	0.71	0.90	4.50	4.50	0.70
<b>8</b>	1.01	5.05	4.47	0.69	0.99	4.76	4.65	0.72
<b>9</b>	1.03	5.89	3.93	0.61	1.00	5.62	4.01	0.62
<b>10</b>	1.17	7.49	3.51	0.54	1.16	8.01	3.24	0.50
<b>LS</b>	1.03	2.82	8.16	1.26	1.20	5.04	5.35	0.83
<b>Panel (c): Large stocks</b>								
<b>1</b>	0.34	8.14	0.92	0.14	0.25	8.79	0.63	0.10
<b>2</b>	0.72	5.99	2.69	0.42	0.63	6.27	2.24	0.35
<b>3</b>	0.72	5.04	3.20	0.49	0.73	5.14	3.20	0.49
<b>4</b>	0.73	4.52	3.63	0.56	0.81	4.70	3.86	0.60
<b>5</b>	0.79	3.90	4.55	0.70	0.86	4.23	4.57	0.71
<b>6</b>	0.85	4.25	4.48	0.69	0.73	4.25	3.88	0.60
<b>7</b>	0.85	4.43	4.33	0.67	0.86	4.48	4.30	0.66
<b>8</b>	0.92	4.96	4.17	0.64	0.86	4.66	4.12	0.64
<b>9</b>	0.96	5.51	3.90	0.60	0.87	5.28	3.71	0.57
<b>10</b>	0.98	7.10	3.11	0.48	1.00	7.41	3.04	0.47
<b>LS</b>	0.65	3.11	4.67	0.72	0.76	5.49	3.09	0.48

**Table A2-** The performance of equally-weighted portfolios with 94 characteristics in 1980-2021

This table reports the average monthly out-of-sample performance of equally-weighted portfolios based on the all 94 characteristics listed in table A1. The four left columns show the results when the in-sample labels are ranked based on the realized returns (according to equation 4), while the four right columns show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). Columns titled "mean" show the average monthly excess returns of created portfolios from Jan 1980 to Dec 2021 in percentage. std is the monthly standard deviation of portfolios.  $t$  stat shows if the risk premiums are significantly different from zero and SR demonstrates the annualized Sharpe ratio. Panel (a) considers all data except the 5% tiniest for creating portfolios, while panel (b) and (c) includes all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median), respectively. LS shows the performance of a long-short portfolio which goes long (short) in decile 10 (1). The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The value-weighted counterparts are shown in table 1.

	Target variable: $r_t$			Target variable: $E(r_t)$		
	All data	All but tiny	Large stocks	All data	All but tiny	Large stocks
$\alpha_{\text{CAPM}}$	1.75	1.17	0.80	1.32	1.38	0.96
<b>t-stat</b>	12.21	9.58	5.96	5.41	6.18	3.96
<b>adj <math>R^2</math></b>	0.00	0.09	0.09	0.01	0.04	0.05
$\alpha_{\text{FF3}}$	1.77	1.15	0.80	1.44	1.52	1.15
<b>t-stat</b>	12.30	9.43	5.92	5.99	7.06	5.06
<b>adj <math>R^2</math></b>	0.00	0.09	0.08	0.05	0.12	0.17
$\alpha_{\text{Carhart}}$	1.43	0.81	0.46	0.69	0.79	0.40
<b>t-stat</b>	11.92	8.86	4.23	4.36	6.35	2.91
<b>adj <math>R^2</math></b>	0.33	0.50	0.43	0.61	0.71	0.71
$\alpha_{\text{FF5}}$	1.53	0.93	0.60	1.23	1.30	0.99
<b>t-stat</b>	10.63	7.63	4.38	4.93	5.88	4.18
<b>adj <math>R^2</math></b>	0.09	0.16	0.13	0.07	0.13	0.17
$\alpha_q$	1.31	0.74	0.42	0.69	0.81	0.49
<b>t-stat</b>	9.98	6.70	3.32	3.05	4.07	2.25
<b>adj <math>R^2</math></b>	0.26	0.33	0.26	0.24	0.31	0.30
$\alpha_{q^5}$	1.16	0.63	0.27	0.50	0.61	0.20
<b>t-stat</b>	8.32	5.39	1.98	2.08	2.87	0.86
<b>adj <math>R^2</math></b>	0.27	0.33	0.28	0.25	0.32	0.32

**Table A3-** Risk adjusted returns for equally-weighted long-short portfolios

this table reports monthly alphas,  $t$  values and adjusted  $R^2$  for out-of-sample performance of equally-weighted long-short portfolios. The three left columns show the results when the in-sample stocks are ranked based on the realized returns (according to equation 4), while the three right columns show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). All data includes the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The value-weighted counterparts are shown in Table 2.

