Machine Learning Predictions of International Stock Returns

Ondrej Tobek

Corpus Christi College and Faculty of Economics
University of Cambridge

Supervisor: Professor Oliver Linton
Research Advisor: Professor Mark Salmon

This dissertation is submitted for the degree of Doctor of Philosophy

July, 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

To my parents, Hana Tobkova and Ladislav Tobek

Preface

This dissertation is broadly describing predictability of returns on individual stocks in international context. The first chapter covers required prerequisites for any study of fundamental anomalies outside the US. The second chapter studies the predictability of stock returns at an annual frequency. The last chapter then looks at possible profitability of the predictability on liquid universe of stocks at monthly frequency.

In the first chapter, we study the role of the choice of a fundamental database on the portfolio returns of a set of 74 fundamental anomalies. We benchmark Compustat by comparing it to Datastream in the US and find systematic differences in the raw financial statements across the databases. These differences only have a small effect on the returns of anomalies when they are constructed on stock-months existing in both databases. Different stock coverage across the databases, however, leads to large statistically and economically significant disparities in the returns. Profitability anomalies yield negative returns on the Datastream universe.

In the second chapter, we study statistical significance of 93 fundamental anomalies published in academic journals in a multiple hypothesis setting. We generate a universe of 48,387 data-mined fundamental strategies in order to overcome a problem of not being able to observe strategies that were tried but not published. The multiple hypothesis tests reveal that the number of significant anomalies heavily depends on the precise specification of the tests. We show that the adjustment of standard errors on portfolio returns for heteroskedasticity and autocorrelation is of first order importance and t-statistics on the portfolio returns may not have critical values of the normal distribution.

In the third chapter, we study out-of-sample returns on 153 anomalies in equities documented in academic literature. We show that machine learning techniques that aggregates all the anomalies into one mispricing signal are 4 times more profitable than a strategy based on individual anomalies and survive on a liquid universe of stocks. The machine learning also leads to 2 times larger Sharpe ratios with respect to the corresponding standard finance methods. We next study the value of international evidence for selection of quantitative strategies that outperform out-of-sample. Past performance of quantitative strategies in the regions other than the US does not help to pick out-of-sample winning strategies in the US. Past evidence from the US, however, captures most of the predictability within the other regions. The value of international evidence in empirical asset pricing is therefore very limited.

Acknowledgements

I would like to sincerely thank my supervisor, Oliver Linton, for guiding my research and academic progress. I have learned a lot from his professionalism and uncompromising standards for high-quality research. I would also like to thank my research advisor, Mark Salmon, who provided me with an essential feedback needed to develop this text. His breadth of knowledge of financial markets and state of art interdisciplinary academic research will continue to inspire me during my professional career.

I am also grateful to Martin Hronec who was a great friend during my PhD studies and who contributed to development of the universe of anomalies studied here. The first and third chapter originally started as our joint project and benefited greatly from his comments. The construction of the anomalies essentially corresponds to a replication of the relevant published papers, a process which is prone to errors due to the large number of replicated papers. Martin has aided me in coding of many of the fundamental anomalies and has provided an indispensable four eye check for the rest. He was also responsible for drafting descriptions of the fundamental anomalies in Appendix B.

Chapter 1

Does the Source of Fundamental Data Matter?

Most of the research in accounting and finance relies only on two databases, the Center for Research in Security Prices (CRSP) and Compustat, since they are the most easily available to academics. However, these databases are not as heavily used outside academia and are not error-proof. Can these errors create significant biases across studies or are the errors idiosyncratic and no cause for worry? We test this question by looking at the performance of 74 fundamental anomalies published in finance and accounting journals when they are constructed in the Compustat universe or alternatively in the Reuters Datastream universe.[1] We also test the role of trade data by comparing portfolio returns on the anomalies constructed with individual stock returns from Datastream or CRSP and fundamental signals constructed in Compustat.

The fundamental anomalies in this study describe characteristics related to individual stocks that can predict their future returns. No distinction is being made between characteristics that are related to risk premia and variables that are related to mispricing due to frictions or other market imperfections. The studied anomalies are, for example, accruals of Sloan (1996), earnings over price of Basu (1977), composite equity issuance of Daniel and Titman (2006), and R&D over Market Equity of Chan et al. (2001).

Another crucial aspect of the individual databases is the composition of the universe of stocks there. Academic studies mostly focus only on common stocks listed on countries' main exchanges, but this focus requires a classification by data vendors that is often wrong in earlier years. Some databases might also suffer from incomplete coverage for the stocks with low capitalization and the less frequently traded stocks. We study the implications of these differences among the databases for quantitative strategies. CRSP and Compustat are the primary source of data only in academia. Other data sources are more common in the industry. It is not obvious if the academic findings using CRSP and Compustat hold

[1] We sometimes call the Compustat universe as CRSP and Compustat universe since Compustat does not include trade data whereas Datastream contains both market and fundamental data. The fundamental sub-database in Datastream is called Worldscope and we denote it interchangeably as Datastream throughout this text.

on the other data sources. This study tries to provide quantitative evidence to bridge this gap in knowledge.

We first study the fundamental anomalies on a sample of stocks in CRSP that can be matched to fundamental data in both Datastream and Compustat. We start by comparing the individual raw items on the financial statements that are required for constructing the anomalies.[2] We find that the items can substantially differ across the two databases. There are some apparent patterns in the differences. They tend to cluster in areas where the data vendors require specific methodologies to be applied. Some examples include the treatment of short-term versus long-term debt, long-term leases, or financing items on cash flow statements. These substantial differences in raw items, however, rarely translate to differences in the portfolio returns on fundamental anomalies in the matched Datastream and Compustat sample of firms. Average correlation between portfolio returns on the anomalies created based on Datastream and portfolio returns on the anomalies created based on Compustat is 95.9%. There are also no apparent economically significant differences between the returns on the anomalies across the two databases.

The discrepancies are, however, substantially larger once we move outside the matched sample and construct anomalies on the full samples of companies in each fundamental database. We partially explain this outcome by the lower coverage of stocks with lower capitalization in Datastream in the earlier period, but some economically and statistically significant differences nonetheless remain.[3] The discrepancies are huge when the individual quantitative strategies are considered.[4] 41 of the 74 anomalies are significant at the 5% level in CRSP plus Compustat and 39 in Datastream over the 1990 to 2016 period. There are, however, only 29 anomalies that are significant in both. Inference for individual strategies thus suffers from large biases. The discrepancies are, however, much smaller for grouped anomalies. The average return on all 74 fundamental anomalies is almost identical among the two databases. Datastream and other alternative data sources are thus safe to use in the aggregate analysis of returns on anomalies, especially when micro-caps are excluded from the sample.

The fundamental coverage in Datastream significantly predicts expected returns on stocks in CRSP. Stocks without the fundamental coverage significantly underperform those with the coverage. The fundamental coverage effect on expected returns is closely related to the number of analysts covering effect in Elgers et al. (2001). The underperformance of stocks without the fundamental coverage is especially channeled to stocks with small operating profitability. Operating profitability anomaly yields substantially lower returns in Datastream because the low profitability stocks are less likely to be covered

[2]This comparison was similarly performed in Ulbricht and Weiner (2005), who studied sample differences in fundamental variables in Datastream and Compustat in the US.

[3]Datastream covers 87.5% of the overall capitalization of stocks in Compustat in 1990, but this coverage has increased to essentially 100% since 2005. The two databases, however, continue to cover different sets of stocks labeled as common equity. The differences in returns on anomalies therefore remain substantial even after 2005.

[4]We provide detailed results for each anomaly in the Appendix D.

there. A value-weighted strategy shorting stocks without the fundamental coverage in Datastream that are in the lowest profitability decile in Compustat yields 28% annually over the 2000 to 2016 period.

There are three main sources of the differences in the returns on the anomalies. Firstly, the imperfect coverage causes disparity in portfolio breakpoints across the databases. Using breakpoints from NYSE, or all-but-microcaps universe of stocks with full coverage in each region, elevates this problem. Secondly, the coverage of stocks within the population quantiles may differ. Value-weighting limits this problem since it shifts the focus on stocks that tend to have better coverage in all databases. Lastly, the databases may have idiosyncratic differences due to errors and design choices. Examples include different categorization of the individual securities and companies. These database-specific issues are the hardest to minimize and require a tailored solution every time.

The large discrepancies in the returns before 2005 can have implications for international studies. We show that the problems with coverage are also prevalent in Europe, Japan, and Asia Pacific before 2000. Datastream is widely used in academic international studies. Examples of studies that rely on Datastream include McLean et al. (2009), Hou et al. (2011a), Titman et al. (2013), Watanabe et al. (2013), and Jacobs (2016).[5] The performance of individual strategies constructed on Datastream outside the US can be connected to some biases as was documented in the US. It is therefore important to study if the imperfect coverage is a source of some concerns and what mitigating approaches can be taken. We simulate the effect of imperfect fundamental coverage in Japan and Asia Pacific before 2000 on later sample where there was essentially full coverage. The simulated imperfect coverage leads to the same conclusions as for the US and it can have a large impact on measurement of performance of anomalies in the international setting.

We test two constructions of portfolios that should lower the discrepancies due to fundamental coverage. Both of the methods shift the focus on larger capitalization stocks where the bias is smaller. The first method discards all the micro-caps stocks with capitalization smaller than the bottom decile of the NYSE. The second uses the breakpoint from the 1000 largest stocks in the region to construct the portfolios. We then use value-weighted returns in both of them. The correlation of portfolios between the two databases increases from 80.2% to approximately 86%., but substantial differences remain. There are 11 significant signals in Compustat and 12 in Datastream, but only 6 of those are common across the two databases for the all-but-micro-caps universe of stocks. We conclude that the choice of the fundamental database used can have a large impact on tests of individual quantitative strategies, and researchers should be aware of this impact. Screening out small cap stocks and value-weighting can nonetheless improve robustness of empirical findings.

We next study the implications of the fundamental database choice for a selection of independently significant signals. There is a large amount of recent literature that

[5] Fama and French (2012) and Fama and French (2017) use fundamental data from Datastream to fill in gaps from Bloomberg, but similar patterns in coverage are also expected there.

attempts to shrink the number of anomalies by finding those that are independently significant after controlling for all the others.[6] Here, we follow the methodology from Green et al. (2017) and use Fama and MacBeth (1973a) regressions of individual stock returns on rescaled fundamental characteristics and control for the false discovery rate. The results are overwhelming in the US, as there is only one significant anomaly out of 8 in Compustat that is common between the two databases. Both databases thus lead to very different discoveries. The differences in the US should translate to differences among selected anomalies in different global regions. Jacobs and Müller (2017a) indeed show that significant anomalies are very different across the global regions, and our analysis thus explains this striking inconsistency. Any study attempting to distil which anomalies are significant should therefore be aware that any selection procedure is very unstable and is dependent on the imperfections of the underlying data.

The conclusions of our study are not unique to Datastream but apply to all sources of historical fundamental data for international equities, given that none of them offers perfect coverage of all listed stocks. Dai (2012) documents the gaps in coverage in FactSet Fundamentals, Compustat Global, and Bureau Van Dijks international databases. Fama and French (2012) note gaps in the Bloomberg database. We focus only on anomalies created with fundamental data, but our conclusions are valid for trade data as well. Stocks covered in Datastream in 1990 correspond to 91.5% of the overall capitalization of all the stocks in CRSP, which is better than for fundamental data but is nowhere near perfect.

Ince and Porter (2006) have shown that the Datastream returns data has limitations, and some adjustments need to be applied to limit its errors. We propose several new ways to further limit the errors. We show that there are only a few discrepancies in returns with respect to CRSP after 2000. We recommend that the returns before 1990 should be winsorized at the 0.1% percentile and returns from 1990 to 1999 at the 0.01% percentile. We also propose a new way to correct the returns when there are stale quotes at the time of stock splits and other corporate events. Not implementing them can lead to erroneous returns of several thousand percent.

This study is the first to evaluate the impact of not including delisting returns in Datastream. Shumway (1997) showed that missing delisting returns in CRSP can have a large impact on the returns on some anomalies, such as size. He proposed that the missing performance related delisting returns in CRSP should be filled with -30% return. We revisit his analysis after over 20 years and conclude that the role of missing delisting returns is much smaller than originally documented. We find no economically significant bias in the returns on anomalies from ignoring the missing delisting returns in CRSP. Specifically, we note that omitting all delisting returns in CRSP leads to economically similar returns on our set of fundamental anomalies relative to properly accounting them. Missing delisting returns in Datastream therefore should not be a serious cause for a worry.