	All data		All but tiny		Large stocks	
	3 char	12 char	3 char	12 char	3 char	12 char
<b>Panel (a): Predicting <math>r_t</math></b>						
<b>mean (1)</b>	0.15	-0.28	0.39	0.05	0.49	0.17
<i>t</i> -stat	0.39	-0.71	1.22	0.14	1.70	0.48
<b>mean (10)</b>	1.00	1.63	1.03	1.22	0.99	1.10
<i>t</i> -stat	3.35	4.72	3.61	3.81	3.66	3.58
<b>mean (LS)</b>	0.86	1.91	0.64	1.17	0.50	0.93
<i>t</i> -stat	4.72	14.05	4.33	8.45	3.18	6.82
<b>SR</b>	0.73	2.17	0.67	1.30	0.49	1.05
$\partial_{\text{CAPM}}$	1.01	2.01	0.75	1.32	0.59	1.06
<i>t</i> -stat	5.66	14.88	5.12	9.79	3.75	7.89
$\partial_{\text{FF3}}$	1.06	2.01	0.80	1.30	0.66	1.06
<i>t</i> -stat	5.94	14.88	5.51	9.64	4.39	7.83
$\partial_{\text{Carhart}}$	0.50	1.69	0.30	0.91	0.14	0.70
<i>t</i> -stat	4.26	15.06	3.75	9.43	1.70	6.69
$\partial_{\text{FF5}}$	0.84	1.73	0.64	1.04	0.54	0.83
<i>t</i> -stat	4.61	13.19	4.28	7.74	3.47	6.06
$a_q$	0.45	1.54	0.34	0.82	0.24	0.61
<i>t</i> -stat	2.78	12.85	2.53	6.75	1.68	4.95
$a_{q^5}$	0.36	1.39	0.17	0.69	0.10	0.45
<i>t</i> -stat	2.07	10.98	1.19	5.37	0.67	3.40
<b>Panel (b): Predicting <math>E(r_t)</math></b>						
<b>mean (1)</b>	-0.23	-0.13	-0.03	-0.12	0.20	0.03
<i>t</i> -stat	-0.52	-0.29	-0.08	-0.29	0.56	0.08
<b>mean (10)</b>	1.34	1.04	1.30	1.17	1.12	1.05
<i>t</i> -stat	4.09	2.94	4.11	3.52	3.90	3.43
<b>mean (LS)</b>	1.57	1.17	1.32	1.29	0.92	1.02
<i>t</i> -stat	5.22	4.21	4.66	5.02	3.21	3.86
<b>SR</b>	0.81	0.65	0.72	0.77	0.50	0.60
$\partial_{\text{CAPM}}$	1.73	1.35	1.53	1.51	1.11	1.25
<i>t</i> -stat	5.73	4.86	5.40	5.98	3.89	4.80
$\partial_{\text{FF3}}$	1.89	1.48	1.71	1.69	1.30	1.42
<i>t</i> -stat	6.38	5.43	6.28	6.95	4.75	5.67
$\partial_{\text{Carhart}}$	0.88	0.61	0.69	0.80	0.27	0.52
<i>t</i> -stat	5.23	3.50	6.18	7.02	2.45	4.25
$\partial_{\text{FF5}}$	1.61	1.22	1.42	1.51	1.02	1.24
<i>t</i> -stat	5.26	4.33	5.05	5.99	3.61	4.77
$a_q$	0.91	0.60	0.76	0.91	0.40	0.67
<i>t</i> -stat	3.30	2.35	2.99	3.90	1.52	2.74
$a_{q^5}$	0.69	0.36	0.49	0.57	0.14	0.33
<i>t</i> -stat	2.34	1.32	1.83	2.33	0.51	1.30

**Table A4-** The performance of equally-weighted portfolios based on 3 and 12 characteristics in 1980-2021

This table reports the average monthly out-of-sample performance of equally-weighted portfolios based on 3 and 12 characteristics. Panel (a) show the results when the in-sample labels are ranked based on the realized returns (according to equation 4), while panel (b) show the counterpart results when the labels are based on the average of last year returns ( $T = 12$  in equation 3). Rows titled "mean (1), (10) and (LS)" show the average monthly excess returns of portfolios 1, 10 and long-short portfolio from Jan 1980 to Dec 2021 in percentage and SR demonstrates the annualized Sharpe ratio. The number of neighbours is considered  $k = 1000$  with  $t = 120$  the number of months included in the in-sample data. The value-weighted counterpart results are presented in table 3.

	All data			All but tiny			Large stocks		
	3 char	12 char	94 char	3 char	12 char	94 char	3 char	12 char	94 char
$a_{\text{CAPM}}$	1.74	1.32	1.24	1.61	1.62	1.54	1.21	1.48	1.39
<b><math>t</math>-stat</b>	5.71	4.85	4.77	5.59	6.36	5.80	4.08	5.36	4.14
<b>adj <math>R^2</math></b>	0.02	0.02	0.00	0.04	0.04	0.04	0.03	0.05	0.04
$a_{\text{FF3}}$	1.87	1.45	1.37	1.79	1.78	1.72	1.40	1.63	1.58
<b><math>t</math>-stat</b>	6.25	5.43	5.41	6.46	7.26	6.71	4.94	6.07	4.84
<b>adj <math>R^2</math></b>	0.06	0.06	0.05	0.11	0.11	0.12	0.12	0.11	0.10
$a_{\text{Carhart}}$	0.88	0.61	0.69	0.76	0.92	0.95	0.34	0.75	0.79
<b><math>t</math>-stat</b>	4.89	3.50	3.54	6.46	7.12	5.33	2.85	4.58	2.95
<b>adj <math>R^2</math></b>	0.67	0.60	0.46	0.85	0.76	0.58	0.85	0.68	0.41
$a_{\text{FF5}}$	1.52	1.21	1.15	1.48	1.62	1.54	1.10	1.48	1.49
<b><math>t</math>-stat</b>	4.95	4.37	4.37	5.20	6.36	5.81	3.77	5.31	4.39
<b>adj <math>R^2</math></b>	0.09	0.07	0.07	0.14	0.12	0.12	0.14	0.11	0.10
$a_q$	0.82	0.61	0.66	0.82	1.05	0.97	0.47	0.91	0.94
<b><math>t</math>-stat</b>	2.99	2.45	2.67	3.17	4.38	3.97	1.74	3.42	2.86
<b>adj <math>R^2</math></b>	0.29	0.25	0.20	0.30	0.24	0.28	0.27	0.22	0.18
$a_{q^5}$	0.64	0.40	0.45	0.54	0.71	0.71	0.22	0.55	0.41
<b><math>t</math>-stat</b>	2.19	1.48	1.72	1.97	2.82	2.76	0.75	1.95	1.19
<b>adj <math>R^2</math></b>	0.29	0.26	0.21	0.31	0.26	0.29	0.27	0.24	0.21

**Table A5-** Risk adjusted returns for an equally-weighted long-short portfolio with a stochastic selection

This table reports monthly alphas,  $t$  values and adjusted  $R^2$  for out-of-sample performance of equally-weighted long-short portfolios where there is a stochastic selection in the training sample with  $\rho = 0.05$ . The results are shown for the case three cases when 3 characteristics, 12 characteristics and 94 characteristics are used to find neighbouring assets. The target variables are defined based on the average of last year returns ( $T = 12$  in equation 3). All data includes the universe of stocks except the 5% tiniest for creating portfolios, while all but tiny and large stocks include the assets with above 20% and 50% NYSE market-cap, respectively. I consider six factor models and report alpha and their  $t$  stats based on them. These six models include CAPM, Fama French three factor model (FF3), Carhart four factor model (Carhart), Fama French 5 factor model (FF5),  $q$  model and augmented  $q$  model with expected growth ( $q^5$ ). The number of neighbours is considered  $k = 1000$  with  $t = 120$  of months included in the in-sample data. The value-weighted counterpart results are shown in Table 4.