[6]See, for example, Lewellen et al. (2015), Green et al. (2017), Feng et al. (2017), and Freyberger et al. (2017) for evidence from the US and Jacobs and Müller (2017a) for international evidence.

Our study is the closest to Ulbricht and Weiner (2005), who compared Compustat and Datastream in the US. They focused mainly on summary statistics for individual items on financial statements, while our study focuses on impacts for a large number of fundamental strategies. The studies are thus similar only in the initial step. The imperfect coverage of micro-caps in the US was also previously documented in Ulbricht and Weiner (2005), but we extend this coverage to international evidence and provide a wide assessment of the impacts of this imperfection. Our study is also related to Ince and Porter (2006) in that we propose new quality screens to shrink errors in Datastream. Analysis in Schmidt et al. (2017) is similar in scope in that they demonstrated how to screen fundamental and return data from Datastream to construct risk factors for 23 countries. They have, however, not focused on the role of imperfect fundamental coverage and their documented screens of data have already been previously published in other studies such as Lee (2011) and Griffin et al. (2010).

Our paper also broadly belongs to a class of studies investigating cross-sectional predictability of individual signals outside the US. See, for example, Chui et al. (2010), Barber et al. (2013), McLean et al. (2009), Rouwenhorst (1998), Lam and Wei (2011), Titman et al. (2013), and Watanabe et al. (2013).

We contribute to the academic literature in four ways. First, we propose new adjustments for the data from Datastream that decrease the number of errors there. These can be applied to similar databases facing the same problems. Secondly, we document that missing delisting returns in Datastream are not creating serious biases when constructing portfolios for a wide range of anomalies. This is one key takeaways of this study as the delisting returns are not available in the international sample and any international study therefore has to tacitly rely on this conclusion. Next, we provide robust evidence that the choice of Compustat as the main database in most of the finance and accounting literature is not a source of serious concern due to possible idiosyncratic errors there. This should be a key takeaway for practitioners and researchers interested in international markets as they often rely on different sources of fundamental data. Finally, we document the importance of coverage of listed stocks in fundamental databases. The coverage is especially important in the international setting where there is no single database with full coverage that spans a long time period. The partial coverage can lead to biased and inconsistent results, especially for stocks with smaller capitalization. This outcome is the main takeaway of our study and should serve as a caveat for international studies where the fundamental data is important.

1.1 Data and Initial Adjustments

One of our sources for data on US stocks is the Merged CRSP/Compustat database from Wharton Research Data Services (WRDS). The sample spans the 1963 to 2016 period and contains all NYSE, Amex, and NASDAQ common stocks (CRSP share code 10 or 11). We adjust the returns for delisting following guidance in Shumway (1997) and Hou et al. (2017).[7]

Our second source of US data, and primary source for international data, is Reuters Datastream (Worldscope). The database manual from 2007 states that: "The total universe of companies contained on the database has grown from approximately 4,000 in 1987, to over 51,100 at March 2007. This includes 33,300 currently active companies in developed and emerging markets, representing approximately 95% of global market capitalization." It should thus provide a good comparison for CRSP/Compustat in the US given its wide coverage. We source individual stocks in each country from both alive and dead lists of stocks to limit survivorship bias. We filter the data following Ince and Porter (2006), Lee (2011), and Griffin et al. (2010). The procedure includes manually checking the names of the shares in the database for over 100 expressions that describe their share class. We leave only the primary quotes of ordinary shares of companies with few exceptions where the fundamental data in Datastream is linked with other share classes.[8] We also exclude all Real Estate Investment Trusts (REIT) We require the return index (RI) to be larger than 0.001 on the first day of the month for higher precision. All the returns in this study are converted to US dollars. We set the RI to missing if the price on the first day of the month is larger than $1 million. We delete daily returns for days when the stock market was closed in a given country.

We use the classification of Fama and French (2017), sorting developed countries into 4 groups: (1) North America (United States and Canada); (2) Europe (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom); (3) Japan; and (4) Asia Pacific (Australia, New Zealand, Hong Kong, and Singapore). The Datastream sample starts in 1990, where there was large enough coverage for the USA, Europe, and Japan. The stocks in individual countries are from the largest exchange in the given country with the exception of the US (NYSE, NASDAQ, and Amex) and Japan (Tokyo and Osaka).

1.1.1 Merging Datastream and Compustat in the US

We need to create a merged database from Datastream and Compustat for further analysis. Accordingly, we merge Datastream and CRSP on their main security level identifiers: DSCD and PERMNO. We do this rather than directly merging Datastream fundamental

[7] Specifically, we use the return over the month if the delisting is on the last day of the month. The relevant delisting return is then added as a return over the next month. Then, we use the delisting return (DLRET) from the monthly file if it is not missing. If it is missing, then we use $(1 + ret_{cum}) * (1 + DLRET_d) - 1$, where ret_{cum} is the cumulative return in the month of delisting and $DLRET_d$ is the delisting return from the daily file. Finally, we fill the gaps with $(1 + ret_{cum}) * (1 + DLRET_{avg}) - 1$, where $DLRET_{avg}$ is the average delisting return for stocks with the same first digit of the delisting code (DLSTCD). Hou et al. (2017) applies the average over the past 5 years, but we found this method to be very noisy and a single large outlier had a huge impact on the average value.

[8] We closely follow the description in Griffin et al. (2010) regarding what shares are not common. We also partially rely on the correct classification of stocks in CRSP, as we keep any stock that can be matched to CRSP by CUSIP and filtered by relevant filters there. This selection procedure is not very important in the current work, as stocks with fundamental coverage in Datastream are not plagued by as many errors or missing categorization compared to those without.

(Worldscope) and Compustat because it leads to a larger number of successfully matched stocks in the two databases. This better match is due to the design of Datastream where static data (for example industry classification or tickers) are separated from time-series data (for example prices). Static data then includes only the latest available entries so that if there are any changes over time, these changes are not recorded. The CRSP and Compustat matching table in WRDS reflects the full history of changes. The fundamental data is related to the company and not only to particular share issues so that changes in the currently most relevant traded share class would cause a problem. DSCD is then related to particular share issue and it is assigned when it enters the Reuters platform, as is PERMNO in CRSP. Merging on DSCD and PERMNO thus leads to more precise results. We then connect Datastream with Datastream fundamental (done automatically by Reuters when downloading the data) and CRSP with Compustat (we use the Merged CRSP/Compustat database from WRDS) in the second stage.

We first connect the databases by the 8 digit Committee on Uniform Security Identification Procedures ticker (CUSIP) and then check if it was successful by comparing the exchange tickers and names in the two databases. We discard a few cases where it is evident that the merge was not successful. We then merge on 6 digit CUSIP and again manually check for the success of the merger. In the end, we get 130,000 merged PERMNO-year observations out of approximately 250,000 in Compustat over the 1980 to 2016 period. See Panel A of Figure 1.1 for the number of firms in Datastream fundamental and Compustat and their merge success rate over time. It is evident from the figure that less than half of all firms in Compustat were in the merged sample in 1980. This level increased to approximately 95% in 2015. Panel B shows merge success rate based on market cap of the stocks. The market cap of successfully merged stocks over market cap of all stock in Compustat is higher than in Panel A suggesting the coverage in Datastream was better for larger stocks.

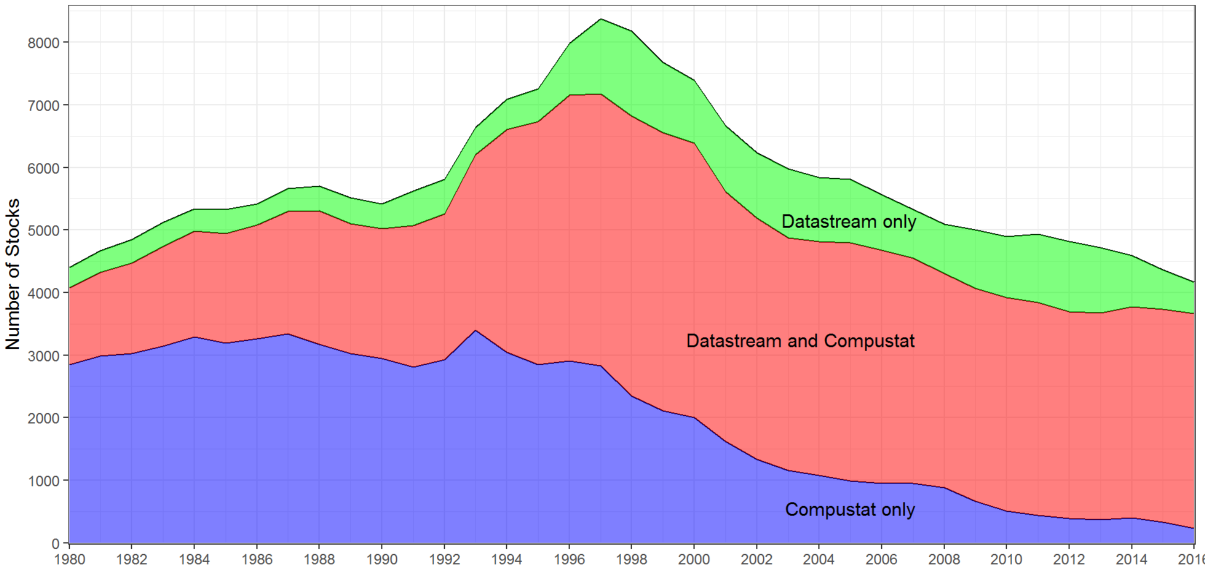
1.1.2 Adjustments of Returns in Datastream

Ince and Porter (2006) provided the first systematic treatment of data quality in the Datastream database. They suggested several adjustments to shrink the size of errors in the database. These adjustments include discarding extreme returns that revert the next month. They also note that dropping stocks with a price lower than $1 decreases the errors, as the mistakes tend to cluster in stocks with a low price. We have at least one decade worth of new data, so we revisit these issues.

Datastream provides stale prices when there is no trade during the day or when the stock is no longer traded so that the price of the last trade is repeated until there is a new trade. We thus delete all observations with stale prices at the end of our sample. We implement a new way to fix returns and prices when there is an event that affects the number of shares outstanding (e.g., stock split), but there are stale quotes of prices at that time.[9] We characterize this event by a concurrent daily return larger than 15%

[9] A natural reaction of price to the 1 to 10 split would be its decrease to 10% of the original price, but

Panel A: Based on Number of Stocks.



Panel B: Based on Market Cap of Stocks.

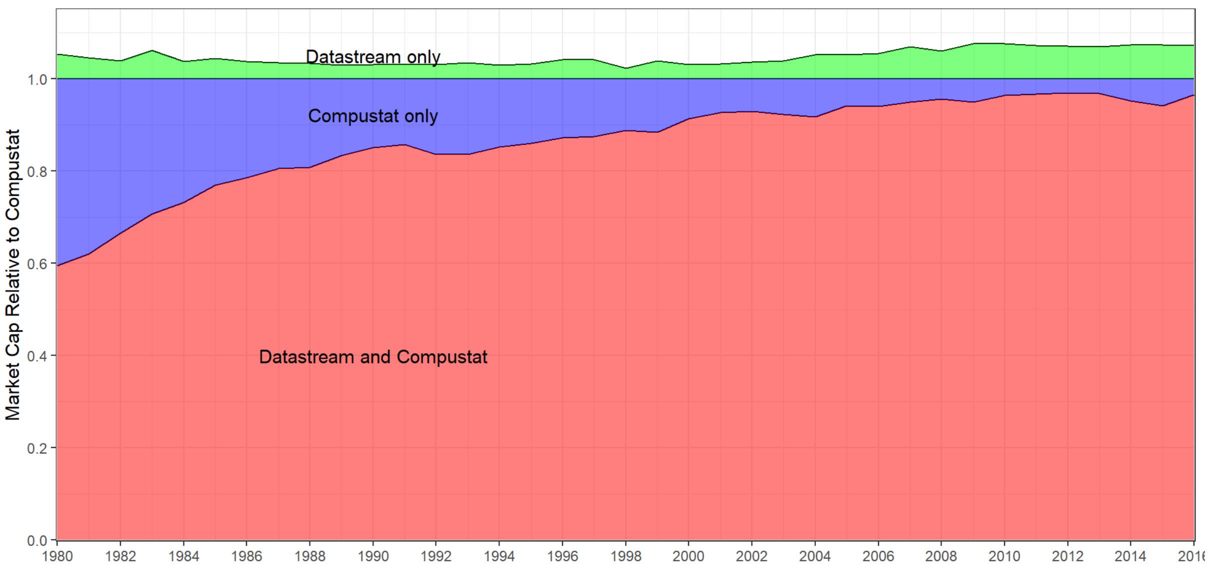


Figure 1.1: Number of Stocks and Their Market Cap with Fundamental Coverage in Compustat and Datastream over Time.