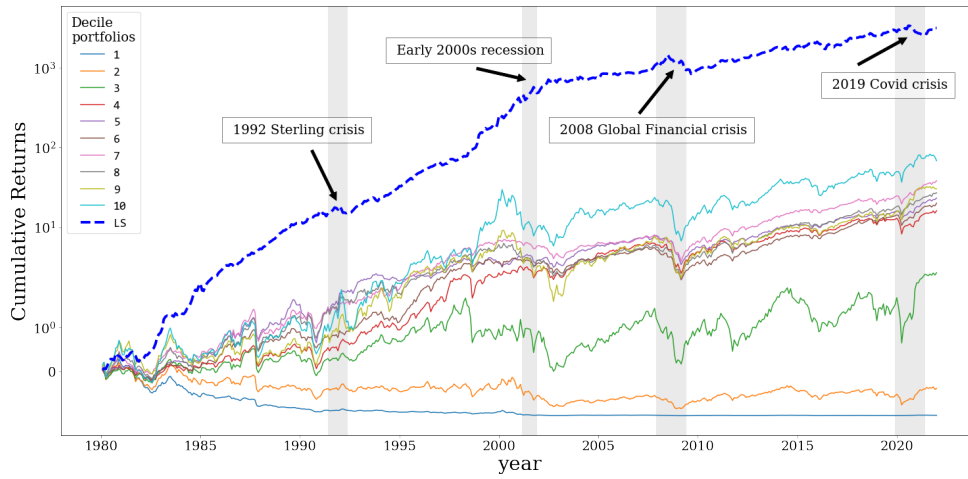
---

Anomaly

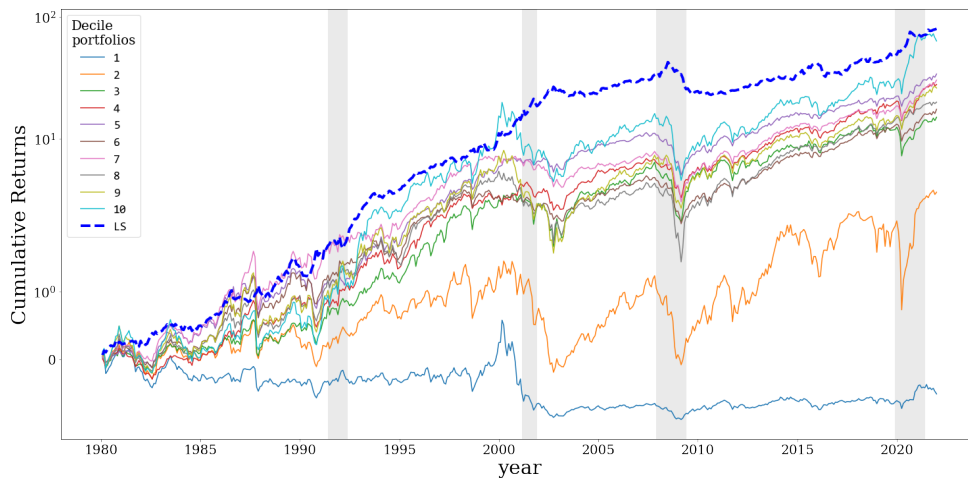
---

- 1- cumulative abnormal returns around earnings announcement dates, 1-month holding period
  - 2- cumulative abnormal returns around earnings announcement dates, 6-month holding period
  - 3- cumulative abnormal returns around earnings announcement dates, 12-month holding period
  - 4- customer industries momentum, 1-month holding period
  - 5- customer industries momentum, 6-month holding period
  - 6- customer industries momentum, 12-month holding period
  - 7- customer momentum, 1-month holding period
  - 8- customer momentum, 12-month holding period
  - 9- changes in analyst earnings forecasts, 1-month holding period
  - 10- changes in analyst earnings forecasts, 6-month holding period
  - 11- changes in analyst earnings forecasts, 12-month holding period
  - 12- industry lead-lag effect in earnings surprises, 1-month holding period
  - 13- industry lead-lag effect in prior returns, 1-month holding period
  - 14- industry lead-lag effect in prior returns, 6-month holding period
  - 15- industry lead-lag effect in prior returns, 12-month holding period
  - 16- industry momentum, 1-month holding period
  - 17- industry momentum, 6-month holding period
  - 18- industry momentum, 12-month holding period
  - 19- the number of quarters with consecutive earnings increase, 1-month holding period
  - 20- 52-week high, 6-month holding period
  - 21- 52-week high, 12-month holding period
  - 22- prior 6-month returns, 1-month holding period
  - 23- prior 6-month returns, 6-month holding period
  - 24- prior 6-month returns, 12-month holding period
  - 25- prior 11-month returns, 1-month holding period
  - 26- prior 11-month returns, 6-month holding period
  - 27- prior 11-month returns, 12-month holding period
  - 28- revisions in analyst earnings forecasts, 1-month holding period
  - 29- revisions in analyst earnings forecasts, 6-month holding period
  - 30- 6-month residual momentum, 6-month holding period
  - 31- 6-month residual momentum, 12-month holding period
  - 32- 11-month residual momentum, 1-month holding period
  - 33- 11-month residual momentum, 6-month holding period
  - 34- 11-month residual momentum, 12-month holding period
  - 35- revenue surprises, 1-month holding period
  - 36- supplier industries momentum, 1-month holding period
  - 37- supplier industries momentum, 12-month holding period
  - 38- segment momentum, 1-month holding period
  - 39- segment momentum, 12-month holding period
  - 40- standard unexpected earnings, 1-month holding period
  - 41- standard unexpected earnings, 6-month holding period
- 

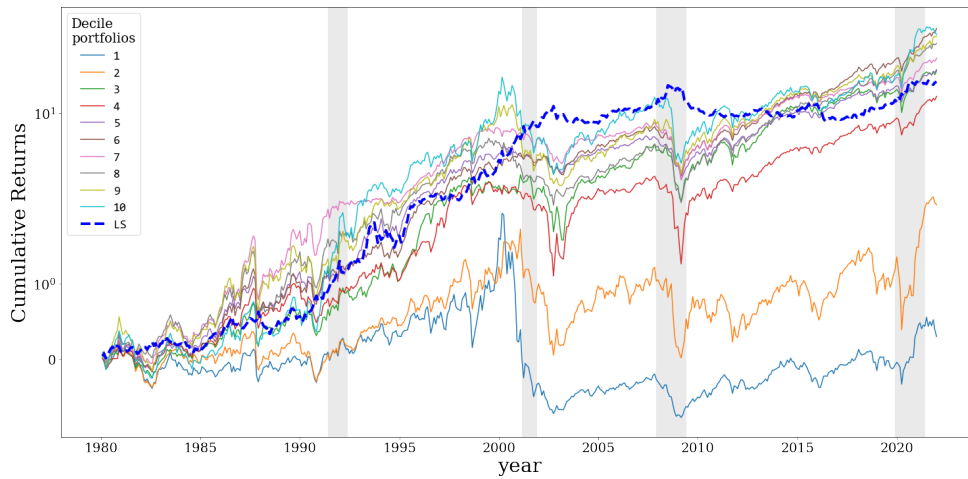
**Table A6-** List of momentum anomalies provided by [Hou et al. \(2020\)](#)