(lower than -15%), an increase in the daily adjustment factor (Datastream variable AR) by 15% (decrease by 15%), and zero volume (if Datastream variable UVO is missing). We delete the latest observations of price with no trading and backfill the correct prices from their first new quote if it arrives in less than 30 days after the event.

Following Ince and Porter (2006), we set as missing those monthly returns over 300% that revert back over the next month. We only discard returns which we failed to correct in our previously described procedure.[10] This adjustment leads to closer returns with respect to CRSP, and we have not found any way to improve it. We also set the RI to missing if the daily return is larger than 500%. We set any monthly return larger than 2000% as missing. There is only one such case in CRSP, but there are many in DS for the US.

Table 1.1 presents correlations between monthly returns in Datastream and CRSP depending on the percent of observations winsorized and the minimum price of a stock at the end of the previous month. We focus on three periods: 1980 to 1989, 1990 to 1999, and 2000 to 2016. We expect that the quality of data will increase over time so that lower adjustment amounts are needed. It is indeed the case and the most recent period does not require any filters or adjustments with 99.6% correlation of the returns. The most successful adjustment in the earliest period is winsorizing the highest and lowest 0.1% of all returns, or approximately 40 stocks, in a given month. We adjust only 0.01%, or approximately 4 observations every month, in the 1990 to 1999 period. There is no need for price filters in the latest period but limiting extreme returns on the stocks with the lowest price helps in the earlier periods. To summarize, we start with adjustments for large daily and monthly returns that revert back by first trying to fix them and then discarding the rest. We then winsorize the resulting returns at different levels depending on the period. Winsorization of the returns does not have a significant impact on our findings but it helps to make the comparison across Datastream and Compustat more robust since the results will not be as easily driven by few outliers.

1.1.3 Construction of Anomalies and Portfolios

To study the role of the source of the accounting information, we primarily focus on the performance of fundamental anomalies. The main reason for this is that it is easy to quantify their differences across databases and this is possible in a systematic way across a large set of published studies. It should also be of the first order importance to any quantitative investor. We have tried to study the largest set of published anomalies possible. We have included all fundamental anomalies that we have found in the literature and that could be implemented in both Compustat and Datastream.[11] Specifically, we

if there has been no trade since the split, the old price is still displayed in Datastream. This outcome results in an incorrectly displayed return of 900%.

[10]Specifically, we set as missing returns for two consecutive months if the return in the first was larger than 300% and the overall return over the two months was lower than 50%.

[11]Some anomalies cannot be replicated with Datastream because it does not contain some needed items. Examples are anomalies based on advertising expense.

Table 1.1:
Quality of Returns in Datastream

The table shows the correlation between returns in Datastream and CRSP in the US depending on the stock price at the end of the previous month and the fraction of returns that are winsorized every month. We separately focus on 3 periods: 1980 to 1989, 1990 to 1999, and 2000 to 2016.

	1980 - 1989				1990 - 1999				2000 - 2016			
Winsorize	All	$.25+	$1+	$5+	All	$.25+	$1+	$5+	All	$.25+	$1+	$5+
None	0.930	0.946	0.961	0.970	0.966	0.972	0.987	0.992	0.996	0.996	0.996	0.996
.01%	0.937	0.950	0.962	0.971	0.973	0.978	0.989	0.992	0.995	0.995	0.996	0.996
.1%	0.953	0.960	0.968	0.976	0.961	0.976	0.987	0.991	0.977	0.979	0.984	0.994
1%	0.935	0.943	0.958	0.974	0.927	0.947	0.967	0.981	0.937	0.942	0.953	0.978

have tried to implement all fundamental anomalies documented in Harvey et al. (2016), McLean and Pontiff (2016), and Hou et al. (2017). We considered 93 anomalies initially, but excluded 19 that we failed to replicate within the original sample of the studies. The final sample therefore constitutes 74 anomalies. We list only the remaining 74 anomalies in our analysis.[12] We have grouped the anomalies into 5 categories and our main analysis then focuses only on these categories. The detailed results are provided in the Appendix D. The groups are: accruals, profitability, value, investment, and intangibles. A detailed list of anomalies is provided in Table A.1 in the Appendix A. A detailed description of how we construct the anomalies is provided in the Appendix B.

We follow the original papers' guidance on the sample construction of individual anomalies. Most of the portfolios on the anomalies are equal-weighted except the cash-based operating profitability of Ball et al. (2016), which is value-weighted. We construct returns on zero-cost portfolios as returns on stocks in the top quintile of each signal minus returns on the bottom quintile of each signal. The portfolios sorted on annual fundamental signals are rebalanced annually at the end of June every year, based on signals from business year ending in the previous calendar year. We also follow the original studies in direction of the anomalies and change sign for the signals where required so that all the anomalies should yield positive returns.[13]

Some anomalies require the classification of industries, such as Hou and Robinson (2006). The choice in the original papers is mostly with respect to Standard Industrial Classification (SIC) industry classification. We apply third level Datastream classification, which sorts industries into 19 groups instead for two main reasons. First, the coverage in Datastream is not the same as that in Compustat and this would create a huge difference for fundamental signals dependent on the industries if there are more than 100 industries. Second, the industry classification in Datastream is available only from the static file, which means that only the latest value is available. Variation over time for individual firms between closely related SIC codes would thus again cause problems. We provide the

[12]The full list of the 93 anomalies is available in Appendix G.

[13]The code for creation of the anomalies has undergone a four eye consistency check relative to the original studies. The discrepancies were further benchmarked with results in Hou et al. (2017).

transition between SIC classification and Datastream classification in the Appendix C.

1.1.4 Methodology for Testing Differences in Returns on the Anomalies

We will now describe the tests used to compare returns on anomalies across two different databases.[14] We test significance of returns on individual anomalies with a simple t-test. We adjust standard errors in the t-test for autocorrelation and heteroskedasticity as in Newey and West (1987) with 12 lags. We then compare number of anomalies significant at 5% level across the two databases. Differences in returns on individual anomalies over the two databases are again tested with a t-test. One caveat here is that the differences tend to be heavily significant even if the economic difference is negligible. The large significance occurs when difference in returns over the two databases are consistent over time, which leads to their small standard error. This arises, for example, for some anomalies when delisting returns are omitted. A better indication of meaningful differences in returns over the databases is economic significance (absolute difference in mean returns) and comparison of size of t-statistics on the anomalies. Different t-statistics can lead to different research inference when the anomaly is significant in one database but not in the other.

We test significance of returns on groups of anomalies in panel linear regressions with only intercept as explanatory variable. We use Driscoll and Kraay (1998) heteroskedasticity and autocorrelation robust errors. The difference in returns of groups of anomalies are again tested in panel setting where the dependent variable is returns on anomalies in one database minus returns in the second database.

1.1.5 Role of Delisting Returns

One shortfall of Datastream, and most of the other sources of returns for equities, is that it does not include the delisting return after the stock is removed from the exchange. Shumway (1997) showed that there could be a large bias in returns on portfolios constructed from CRSP data due to missing delisting returns from performance related delistings at the time of publication of his study. The missing delisting returns have created an upward bias for returns on small cap stocks to the point that one half of size anomaly could be explained by it.

There are several frequent reasons for delisting of a stock which can determine the expected delisting return. Mergers and acquisitions are usually connected to positive delisting returns since the buyer has to pay premium to buy publicly traded shares. Performance related delisting can then lead to heavily negative return depending on success of restructuring of the company. Shumway (1997) precisely showed that missing performance related delisting returns in CRSP tend to be heavily negative when he tracked the true delisting returns in an alternative database. He then suggested that the missing

[14] Alternatively, the same approach is also applies within one database but for across two different ways of how to construct the portfolios.

performance related delisting returns should be filled with -30%, which he estimated as mean delisting return in his alternative dataset.[15] Many authors have then adopted his suggestion in the literature.

The quality of CRSP has increased since 1990s so that most of the delisting returns are no longer missing. There are 20 680 delistings in CRSP, with just 2 742 of them missing as of 2017. We revisit the role of missing delisting returns by investigating the returns on portfolios based on our set of anomalies with various delisting return methodologies. We do not opt for the alternative data source on delisting, as Shumway did, but we will rather compare the returns on the portfolios with all the correct adjustments in CRSP and with completely omitted delisting returns. The goal is to see if excluding the delisting returns, as is tacitly done in Datastream, leads to systematic biases.

Table 1.2 provides the results for the 5 categories of fundamental anomalies.[16] It is apparent that there are some differences, but they are far smaller than what Shumway (1997) suggested. They are not systematic in the sense that they would cluster in certain types of anomalies, with the exception of some profitability anomalies that tend to short stocks that go bankrupt with negative delisting returns. Omitting delisting returns then leads to approximately 5% lower estimated returns on them. The differences are small even for size and liquidity anomalies, where they are expected to be the largest. We can therefore conclude that omitting delisting returns is not a cause for serious concern when using Datastream, and other factors play a far larger role. This is a different conclusion with respect to Shumway (1997), but it is hardly surprising. The average return over all delistings that were performance related is very close to zero in our sample, which is strikingly different from the -40% found in his study. His recommendation was to substitute the missing delisting returns for performance reasons by -30% return, which we do in our second comparison in the table.[17] The difference in returns is again tiny and the choice of how to adjust for delisting returns is thus not important.[18]

1.2 Similarity of Financial Statements

We start our comparison of Compustat and Datastream by looking at raw financial statements. The corresponding items between fundamental databases should be very similar as most of the items can be obtained without any adjustment directly from statements

[15] The most cautious approach for long-only portfolios would be to set the missing delisting returns to -100% which would provide the most adverse conditions for portfolio returns possible. The situation is, however, more complex for long-short portfolios since the performance related delistings could be clustered in short leg of the portfolios. One example of such strategy is profitability related anomalies.

[16] The detailed results for each anomaly are provided in the Appendix D.

[17] Delistings for a performance reason have the delisting codes: 500, 520, 551 to 574, 580, and 584 in CRSP.

[18] We have also tried several ways to interpolate the data on delistings from CRSP, but it did not lead to any meaningful improvements relative to omitting the delisting returns. It is possible to sort delistings in Datastream into several categories based on what is included in the names of the shares. Approximately half of all delisted stocks have some indication added to their name, such as 'DELIST' or 'MERGER'. Matching relevant firms in CRSP and computing the average delisting return for the categories, however, yields an average return that is close to zero.

Table 1.2:
Impact of Omitting Delisting Returns in CRSP
The tables show returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. We compare two ways of adjusting for delisting returns with respect to our adjustment. The first one is with all delisting returns set equal to zero and the second one follows Shumway (1997). We also show the correlation between portfolios in the two comparisons. The list of anomalies is provided in Appendix A. The source of fundamental data is Compustat. The portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The sample period is July 1963 to December 2016. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

	Our delisting Adjustment Versus No Delisting				Our Delisting Adjustment vs Shumway (1997)			
	Corr	Our	No Delisting	Diff	Corr	Our	Shumway	Diff
Accruals	0.998	0.53 (6.19)	0.54 (6.35)	(3.16)	0.999	0.53 (6.19)	0.53 (6.12)	(-1.89)
Intangibles	0.999	0.40 (4.23)	0.41 (4.32)	(3.33)	1.000	0.40 (4.23)	0.40 (4.19)	(-2.05)
Investment	0.999	0.48 (8.97)	0.48 (8.99)	(0.88)	1.000	0.48 (8.97)	0.48 (8.96)	(-1.70)
Profitability	0.999	0.38 (3.95)	0.36 (3.78)	(-6.17)	1.000	0.38 (3.95)	0.38 (4.01)	(3.30)
Value	1.000	0.66 (5.66)	0.66 (5.71)	(2.03)	1.000	0.66 (5.66)	0.66 (5.66)	(-0.00)
All	0.999	0.49 (9.07)	0.50 (9.15)	(1.61)	1.000	0.49 (9.07)	0.49 (9.04)	(-1.30)

provided by the companies in their regulatory filings. This, however, is not necessarily the case. We show that specific methodologies chosen by the data vendors can lead to large differences. We focus on reduced versions of the financial statements that include only items that were used in the construction of signals for fundamental anomalies in our reviewed literature. This is only a fraction of the variables, as there are 151 items in financial statements in Datastream with wide coverage from 1995 and over 200 items in Compustat. We focus only on the most important subset for the sake of brevity and because it is often difficult to find close matches for the other variables.

Table 1.3 shows the time series averages of cross-sectional Pearsons and Spearman's correlations between items in the two databases. We also specify how we construct the corresponding items in Datastream in the last column. Some transitions can be done directly by simply matching items, but others have to be done by more complicated transformations. There are some visible patterns in the discrepancies between the databases. First, variables in the current working capital that are part of accruals tend to differ a great deal. Next, there are differences in the classification of leases in Property Plant and Equipment and the classification of long-term versus short-term debt. This is due to the different methodologies of data vendors and their interpretations of the raw statements provided by companies. Other notable differences are among the items in financing cash flows. This is again due to different methodologies by the vendors. To conclude, there are some notable differences across the databases that could create a systematic bias for the fundamental signals constructed from them.

Table 1.3:
Variables from Compustat Mapped onto Datastream

The table shows all fundamental variables that were required for construction of our fundamental anomalies. We first specify their name in Compustat and then document how we construct them in Datastream. We also show Pearson's and Spearman's correlation coefficients between variables in the two databases in our merged sample. The sample spans from January 1989 to December 2016.

		Pearson	Spearman	
BALANCE SHEET				
ASSETS				
Current Assets				
Cash and Short-Term Investments	CHE	0.619	0.990	WC02001
Short-Term Investments	IVST	0.556	0.764	WC02008
Receivables - Total	RECT	0.770	0.984	WC02051
Inventories - Total	INVT	0.824	0.972	WC02101
Current Assets - Other - Total	ACO	0.804	0.964	WC02149 + WC02140
Prepaid Expenses	XPP	0.912	0.911	WC02140
Current Assets - Total	ACT	1.000	1.000	WC02201
Non-Current Assets				
Long-Term Investments	IVAO	0.866	0.745	WC02258 + WC02250
Property Plant and Equipment - Total (Net)	PPENT	0.993	0.997	WC02501
Property Plant and Equipment - Total (Gross)	PPEGT	0.997	0.998	WC02301
Property Plant and Equipment Buildings at Cost	FATB	0.997	0.993	WC18376
Property Plant and Equipment Leases at Cost	FATL	0.771	0.754	WC18381
Investment and Advances - Equity	IVAEQ	0.941	0.846	WC02256
Intangible Assets - Total	INTAN	0.994	0.966	WC02649
Goodwill	GDWL			Set equal to 0
Assets - Total	AT	0.982	1.000	WC02999
LIABILITIES AND SHAREHOLDERS' EQUITY				
Current Liabilities				
Debt in Current Liabilities	DLC	0.961	0.953	WC03051
Account Payable/Creditors - Trade	AP	0.884	0.993	WC03040
Current Liabilities - Other - Total	LCO	0.952	0.991	WC03066 + WC03054 + WC03063 + WC03061
Accrued Expenses	XACC			Set equal to 0
Income Taxes Payable	TXP	0.937	0.860	WC03063
Current Liabilities - Total	LCT	1.000	0.999	WC03101
Long-Term Liabilities				
Long-Term Debt - Total	DLTT	0.985	0.988	WC03251
Liabilities - Other	LO	0.633	0.892	WC03273 + WC03262
Liabilities - Total	LT	0.998	0.998	WC03351
Minority Interest - Balance Sheet	MIB	0.763	0.791	WC03426
Shareholders' Equity				
Preferred/Preference Stock (Capital) - Total	PSTK	0.816	0.898	WC03451
Retained Earnings	RE	0.994	0.990	WC03495
Shareholders' Equity - Total	SEQ	0.995	0.999	WC03501 + WC03451
Common/Ordinary Equity - Total	CEQ	0.995	0.998	WC03501
Deffered Revenue Current	DRC			Set equal to 0
Deffered Revenue Long-Term	DRLT	0.307	0.683	WC03262
Preferred Stock Redemption Value	PSTKRV	0.877	0.914	Set equal to PSTK
Preferred Stock Liquidating Value	PSTKL	0.878	0.914	Set equal to PSTK

1.3 Performance of Anomalies in the Same Sample

The previous section has suggested some large differences in financial statements across the two databases. We will now investigate whether these differences translate into returns

Table 1.3 Continued

		Pearson	Spearman	
INCOME STATEMENT				
Revenue - Total	REVT			Set equal to SALE
Sales/Turnover (Net)	SALE	0.999	0.999	WC01001
Cost of Goods Sold	COGS	0.990	0.969	WC01051
Selling, General and Administrative Expenses	XSGA	0.989	0.982	WC01101
Research and Development Expense	XRD	0.986	0.983	WC01201
Earnings Before Interest, Taxes & Depreciation	OIBDP	0.963	0.983	WC01151 + WC01250
Depreciation and Amortization - Total	DP	0.989	0.992	WC01151
Earnings Before Interest and Taxes	OIADP	0.925	0.971	WC01250
Interest and Related Expense	XINT	0.885	0.993	WC01251
Pretax Income	PI	0.994	0.992	WC01401
Income Taxes - Total	TXT	0.997	0.995	WC01451
Income Before Extraordinary Items	IB	0.995	0.990	WC01551
CASH FLOW STATEMENT				
Indirect Operating Activities				
Operating Activities - Net Cash Flow	OANCF	0.990	0.996	WC04860
Investing Activities				
Capital Expenditures	CAPX	0.976	0.992	WC04601
Investing Activities - Net Cash Flow	IVNCF	0.990	0.994	- WC04870
Financing Activities				
Purchase of Common and Preferred Stock	PRSTKC	0.981	0.967	WC04751
Sale of Common and Preferred Stock	SSTK	0.928	0.960	WC04251
Cash Dividends	DV	0.998	0.992	WC04551
Dividends on Common Stock	DVC	0.987	0.985	WC05376
Long-Term Debt - Issuance	DLTIS	0.946	0.944	WC04401
Long-Term Debt - Reduction	DLTR	0.915	0.948	WC04701
Net Changes in Current Debt	DLCCH			WC04821
Financing Activities - Net Cash Flow	FINCF	0.987	0.991	WC04890
OTHER ITEMS				
Book Value per Share	BKVLPS	0.921	0.982	WC05476
SIC Industry Classification	SIC			WC07023
Earnings per Share	EPSPX	0.956	0.983	WC05210
Earnings per Share after Extraordinary Items	EPSPI	0.942	0.987	WC05230
Employees	EMP	0.937	0.992	WC07011
Net Income	NI			Set equal to IB
Preffered Dividends in Arrears	DVPA			Set equal to 0
Treasury Stock - Preferred	TSTKP			Set equal to 0

on the anomalies that are constructed from them. We start with a comparison within the sample of stocks that can be matched between the two databases in this section and follow with full samples in the individual databases in the next one.

We test the differences in two settings. First, we compare the similarities in the fundamental signals themselves, and then we turn to the returns on portfolios created based on them. Panel A of Table 1.4 first looks at time series average of cross-sectional correlations between signals created from either Compustat or Datastream. Pearson's correlations can be very low for some signals, but the similarity in rankings based on the signals are much higher, with an average Spearman's correlation of 93.9%. This is mainly caused by outliers where few observations can completely dominate the correlations. The signals tend to have large tails and non-normal distribution so ranks are better at capturing the dependence structure.

Table 1.4:
Datastream versus Compustat in the Common Sample
The tables shows returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. We consider 3 cases for the comparison. First, we compare portfolios created with CRSP & Compustat or with just Datastream in Panel A. We then decompose the overall difference in Panel B by using CRSP returns for both sources of fundamental data or Compustat fundamental signals for both sources of data on returns. We also show the correlation between the two cases. The list of anomalies is provided in Appendix A. The source of fundamental data is either Compustat (CS) or Datastream (DS). The portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The sample period is July 1990 to December 2016. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

Panel A: Compustat with CRSP or Full Datastream

	Signals		Portfolios			
	Pears Corr	Spear Corr	Corr	CT	DS	Diff
Accruals	0.917	0.934	0.950	0.60 (4.87)	0.59 (4.90)	(-1.28)
Intangibles	0.809	0.881	0.909	0.51 (3.49)	0.51 (3.08)	(-0.01)
Investment	0.905	0.960	0.977	0.37 (5.54)	0.36 (5.49)	(-1.02)
Profitability	0.837	0.951	0.964	0.01 (0.05)	0.00 (0.02)	(-0.36)
Value	0.707	0.972	0.994	0.64 (3.44)	0.62 (3.31)	(-2.50)
All	0.841	0.939	0.959	0.45 (6.07)	0.44 (5.96)	(-1.24)

Panel B

	CRSP Returns				Compustat Signals			
	Corr	CT	DS	Diff	Corr	CT	DS	Diff
Accruals	0.957	0.60 (4.87)	0.60 (4.91)	(-0.92)	0.990	0.60 (4.86)	0.60 (4.84)	(-1.21)
Intangibles	0.912	0.51 (3.49)	0.51 (3.03)	(0.02)	0.998	0.51 (3.50)	0.50 (3.51)	(-0.79)
Investment	0.982	0.37 (5.54)	0.36 (5.54)	(-1.20)	0.994	0.36 (5.54)	0.36 (5.47)	(-0.05)
Profitability	0.969	0.01 (0.05)	0.00 (0.02)	(-0.37)	0.995	0.01 (0.06)	-0.00 (-0.01)	(-1.09)
Value	0.995	0.64 (3.44)	0.62 (3.35)	(-2.17)	0.997	0.64 (3.42)	0.63 (3.38)	(-1.47)
All	0.963	0.45 (6.07)	0.44 (6.00)	(-1.01)	0.994	0.45 (6.07)	0.44 (5.98)	(-1.34)

The rest of Panel A presents the discrepancies in the returns of the portfolios created either with CRSP and Compustat or with Datastream only. Panel B then decomposes the differences in the returns of the portfolios into two components created either by differences in returns (Compustat signals) or differences in signals (CRSP returns) across the two data sources. We do this by matching the fundamental signals from both Compustat and Datastream with the returns from CRSP. Alternatively, we take the fundamental signals from Compustat and merge them with the returns from either CRSP or Datastream. We then create portfolios and compare their returns. The table shows that there are some discrepancies for some signals, but they do not lead to any systematic biases. The lowest differences are in the value category, with a 99.5% average correlation occurring between the portfolios in this category. The largest differences are for intangibles. It is evident that the returns from both CRSP and Datastream provide almost identical portfolios for the same fundamental signals. This is documented by their average correlation of 99.4%. There are no strong systematic differences across the anomalies. The differences in quantitative portfolios between the two databases are therefore mainly due to distinct fundamental signals in each of them.

One thing to notice is that the average return on profitability anomalies is not positive for the joint sample of stocks from Datastream and Compustat. We will cover this discrepancy in greater detail in a Section 1.6.1 later.

1.4 Performance of Anomalies in Separate Samples

We now turn to problems with distinct samples that emerge when Datastream and Compustat are not matched. That is, we look at differences across the two databases if portfolios are created solely from the data in each of them. We first start with the US and then widen the scope to international markets in the next section.

Table 1.5 compares the performance of the fundamental anomalies in the two databases without restriction on their joint coverage. We first focus on the case when there are no further filters on the universe of stocks and then try to test if the differences are smaller with some filters. The average return for all the anomalies is practically the same in Datastream and Compustat. The average return is, however, not similar across all the categories. The returns on profitability anomalies drop the most with their average return going from 0.36% to -0.01% monthly and the average t-statistic on individual anomalies going from 1.41 to 0.35. The t-statistic on returns of all profitability anomalies goes from 1.43 to -1.09. The difference between returns in Compustat and Datastream is significant at 1% level. The changes in other categories are statistically significant only for Intangibles, but changes in individual anomalies can be substantial in all the groups. A large difference is, for example, in operating profitability over assets, which would yield 0.93% monthly according to Compustat but only -0.10% monthly according to Datastream. This difference is significant at the 0.05% level.

43 of the anomalies have a difference in the mean returns that is significant at the 5% level. There are 41 significant anomalies with Compustat and CRSP and 39 with Datastream. This is the same as in the common sample, but there are only 29 anomalies that are significant in both databases. Thus, one-quarter of all the anomalies cannot be consistently replicated across the two databases. This leads us to conclude that both databases can convey substantially different results due to their different coverage and classification of stocks when one considers individual anomalies. The differences are, however, much smaller if one focuses on groups of anomalies.