(a) All data



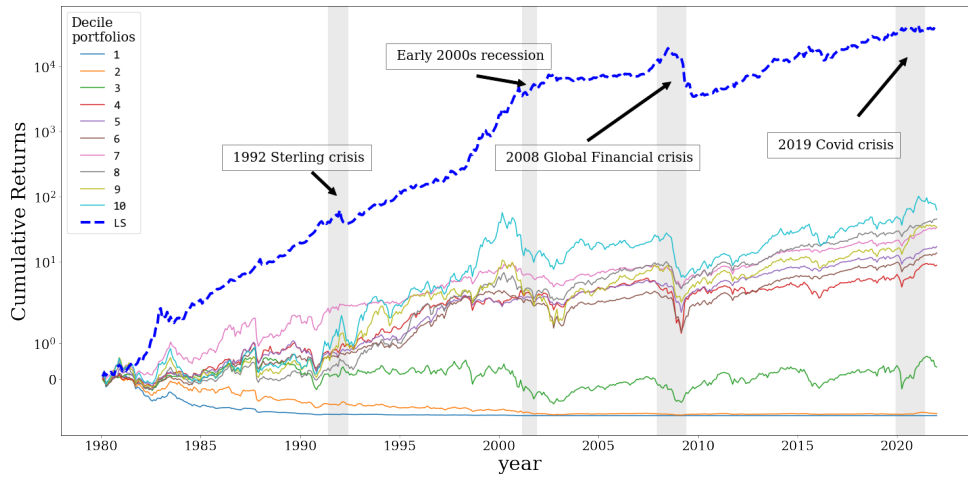
(b) All but tiny stocks



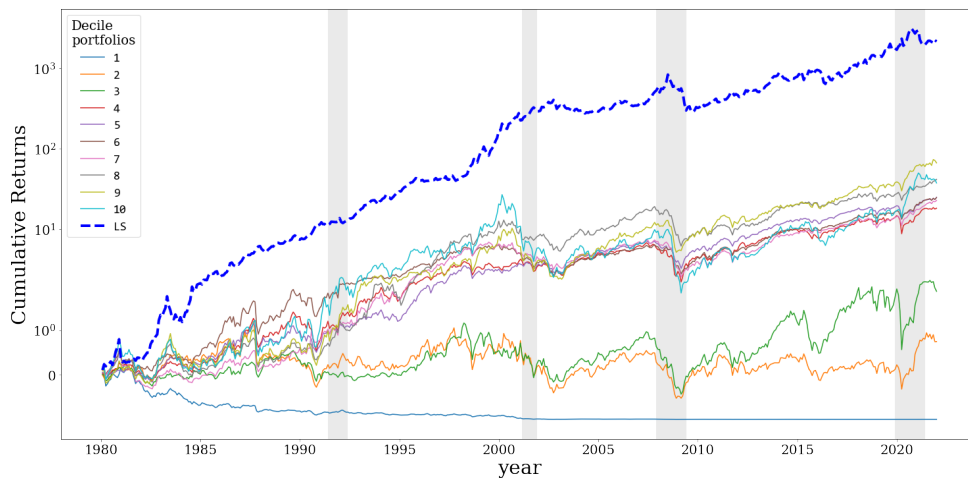
(c) Large stocks

**Figure A1.** Cumulative returns of decile portfolios for the period 1980-2021 when the target variable is realized returns

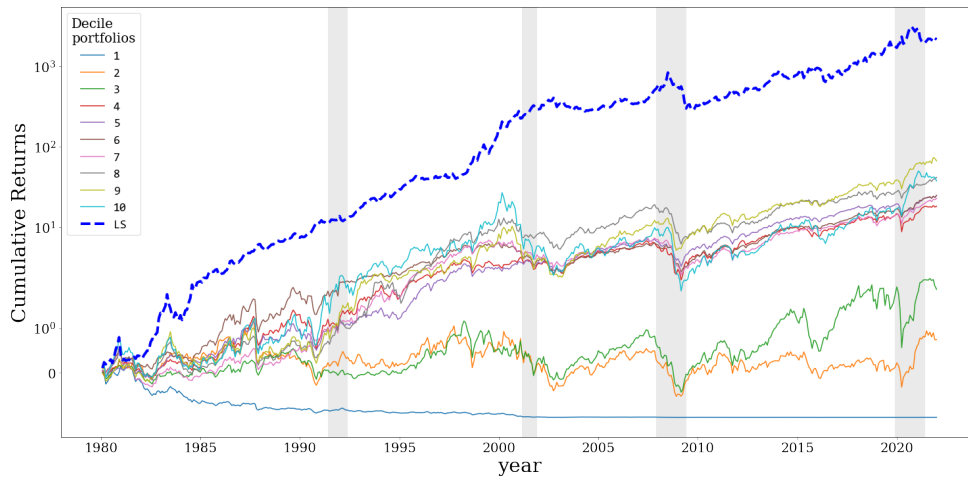
This figure shows the cumulative returns of all portfolios from 1980-2020 in the main analysis when  $k = 1000$  and the target variables are defined based on the realized returns ( $T = 1$  in equation 3). Panel (a) shows portfolios consisting all but 5% tiniest ones. Panel (b) and (c) show portfolios containing all but tiny (above 20% NYSY percentile) and large stocks (above NYSE median). Four crisis periods are shown in gray and they include 1992 Sterling crisis, Early 2000s recession, 2008 global financial crisis and 2019 Covid crisis. The long-short portfolio strategy reacts to the crisis periods. The y-axis is in a logarithmic scale.