We next try to look at a reduced set of stocks that would suffer from smaller disparities. Figure 1.1 has documented that the coverage on Datastream was not ideal in earlier periods, especially for small stocks. Reuters provides different depths of fundamental coverage for companies in Datastream. Smaller companies that do not meet certain criteria are available only with a reduced set of items on their financial statements and all anomalies thus cannot be constructed for them. The Worldscope manual reports that \$100 million market capitalization is the required threshold for the full coverage in some regions. This could be binding, especially historically. There are also differences in the way that Datastream and Compustat treat financial firms. The financial firms in Datastream have a special template for their financial statements, which is comprised of items that are different relative to industrial firms. This could lead to problems, as some signals cannot be constructed for them. Another important factor, which we consider, is time, as the coverage in Datastream has improved steadily.

We therefore provide results for a restricted sample that contains only non-financial stocks with capitalization over \$100 million and that spans the 2000 to 2016 period. The \$100 million capitalization requirement is very similar to discarding the stocks with a size lower than the bottom decile in the NYSE, which has been widely used throughout the literature.[19] We then construct the fundamental signals on this reduced sample but create portfolios only from July 2010. Specifically, we censor all fundamental information from the time when the capitalization was lower than \$100 million and before 2000 so that the signals are constructed only using a similar information set. This leads to samples in Compustat and Datastream that are very similar in size, and there are no obvious biases across capitalization quintiles in Datastream.

It is evident that the similarity of portfolios has increased, with the average correlations between returns increasing from 80.2% to 90.3%, but the differences remain substantial for some anomalies. 90.3% is still substantially smaller than 95.9% for stocks matched in the common sample, which implies that the classification of stocks in individual databases can have a substantial impact. The large difference in operating profitability over assets has virtually disappeared and would yield a 0.51% monthly average return according to Compustat and 0.44% according to Datastream. There are still 14 anomalies with differences in returns across the two databases that are significant at the 5% level. Significant anomalies again differ across the two databases. There are 6 significant anomalies with Compustat and 8 with Datastream, but only 4 of those are common across the two databases.

1.4.1 What Drives the Differences?

We now study in more detail whether the missing fundamental coverage for stocks with smaller market capitalization can explain the discrepancy in the profitability of anomalies across the two databases. Figure 1.2 maps the proportion of stocks within a given size quintile in CRSP that has fundamental coverage in Datastream. We also include the lowest size quintile in Compustat for comparison. It is evident that the coverage has been very uneven over time and for different size quintiles. The smallest half of stocks suffered

[19] See, for example, Hou et al. (2017) and Green et al. (2017).

Table 1.5:
Datastream versus Compustat in Their Own Full Samples
The table shows the returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. We compare portfolios created with CRSP & Compustat or with just Datastream for either all available stocks or for a reduced sample. The full sample starts in July 1990 and ends in December 2016. The reduced sample begins in July 2010 and omits all financial stocks or those with capitalization under $100 million. We also show correlation between the two cases. The list of anomalies is provided in Appendix A. The portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

	Full Samples				Cap over $100 million & No Financial & 2010+			
	Corr	CT	DS	Diff	Corr	CT	DS	Diff
Accruals	0.762	0.56 (4.39)	0.62 (5.05)	(1.03)	0.883	0.08 (0.79)	0.10 (0.94)	(0.47)
Intangibles	0.714	0.41 (2.55)	0.58 (3.48)	(3.81)	0.852	0.06 (0.61)	0.04 (0.42)	(-0.42)
Investment	0.815	0.49 (6.89)	0.46 (6.36)	(-0.92)	0.909	0.17 (2.14)	0.15 (1.85)	(-0.85)
Profitability	0.841	0.36 (2.11)	-0.01 (-0.04)	(-6.66)	0.915	0.36 (2.96)	0.26 (1.99)	(-2.45)
Value	0.899	0.64 (3.95)	0.63 (3.39)	(-0.24)	0.968	0.35 (2.05)	0.40 (2.34)	(2.63)
All	0.802	0.50 (7.26)	0.48 (6.41)	(-0.63)	0.903	0.19 (2.69)	0.18 (2.56)	(-0.55)

from insufficient fundamental coverage until 2000, and the full coverage only occurred around 2010.[20]

Figure 1.3 further maps a smoothed histogram of the market cap of stocks with fundamental coverage in Compustat and Datastream in 1990 and 2015. It is apparent that the insufficient coverage in Datastream was throughout the whole distribution in 1990 but has virtually disappeared by 2015. There is thus no simple rule regarding how to discriminate based on size to eliminate all the differences in returns on the anomalies.

Table 1.6 tries to explain the differences in returns on anomalies across the two databases. We focus on the full samples without restrictions. We regress the difference in returns on the average cross-sectional quantile of the size of stocks in the respective portfolios. The quantiles are taken with respect to all the stocks in CRSP or Datastream. We also regress the differences in returns on differences in average size. The regression is a simple pooled OLS with standard errors clustered on time periods and anomalies. Both size and difference in size are significant at the 5% level, both individually and jointly. The table thus documents that size is indeed important in explaining the differences and returns on anomalies that are more prevalent in larger stocks, which tend to differ less across the two databases.

1.4.2 Sources of Bias in the Portfolio Returns

There would be no problems with the imperfect fundamental coverage if the stocks would be omitted randomly. The problem is that the coverage is not random, as documented

[20]Note that the proportion for some quintiles in Datastream is larger than 100% in 2010s. This is due to different classification of common stocks in Datastream. There are therefore more stocks with fundamental coverage in Datastream than there are common stocks in CRSP.

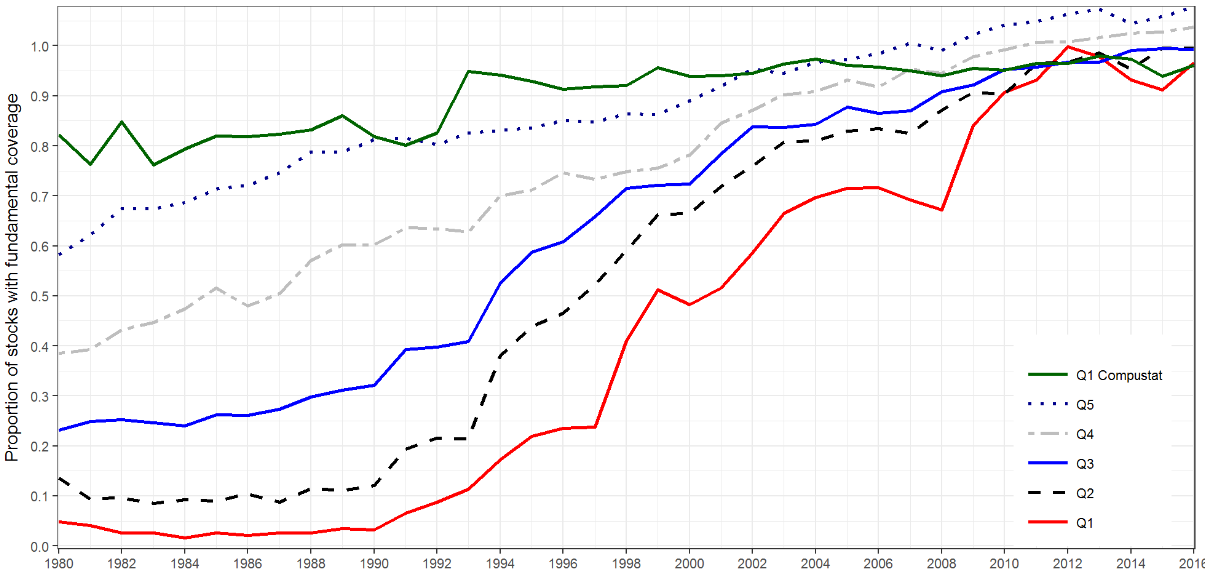


Figure 1.2: Fraction of Stocks in CRSP with Fundamental Coverage in Compustat or Datastream in a Given Size Quintile.

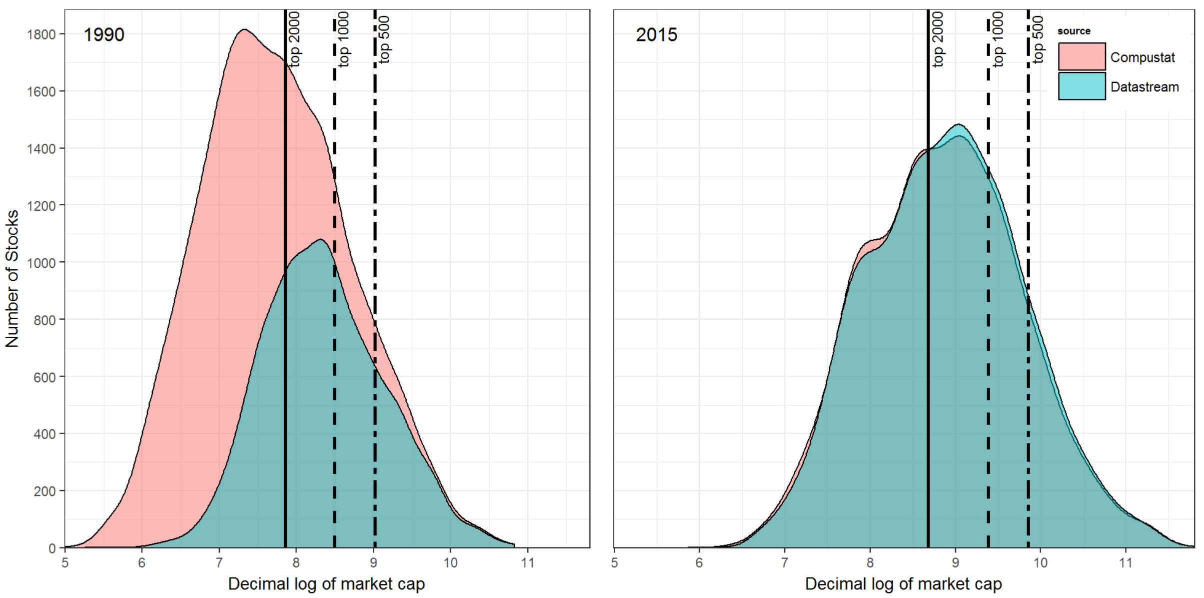


Figure 1.3: Histogram of Market cap of Stocks with Fundamental Coverage in Compustat and Datastream.

Table 1.6:
Explaining the Difference in Returns across Datastream and Compustat
The table shows the results from regressions of differences in the returns of portfolios from alternative databases. The portfolios are created from sorts on fundamental anomalies constructed with data from either CRSP and Compustat or with just Datastream. We then regress the monthly returns from Datastream minus the returns from Compustat on size in Compustat or the difference in size across the two databases. The size is measured as the mean cross-sectional quantile of the size of stocks in the portfolio with respect to the full universe of US stocks at the beginning of each month. The list of anomalies is provided in Appendix A. The portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The sample period is July 1990 to December 2016. The standard errors in regressions are clustered at time and anomaly effects.

	I	II	III
Intercept	-1.48	-0.06	-1.44
	(-3.44)	(-0.67)	(-3.25)
Size	2.69		2.79
	(3.35)		(3.42)
Difference in Size		-1.26	-1.37
		(-2.19)	(-2.50)
R^2	0.0038	0.0016	0.0056

earlier. There are three main sources of the biased returns on portfolios, and we will now cover them in detail.

Firstly, breakpoints on the portfolios are biased since the covered sample of stocks is not randomly sampled from the full population of stocks. The breakpoints are therefore valid only for a given database and not for the full population of the stocks or for other databases. The weighted average of stock returns for a subpopulation bounded by incorrectly specified breakpoints is biased if the bias in breakpoints is related to stock returns. That is, if the biased breakpoints cause omission or addition of stocks with different average return with respect to what subpopulation average for the given portfolio is. We will show that the likelihood of the fundamental coverage in Datastream depends on company size and number of analysts following among other factors. Both size and a number of analysts following has been linked in the literature to stock returns, see Banz (1981) and Elgers et al. (2001).[21] Interactions between the anomalies and the variables driving the coverage is a source of another bias. Fama and French (1992) and Fama and French (2015), for example, document interactions of size and book to value, investments, and profitability. Bias coming from inappropriate breakpoints can be minimized by using breakpoint from all-but-microcaps subpopulation of the stocks where there are only mild

[21] The size premium is almost non-existent since 1990 after accounting for all the biases. The main source of bias is bid-ask spread jump as analyzed in Asparouhova et al. (2010) that is present for equal-weighted returns. Asparouhova et al. (2010) propose ways how to limit the bias. Equal-weighting is nonetheless followed in this study to provide results mimicking the reviewed academic anomalies research where the bid-ask jump issue is always ignored.

coverage issues.