(a) All data



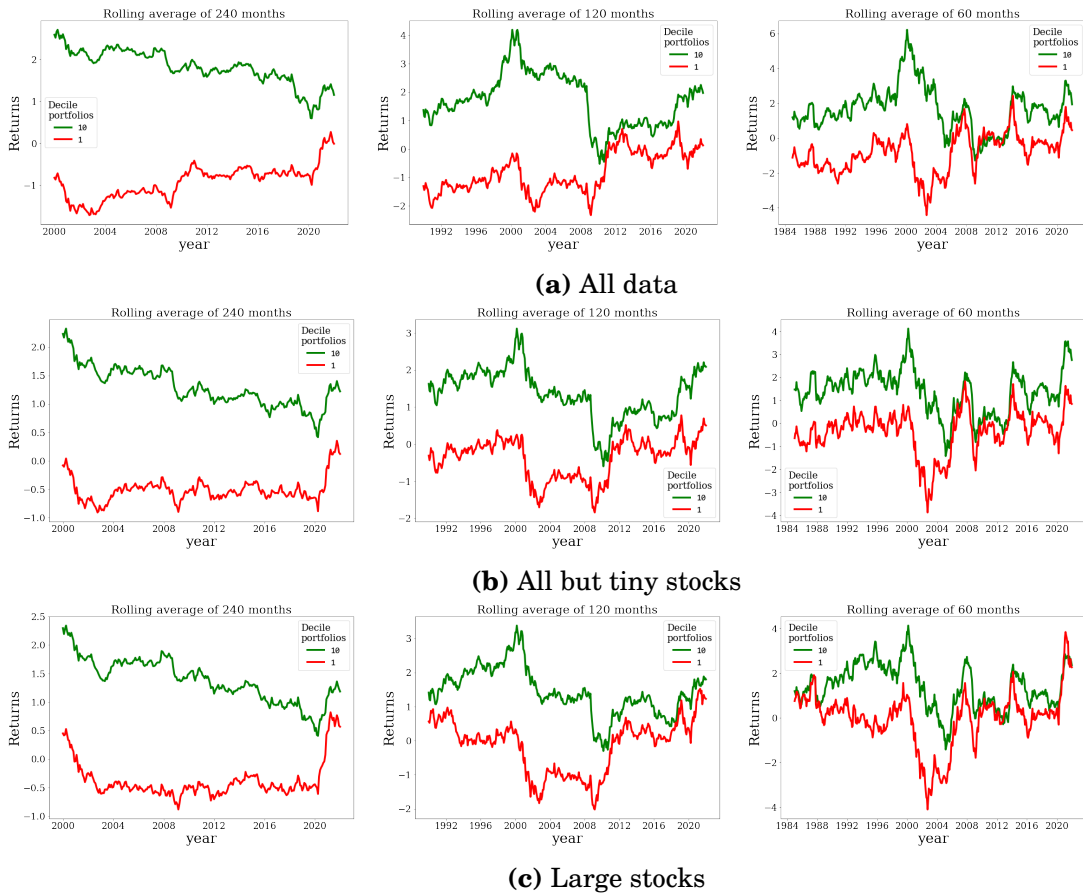
(b) All but tiny stocks



(c) Large stocks

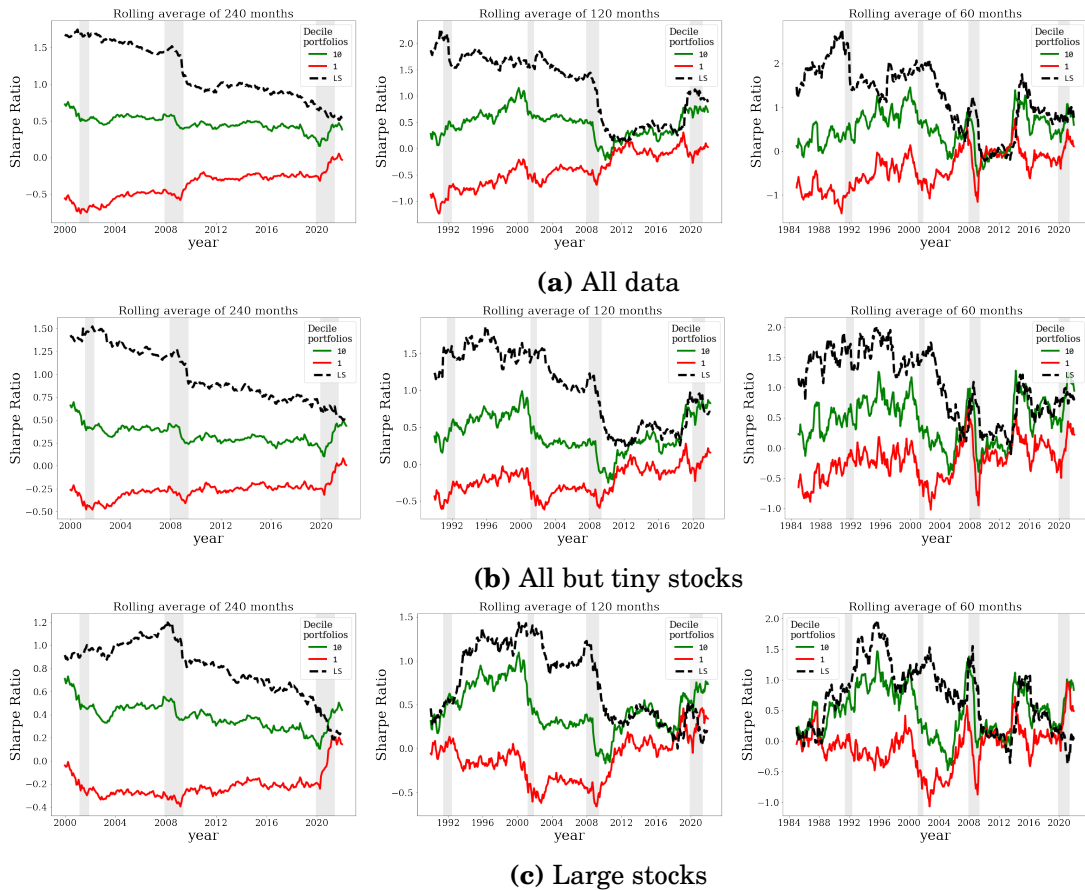
**Figure A2.** Cumulative returns of decile portfolios for the period 1980-2021 with a stochastic selection