Secondly, imperfect coverage for stocks within a given subpopulation bounded by correct breakpoints can be a source of more bias. Suppose that it is possible to precisely specify population breakpoints and the bias discussed in the previous paragraph is completely dissolved. Non-random sampling could still cause problems if the likelihood of stocks omission is related to their expected returns. The argument for the bias is especially strong for interaction effects with size. Smaller stocks tend to be more illiquid and harder to trade in a significant quantity which limits the arbitrage opportunity. Any anomalies due to market frictions should therefore be stronger for the small cap stocks which creates interaction effects with size and problems with the non-random sampling.

Lastly, idiosyncratic differences across the databases can be a source of some bias. Classification of industries and treatment of static and time-series information are good examples. This aspect of the bias can be minimized only through specific treatment in the individual cases.

1.4.3 Portfolio Constructions Limiting the Discrepancies

Is there any way to decrease the differences by choosing an appropriate methodology? This is not very important in the US, but it is of first order importance for international studies since Datastream is the most widely used database there. Figure 1.4 showed that there is a lower discrepancy in coverage for larger stocks. Specifically, the coverage for the 1000 largest stocks is very similar across the databases. We will now look at procedures that filter the universe of stocks based on their size to lower the bias.

Table 1.7 presents the returns and t-statistics for value-weighted portfolios constructed on a all-but-microcaps universe or with portfolio breakpoints from the largest 1000 stocks. The all-but-microcaps universe is defined by stocks with a capitalization larger than that of the smallest decile at the NYSE. The logic behind the first adjustment is to truncate the whole distribution of stocks and discard the part where the difference is the largest. This should not cause any serious problems for measurement of profitability for implementable and scalable strategies as the small stocks constitute only a very small proportion of the overall capitalization of the whole market and it is advocated, for example, in Hou et al. (2017). The second adjustment then again shifts the focus to all-but-microcaps but does not discard the other stocks. The breakpoints based on the largest 1000 stocks and value-weighting guarantees that the largest stocks will dominate the returns of the portfolios. The use of breakpoints on all-but-microcapss is very similar to the use of NYSE breakpoints, which has been applied in many studies and is advocated, for example, in Fama and French (2017).

Both methods lead to significant improvement in the correlation of portfolios across the two databases and provide very similar results. The average correlation has increased from 80.2% to approximately 86%. The discrepancy for the returns on profitability anomalies is now much lower as well, and the average absolute difference in the t-statistics on individual anomalies decreased to almost one third. The difference in the inference on significance of individual anomalies remains substantial nonetheless. There are 11 significant signals in Compustat and 12 in Datastream, but only 6 of those are common across the two databases for the all-but-microcaps universe of stocks. There are 9 significant signals in Compustat and 9 in Datastream, but only 6 of those are common across the two databases for breakpoints based on the 1000 largest stocks. This is an even larger difference in relative terms with respect to considering all the stocks.

Table 1.7:
Portfolio Constructions Reducing the Discrepancy between Databases
The table shows the returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. We compare the portfolios created with CRSP & Compustat or with just Datastream for either the all-but-microcaps universe of stocks or for the full sample of stocks with breakpoints from the largest 1000 stocks. The full sample includes all available stocks while the all-but-microcaps universe is restricted to stocks with capitalization larger than that of bottom decile at NYSE. We also show the correlation between the two cases. The sample starts in July 1990 and ends in December 2016. The list of anomalies is provided in Appendix A. The value-weighted portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

	All-but-microcaps VW				Breakpoints from 1000 Largest Stocks VW			
	Corr	CT	DS	Diff	Corr	CT	DS	Diff
Accruals	0.858	0.23 (2.16)	0.22 (2.03)	(-0.27)	0.868	0.21 (2.10)	0.22 (2.11)	(0.29)
Intangibles	0.762	0.17 (2.41)	0.26 (2.98)	(2.38)	0.778	0.21 (3.14)	0.23 (2.73)	(0.54)
Investment	0.851	0.26 (3.31)	0.23 (2.89)	(-0.89)	0.854	0.22 (2.99)	0.21 (2.62)	(-0.65)
Profitability	0.869	0.24 (2.06)	0.17 (1.34)	(-1.26)	0.875	0.22 (2.39)	0.15 (1.42)	(-1.51)
Value	0.948	0.19 (1.09)	0.23 (1.27)	(1.14)	0.942	0.19 (1.03)	0.20 (1.09)	(0.49)
All	0.857	0.22 (3.02)	0.22 (2.99)	(0.32)	0.863	0.21 (2.96)	0.20 (2.79)	(-0.44)

1.5 Implications for Studies of International Markets

We have shown that fundamental coverage in Datastream in the US is not complete and this can have a large consequence on the measurement of performance of the anomalies. We will now focus on its coverage in different countries, as it is often the first database that researchers go to for international data. The US evidence serves as a great testing ground because it includes a large number of stocks, and its implications should be valid elsewhere as well. It is thus important to study imperfections in the coverage, as they could lead to biased estimates in these studies.

1.5.1 Fundamental Coverage Around the Globe

Figure 1.4 presents a fraction of stocks with fundamental coverage depending on the size quintile in Japan, Europe, and Asia Pacific. It is evident that the imperfect coverage is as much present internationally as it is in the US. We next look for support of this imperfect fundamental coverage in Datastream and guidance regarding what patterns to expect from its manual. The Worldscope's manual states that: "In 1987, Worldscope established a second research center in Shannon, Ireland, to maintain and develop the database. In 1995, Worldscope established a third major research and data collection center in Bangalore, India. A fourth major research and data collection center in Manila, Philippines was added with Primark's 1999 acquisition of the Extel company database.... Today, the database operations group, which supports the Worldscope database, employs over 500 people mainly located in 3 collection centers located in Bangalore (India), Shannon (Ireland), and Manila (The Philippines)." It is thus very likely that the quality of data has been changing over time as new research centers have been established. We show precisely this in Figure 1.4. The coverage in Australia, New Zealand, Hong Kong, and Singapore was very uneven until 2001 and is close to 100% after that. Similarly, in Japan, Datastream fully covered only companies with large capitalization until 1998. The coverage is not complete in few European countries even as of 2017, but companies outside the lowest size quintile are generally fully covered from 1997. This is partly due to the inclusion of stocks outside the primary trading venue in each stock exchange. These stocks tend to be very illiquid and have only tiny market capitalization. They are thus not a source of serious concern, as any quantitative investor would exclude them from their investment universe anyway.

1.5.2 Determinants of the Coverage

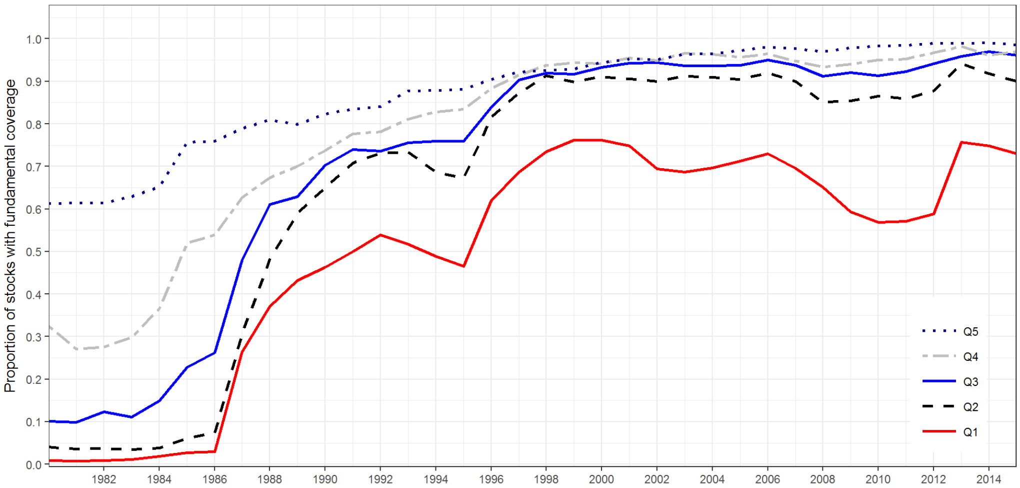
We have shown that the dependence of fundamental coverage on the market cap of individual stocks can have an impact on the measurement of performance of individual anomalies in the case of the US. Are there any other confounding variables that a researcher should be aware of? The Worldscope manual from 2007 describes its content coverage in the following way: "A fully detailed analysis is required for all companies within the following countries: the United Kingdom, and the U.S. For all other countries, fully detailed analysis is required if any of the following criteria is fulfilled:

- Company is a constituent of the, FTSE ALL World, Dow Jones Global, MSCI World, MSCI EMF, S&P Global, S&P/Citigroup or a selected local index.
- Company has 5 or more broker estimates.
- Company has a market capitalization of greater than 100 million dollars (exception Japan, China & Taiwan).
- Legacy companies from Extel database[22]."

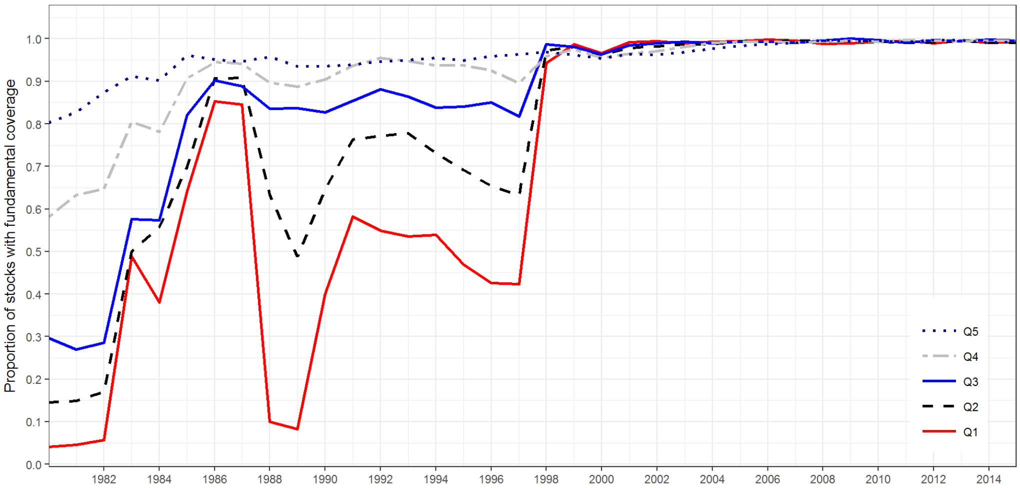
This description suggests that the number of analysts following can have a role very similar to size if it is related to expected returns on individual stocks. Elgers et al. (2001) show that this is indeed the case. Constituency in the indexes is more difficult to measure, but it is usually closely connected to size, which will capture most of its effect.

[22]The Extel database was acquired by Worldscope in 1999 and covered stocks in Asia.

Panel A: Europe.



Panel B: Japan.



Panel C: Asia Pacific.

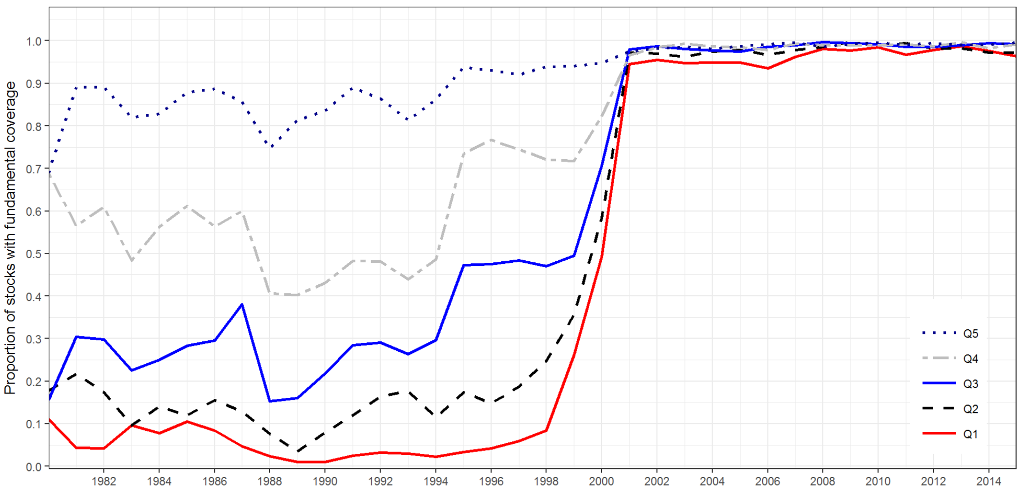


Figure 1.4: Fraction of Stocks with Fundamental Coverage in Datastream in a Given Size Quintile.

Table 1.8 presents logit regressions predicting fundamental coverage with the size quantile and analyst followings in individual countries

$$\begin{aligned} Fundamental\ Coverage_{it} &= \beta_0 + \beta_1 \mathbb{1}\{Size_{it} > \$100M\} \\ &+ \beta_2 \mathbb{1}\{Analysts\ Following_{it} \geq 5\} + \beta_3 Size\ Quantile_{it} \\ &+ \beta_4 (Size\ Quantile_{it} - Size\ Quantile_t^{\$100M}) \mathbb{1}\{Size_{it} > \$100M\} + \epsilon_{it}. \end{aligned} \quad (1.1)$$

where $Fundamental\ Coverage_{it}$ is equal to one when the given stock i has fundamental information in Datastream in month t and zero otherwise. $\mathbb{1}\{Size_{it} > \$100M\}$ is a dummy that is equal to one when market cap of the stock i in month t is larger than 100 million US Dollars and zero otherwise. $\mathbb{1}\{Analysts\ Following_{it} \geq 5\}$ is a dummy equal to one when number of analysts following the company i in month j in I/B/E/S is larger than or equal to 5 and zero otherwise. $Size\ Quantile_{it}$ describes cross-sectional quantile (0 to 1) of market capitalization of the stock i relative to all the stocks in country of its listing in month t. $(Size\ Quantile_{it} - Size\ Quantile_t^{\$100M})\mathbb{1}\{Size_{it} > \$100M\}$ captures change in slope of the cross-sectional market cap quantile when the market cap is larger than 100 million US Dollars. Cross-sectional market capitalization quantile of stock with 100 million US Dollars $Size\ Quantile_t^{\$100M}$ is subtracted from $Size\ Quantile_{it}$ so that $\mathbb{1}\{Size_{it} > \$100M\}$ retains its interpretation and is not distorted by the possible kink in predictive impact of $Size\ Quantile_{it}$. The regressions are separately estimated on all available stocks in Datastream in each individual country in 1990 to 2002 period and 2003 to 2016 period. The fundamental coverage is worse in the earlier period. Change in coefficients in the later period helps illuminate if the problems with imperfect coverage improved for the more recent data.

Coefficients from the fitted regressions should not be interpreted as having a causal relationship. There are unobserved characteristics that can determine both the explanatory variables and the dependent variable. One such example is constituency of a given stock in global indexes which leads to larger attention by both analysts and providers of fundamental data. The pooled panel regressions should merely answer the question whether there is a potential for problems with confounded variables. All the standard errors in the reported t-statistics are HAC robust.

β_0 is proportional to unconditional coverage. That is, the higher it is, the better the fundamental coverage for stocks of all sizes. It has increased from the 1990-2002 period to the 2003-2016 period for almost all the regions, and the increase has been substantial for the US and countries in the Asia Pacific, as would be expected from the previous graphs. β_1 then captures the discrete change in coverage at approximately \$100 million. It is insignificant or close to zero almost everywhere, with the exception of the US in the earlier period. This documents that the coverage of the stocks has indeed been only selective and not full in the US. On the other hand, the quantiles of size and more than 4 analysts following are significant almost everywhere. This means that both of them can lead to spurious results if the effect under study is somehow related to them. The size quantile tends to have a lower effect after the \$100 million threshold, as is evident from a mostly insignificant $\beta_3 + \beta_4$ measuring slope on on size for stocks with capitalization larger than \$100 million.

Table 1.8:

Predicting Fundamental Coverage

The table reports the estimated coefficients and corresponding t-statistics for the stock-month level logit regression of fundamental coverage on its explanatory variables

$$\begin{aligned} Fundamental\ Coverage_{it} &= \beta_0 + \beta_1 \mathbb{1}\{Size_{it} > \$100M\} + \beta_2 \mathbb{1}\{Analysts\ Following_{it} \geq 5\} \\ &+ \beta_3 Size\ Quantile_{it} + \beta_4 (Size\ Quantile_{it} - Size\ Quantile_t^{\$100M}) \mathbb{1}\{Size_{it} > \$100M\} + \epsilon_{it} \end{aligned}$$

where $Fundamental\ Coverage_{it}$ is equal to one when the given stock i has fundamental information in Datastream in month t and zero otherwise. $\mathbb{1}\{Size_{it} > \$100M\}$ is a dummy that is equal to one when market cap of the stock i in month t is larger than 100 million US Dollars and zero otherwise. $\mathbb{1}\{Analysts\ Following_{it} \geq 5\}$ is a dummy equal to one when number of analysts following the company i in month j in I/B/E/S is larger than or equal to 5 and zero otherwise. $Size\ Quantile_{it}$ describes cross-sectional quantile (0 to 1) of market capitalization of the stock i relative to all the stocks in country of its listing in month t. $(Size\ Quantile_{it} - Size\ Quantile_t^{\$100M}) \mathbb{1}\{Size_{it} > \$100M\}$ captures change in slope of the cross-sectional market cap quantile when the market cap is larger than 100 million US Dollars. We also report the Nagelkerke et al. (1991) R^2 index to measure goodness of fit. The standard errors in the reported t-statistics are HAC robust. The regression results are estimated on sample either from 1990 to 2002 or from 2003 to 2016.

	1990 - 2002						2003 - 2016					
	β_0	β_1	β_2	β_3	$\beta_4+\beta_3$	R^2	β_0	β_1	β_2	β_3	$\beta_4+\beta_3$	R^2
Australia	-1.87	0.21	2.12	2.92	4.01	0.40	1.35	-0.22	1.72	3.12	-1.69	0.08
	(-26.38)	(1.16)	(9.74)	(14.69)	(2.82)		(23.64)	(-1.10)	(4.20)	(13.77)	(-1.49)	
Austria	0.71	0.45	0.59	-0.15	1.12	0.06	-0.69	-0.09	2.74	5.06	0.71	0.31
	(2.00)	(1.12)	(1.77)	(-0.14)	(0.50)		(-1.54)	(-0.17)	(2.75)	(3.57)	(0.33)	
Belgium	-1.00	-0.61	2.44	4.69	0.02	0.31	-1.40	-0.30	0.41	7.90	3.09	0.43
	(-3.37)	(-1.39)	(3.96)	(4.95)	(0.01)		(-4.14)	(-0.55)	(0.64)	(7.01)	(1.58)	
Canada	-2.11	0.35	1.24	4.25	3.99	0.45	0.60	-0.10	1.15	6.53	1.18	0.17
	(-28.16)	(2.33)	(8.74)	(14.56)	(4.94)		(6.50)	(-0.68)	(4.63)	(9.68)	(2.21)	
Denmark	0.06	-0.55	1.76	4.04	-1.70	0.16	0.96	-1.04	2.13	7.31	-1.83	0.15
	(0.26)	(-1.12)	(4.39)	(5.07)	(-0.68)		(3.10)	(-1.61)	(3.11)	(6.02)	(-0.83)	
Finland	0.38	-0.95	1.20	3.13	-0.05	0.10	0.69	-0.86	1.94	6.54	-1.77	0.15
	(1.05)	(-2.27)	(3.33)	(2.61)	(-0.02)		(1.39)	(-1.34)	(2.78)	(3.64)	(-0.81)	
France	-0.36	-0.59	2.09	2.73	-0.26	0.16	-0.52	-0.23	1.26	5.46	1.45	0.27
	(-2.88)	(-3.65)	(9.74)	(7.48)	(-0.42)		(-4.79)	(-1.18)	(3.54)	(14.06)	(1.71)	
Germany	0.50	-0.43	1.58	1.92	0.82	0.10	-1.26	-1.18	1.71	7.80	3.07	0.45
	(2.79)	(-2.21)	(6.73)	(3.65)	(1.28)		(-14.00)	(-5.64)	(3.86)	(19.87)	(2.98)	
Greece	-0.46	-0.24	1.82	3.36	1.25	0.17	1.19	0.03	0.56	3.78	1.51	0.09
	(-2.11)	(-1.00)	(5.08)	(5.87)	(1.12)		(4.63)	(0.06)	(1.43)	(3.73)	(0.72)	
Hong Kong	-1.85	-0.25	0.27	3.31	6.27	0.19	1.00	-0.10	2.01	3.02	-0.36	0.07
	(-9.02)	(-1.89)	(1.86)	(9.09)	(6.71)		(5.59)	(-0.69)	(6.37)	(7.88)	(-0.57)	
Ireland	-1.45	-0.63	1.00	8.41	-0.17	0.44	1.40	-1.14	0.18	3.45	2.20	0.04
	(-4.52)	(-1.01)	(1.56)	(7.12)	(-0.08)		(2.00)	(-1.96)	(0.26)	(1.93)	(1.14)	
Italy	-0.08	-0.06	2.17	2.83	1.15	0.19	-0.98	-0.21	1.49	7.67	0.67	0.28
	(-0.22)	(-0.24)	(6.04)	(3.28)	(1.12)		(-3.24)	(-0.61)	(2.13)	(8.87)	(0.37)	
Japan	0.03	-0.73	2.20	5.41	3.14	0.20	3.39	0.15	1.51	0.75	-1.53	0.01
	(0.39)	(-10.46)	(6.48)	(20.97)	(12.98)		(33.64)	(0.96)	(6.66)	(2.06)	(-3.12)	
Netherlands	-0.24	-1.06	2.74	3.31	0.04	0.32	-0.90	-2.75	1.61	12.05	-1.91	0.34
	(-0.79)	(-2.38)	(9.24)	(3.45)	(0.04)		(-2.75)	(-3.39)	(3.11)	(7.32)	(-1.32)	
New Zealand	-2.91	0.14	1.92	4.86	-4.26	0.48	0.69	0.04	0.76	2.41	5.47	0.16
	(-9.30)	(0.30)	(5.87)	(7.10)	(-1.82)		(3.00)	(0.07)	(0.93)	(3.71)	(2.07)	
Norway	0.34	-0.15	2.18	2.40	-0.45	0.14	0.99	-0.49	1.97	3.46	-0.75	0.10
	(1.59)	(-0.52)	(4.82)	(3.78)	(-0.28)		(3.22)	(-1.56)	(4.84)	(3.73)	(-0.67)	
Portugal	-1.48	-0.12	2.44	5.02	-1.54	0.36	-0.42	0.74	-0.11	5.10	3.64	0.38
	(-8.52)	(-0.29)	(4.79)	(7.05)	(-0.88)		(-1.28)	(0.90)	(-0.07)	(3.88)	(0.64)	
Singapore	-0.69	-0.55	0.71	1.95	6.14	0.13	2.02	-0.42	1.79	1.58	3.38	0.04
	(-2.10)	(-3.00)	(3.31)	(3.51)	(3.92)		(6.51)	(-1.78)	(2.70)	(2.40)	(2.28)	
Spain	-0.06	-0.72	1.83	2.43	1.81	0.23	-1.02	1.36	2.27	6.06	-4.59	0.33
	(-0.14)	(-1.78)	(4.74)	(2.26)	(1.45)		(-2.00)	(2.22)	(3.47)	(4.24)	(-2.20)	
Sweden	-0.36	-0.36	3.01	3.43	0.04	0.23	0.38	-0.56	1.03	6.31	-1.21	0.19
	(-2.52)	(-1.39)	(6.44)	(7.08)	(0.04)		(2.92)	(-1.98)	(2.05)	(11.78)	(-1.18)	
Switzerland	-0.37	-0.75	1.66	3.51	0.76	0.19	1.71	0.37	1.23	2.81	-1.49	0.05
	(-1.34)	(-2.52)	(5.76)	(4.75)	(0.67)		(3.43)	(0.62)	(2.12)	(1.95)	(-0.60)	
UK	0.17	0.16	0.68	2.88	-0.31	0.11	0.68	-0.48	0.53	3.71	0.87	0.09
	(2.31)	(1.28)	(4.56)	(11.10)	(-0.61)		(9.88)	(-3.94)	(3.33)	(14.81)	(1.96)	
USA	-0.33	0.57	1.09	1.87	1.35	0.22	1.22	-0.40	0.56	4.57	0.74	0.05
	(-6.83)	(10.68)	(15.11)	(12.93)	(5.25)		(15.66)	(-5.23)	(7.09)	(12.00)	(4.00)	

1.5.3 Impact on Selection of Individually Significant Signals

We will now test the effect of imperfect fundamental coverage in international markets. Figure 1.4 showed that fundamental coverage in Datastream in Japan and Asia Pacific has changed to essentially 100% after 2000. These markets can therefore serve a testing ground for the impact of missing fundamental coverage. It is possible that the evidence documented so far is valid only in the US and it is therefore important to provide evidence in the other regions where often used as the primary source of fundamental information.