This figure shows the cumulative returns of all portfolios from 1980-2020 in the main analysis when  $k = 1000$ ,  $t = 120$ , and  $\rho = 0.05$  and the target variables are defined based on the average of last year returns ( $T = 12$  in equation 3). Panel (a) shows portfolios consisting all but 5% tiniest ones. Panel (b) and (c) show portfolios containing all but tiny (above 20% NYSE percentile) and large stocks (above NYSE median). Four crisis periods are shown in gray and they include 1992 Sterling crisis, Early 2000s recession, 2008 global financial crisis and 2019 Covid crisis. The long-short portfolio strategy reacts to the crisis periods. The y-axis is in a logarithmic scale.



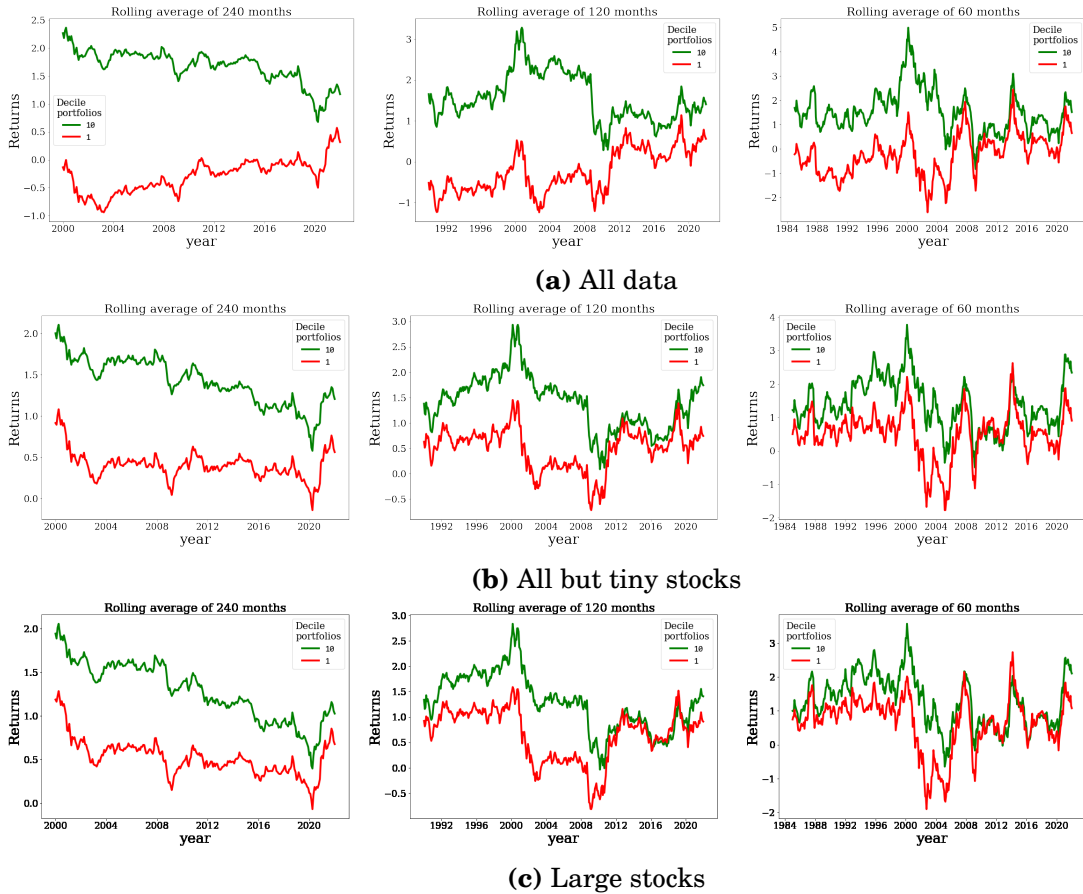
**Figure A3.** Rolling average of spreads between decile 1 and 10 in the period 1980-2021

This figure shows the rolling average of the spread generated from decile portfolio 1 and 10. Left column show the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the average of last year returns ( $T = 12$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ . The setting is  $\rho = 0.05$ .



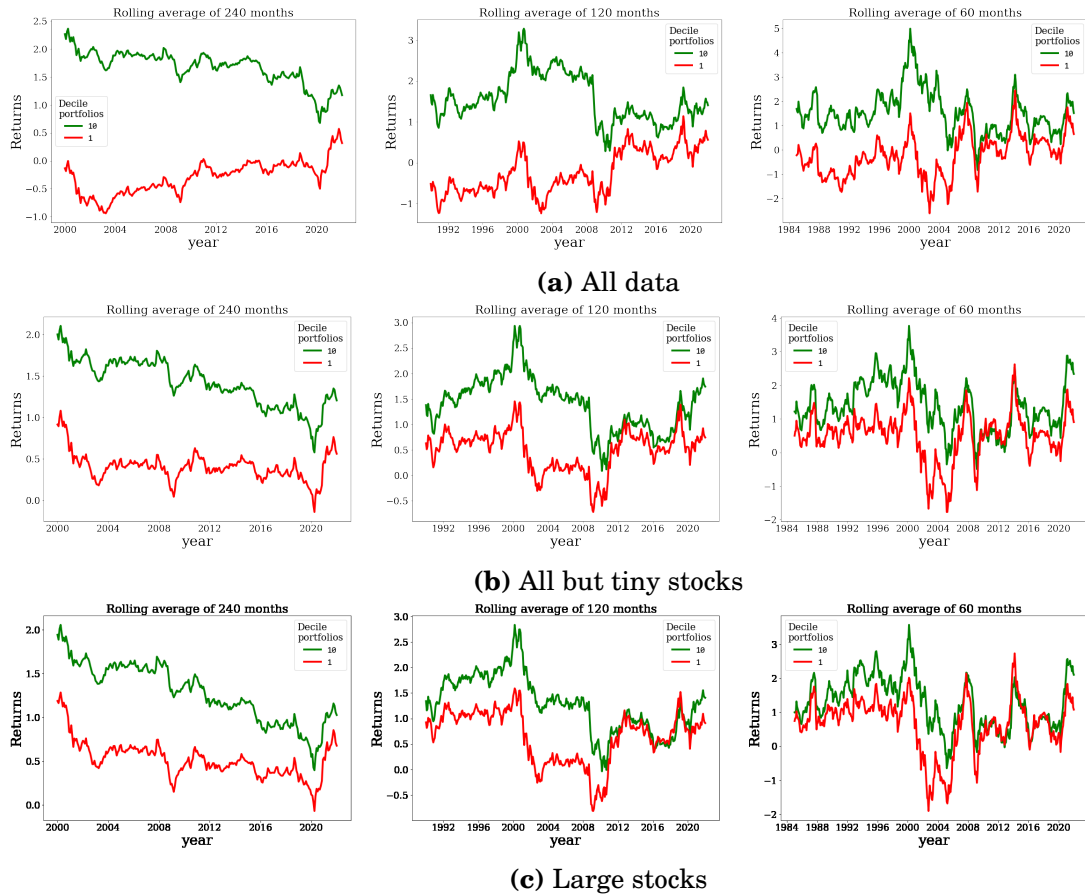
**Figure A4.** Rolling average of Sharpe Ratio for decile 1 and 10 and a long-short portfolio in the period 1980-2021 with a stochastic selection

This figure shows the rolling average of the Sharpe Ratios generated from decile portfolio 1 and 10 and also a long-short portfolio. Left column show the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the average of last year returns ( $T = 12$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ . The crisis periods are shown in gray. The setting is  $\rho = 0.05$ .



**Figure A5.** Rolling average of spreads between decile 1 and 10 in the period 1980-2021 when predicting based on realized returns

This figure shows the rolling average of the spread generated from decile portfolio 1 and 10. Left column show the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the realized returns ( $T = 1$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ .



**Figure A6.** Rolling average of Sharpe Ratio for decile 1 and 10 and a long-short portfolio in the period 1980-2021 when predicting based on realized returns

This figure shows the rolling average of the Sharpe Ratios generated from decile portfolios 1 and 10 and also a long-short portfolio. The left column shows the rolling average when the rolling window is 240 months, while the middle and right columns show the counterpart 120 and 60 months of rolling windows. Panel (a) shows the decile portfolios created by all except 5% tiniest stocks, panel (b) shows the portfolios with all but tiny stocks, and panel (c) includes only large stocks. In this graph the value-weighted portfolios are created with considering 94 characteristics. The target variables are defined based on the realized returns ( $T = 1$  in equation 3) with  $t = 120$  the number of months included in the in-sample data, and  $k = 1000$ . The crisis periods are shown in gray.