Table 1.9 tests the impact of imperfect fundamental coverage in Japan and Asia Pacific. Logit regressions estimating probability of no fundamental coverage for a given stock are fitted over the periods where the coverage was only partial, that is, July 1990 to July 1998 in Japan and July 1990 to July 2000 in Asia Pacific. The specification of regressions is the same as in equation (1.1). The fitted specification in Japan with 0.288 pseudo R^2 is

$$\begin{aligned} \mathit{Fundamental\ Coverage}_{it} &= -0.83(0.10) + 0.10(0.08)\mathbb{1}\{\mathit{Size}_{it} > \$100M\} \\ &+ 2.00(0.39)\mathbb{1}\{\mathit{Analysts\ Following}_{it} \geq 5\} + 5.02(0.32)\mathit{Size\ Quantile}_{it} \\ &- 1.56(0.35)(\mathit{Size\ Quantile}_{it} - \mathit{Size\ Quantile}_t^{\$100M})\mathbb{1}\{\mathit{Size}_{it} > \$100M\}. \quad (1.2) \end{aligned}$$

The fitted specification in Asia Pacific with 0.481 pseudo R^2 is

$$\begin{aligned} \mathit{Fundamental\ Coverage}_{it} &= -3.29(0.11) - 0.26(0.10)\mathbb{1}\{\mathit{Size}_{it} > \$100M\} \\ &+ 1.16(0.11)\mathbb{1}\{\mathit{Analysts\ Following}_{it} \geq 5\} + 4.99(0.22)\mathit{Size\ Quantile}_{it} \\ &+ 0.91(0.75)(\mathit{Size\ Quantile}_{it} - \mathit{Size\ Quantile}_t^{\$100M})\mathbb{1}\{\mathit{Size}_{it} > \$100M\}. \quad (1.3) \end{aligned}$$

We then predict the probability of no coverage for each stock in the period where the coverage was perfect, that is, July 2000 to December 2016 in Japan and July 2002 to December 2016 in Asia Pacific. We randomly discard stocks from the sample with perfect coverage according to the fitted probability of no coverage. The goal is to simulate what would happen if the imperfect fundamental coverage remained after 2000. Portfolios are then formed on the remaining sample of stocks that survived. The portfolios are created from different random samples. Average monthly returns over 10 random draws are then used in the analysis to add robustness. The portfolios on random subsample (Partial category in the table) are finally compared with the original full coverage sample (Full category).

There are statistically and economically significant differences for both individual and grouped anomalies in both Japan and Asia Pacific. The partial coverage portfolios yield one third lower returns than the full sample portfolios in Japan. There are 6 anomalies with significant differences in returns in Japan and 11 in Asia Pacific. There are 12 significant anomalies on the full sample in Japan, 7 on the partial sample, and 6 on both samples. There are 35 significant anomalies on the full sample in Asia Pacific, 32 on the partial sample, but only 26 on both samples.

The imperfect historical coverage in international markets has implications for returns on portfolios there in the very same way as in the US as argued in Section 1.4.2. Our study

Table 1.9:
Effect of the Imperfect Coverage Outside the US
The table shows returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. The probability of a stock having fundamental coverage in Datastream is estimated with logit model as in table 1.8 over July 1990 to July 1998 in Japan and July 1990 to July 2000 in Asia Pacific. The stocks are then randomly sampled from their full population with the fitted coverage probability over July 2000 to December 2016 in Japan and July 2002 to December 2016 in Asia Pacific. Portfolios on anomalies are created from the sampled stocks. Portfolio returns on anomalies created based on all the stocks (Full category) are compared with returns on portfolios created on the random subsample (Partial category). The partial category is based on mean monthly returns from 10 random draws of the stocks. The list of anomalies is provided in Appendix A. The equal-weighted portfolios are constructed by buying stocks in the top quintile of the signal and shorting stocks in the bottom quintile of the signal. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

	Japan				Asia Pacific			
	Corr	Full	Partial	Diff	Corr	Full	Partial	Diff
Accruals	0.829	0.08 (0.74)	0.04 (0.34)	(-1.13)	0.764	1.11 (3.66)	0.99 (3.90)	(-1.19)
Intangibles	0.844	0.15 (2.72)	0.05 (0.92)	(-2.54)	0.744	0.84 (3.13)	0.57 (2.99)	(-1.79)
Investment	0.843	0.10 (1.54)	0.06 (0.84)	(-1.50)	0.763	0.70 (4.76)	0.77 (5.14)	(0.90)
Profitability	0.841	0.13 (1.75)	0.18 (1.85)	(0.97)	0.762	-0.22 (-0.57)	0.04 (0.14)	(1.93)
Value	0.876	0.47 (3.31)	0.38 (2.30)	(-1.17)	0.799	0.77 (3.66)	0.97 (3.81)	(2.55)
All	0.845	0.18 (3.16)	0.13 (2.10)	(-2.07)	0.766	0.70 (4.20)	0.71 (4.42)	(0.21)

is therefore overwhelmingly showing that there could be a huge bias when looking at the performance of individual quantitative strategies in international markets in periods of imperfect fundamental coverage. The bias can completely distort the statistical inference and lead to findings of patterns that are only its artifacts. The simple remedies of focusing on universe of stocks that excludes microcaps proposed earlier can correct for a part of the bias, but they cannot control for all of it.

1.5.4 Impact on Selection of Independently Significant Signals

The analysis so far has focused on returns on portfolios. We will now show that the same caveats apply in regression setting as well. We follow the methodology from Green et al. (2017) to identify independently significant signals. Table 1.10 presents anomalies that are significant in Fama and MacBeth (1973a) panel regressions of individual stock returns on rescaled anomalies.[23] All the signals are pooled in the regressions, as follows:

$$r_{i,t} = \beta_0 + \sum_{j=1}^{M} \beta_j x_{i,j,t-1} + \epsilon_{i,t}. \tag{1.4}$$

for a given month t and number of signals M. $x_{i,j,t-1}$ is the signal for anomaly j and stock i that was available just before the start of month t. Raw fundamental signals are transformed into cross-sectional quantiles among all the stocks in a given region before

[23]The approach is covered in more detail in Section 2.4.

the regressions are run to limit the effect of outliers. We also remove binary variables and signals where the variance inflation factor is higher than 7.[24] We consider simple ordinary least squares (OLS) regressions (E) and the value-weighted weighted least squares (WLS) regression (V). The weight in the value-weighted WLS regression is proportional to market cap of individual stocks in each cross-section and should therefore limit the effect of small capitalization stocks. The regressions use all stock-month observations from July 1963 to December 2016 in the US and from July 1990 to December 2016 elsewhere. All the standard errors are HAC adjusted, as in Newey and West (1987), with 12 lags. We present the results for all the available stocks (All) and the restricted all-but-microcaps stocks with sizes larger than the bottom decile in the NYSE (Large). U stands for all signals found to be significant while **A** stands for those that remain significant after a correction for a false discovery rate (FDR) at 5%.

The FDR correction is very important since one would tend to find one significant signal in 20 individual tests even if all of them are insignificant in reality. The FDR adjustment follows Benjamini and Yekutieli (2001) and proceeds by first sorting p-values from the smallest to the largest so that $p_1 \leq p_2 \ldots \leq p_i \ldots \leq p_M$. FDR adjusted p-values are determined with backward induction where $p_M^{FDR} = p_M \sum_{1\leq j\leq M} \frac{1}{j}$ and

$$p_i^{FDR} = \min\left\{p_{i+1}^{FDR}, p_i \frac{M}{i} \sum_{1\leq j\leq M} \frac{1}{j}\right\} \tag{1.5}$$

The adjusted p-values p_i^{FDR} are then significant with an FDR of 5% if they are smaller than 5%.

The results for the US look staggering. There is only one common signal out of the 8 that is significant with FDR adjustment for Compustat for the full universe of stocks and OLS regressions. This does not change for all-but-microcaps stocks, with one in 5 signals being common. Value-weighting helps as it selects only one significant signal that is common across all the specifications for both the databases. The one commonly significant anomaly is the earnings predictability of Francis et al. (2004), which is surprisingly not related to any commonly used factor. Omitting FDR correction does not change the inference and there are still huge differences. This suggests that it is virtually impossible to select independently significant signals in the same country using different datasets.

The difference in the selected anomalies across the databases in the US then translates to large discrepancies for the international sample. It is apparent that some of the signals are common for the regions, but the variability is again great. Jacobs and Müller (2017a) conducted a similar exercise in international markets and found only a few signals that would be significant across all the regions. Our analysis here suggests that this result is a consequence of the imperfect coverage of Datastream in the individual regions. It serves as an important caveat that the population of stocks in individual regions and its coverage by data vendors has a substantial impact on research findings and anyone working with

[24] The exclusion of signals is done iteratively, and we primarily discard signals that would not be significant for any specification in the US.

international data should be aware of it.

Table 1.10:
Independently Significant Signals
The table shows signals that independently predict the returns on individual stocks in different regions. We measure predictability by significance of coefficients in the Fama and MacBeth (1973a) regressions. We regress the returns on past quantiles of fundamental signals across all stocks in the given region and month. We then focus on the t-statistics on the time-series mean of these coefficients. We report all signals with t-statistics larger than 2 (U) and those with p-values smaller than 5% after adjusting the original p-values for FDR (**A**). The regressions are either equal-weighted (E, standard OLS) or value-weighted (V, WLS with weights given by market cap). We compare the selected signals for CRSP & Compustat with those for Datastream for either the all-but-microcaps universe of stocks or for the full sample of stocks. The full sample (All) includes all available stocks, while the all-but-microcaps universe (Large) is restricted to stocks with capitalizations larger than that of bottom decile of the NYSE. The sample starts in July 1990 and ends in December 2016. The list of anomalies is provided in Appendix A.

	Compustat				Datastream															
	USA				USA				Europe				Japan				Asia Pacific			
	All		Large		All		Large		All		Large		All		Large		All		Large	
	E	V	E	V	E	V	E	V	E	V	E	V	E	V	E	V	E	V	E	V
EPr	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	-	-	U	-	U	-	U	-
CBOP	**A**	U	U	U	U	U	**A**	U	**A**	-	**A**	U	-	-	U	-	-	-	-	-
NOA	-	U	U	U	**A**	U	**A**	U	**A**	-	-	-	-	-	U	-	**A**	-	-	-
SP	U	U	**A**	U	**A**	-	U	-	**A**	-	-	-	-	-	U	-	**A**	-	-	-
RDM	**A**	-	**A**	-	U	-	U	-	**A**	-	U	-	**A**	-	U	-	-	-	-	-
ChNOA	**A**	-	**A**	-	-	-	U	-	**A**	U	U	U	-	-	-	-	-	-	-	-
PY	U	-	**A**	-	**A**	-	-	-	U	-	U	-	-	-	-	-	-	-	-	-
BM	**A**	-	-	-	U	-	-	-	**A**	**A**	**A**	**A**	-	-	-	-	**A**	U	U	U
WWI	U	-	-	-	**A**	-	-	-	U	-	-	-	-	U	-	U	**A**	-	-	-
CM	-	-	-	-	**A**	-	-	-	-	-	U	-	-	-	-	-	**A**	-	-	-
OL	**A**	U	U	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SaGr	**A**	-	U	-	-	-	U	-	-	-	-	-	U	-	-	-	-	-	-	-
GriI	U	-	-	-	**A**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GrLTNOA	**A**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U	-	-	-
AT	-	-	-	-	**A**	-	-	-	-	-	-	-	-	U	U	U	-	-	-	-
CEI5Y	-	-	-	-	-	-	U	-	**A**	U	**A**	-	-	-	U	-	U	U	U	U
ChGMChS	U	-	U	-	-	-	-	-	-	-	U	-	-	-	-	-	**A**	-	-	-
EP	-	-	-	-	-	U	-	U	**A**	-	U	-	-	-	-	-	U	-	-	-
NEF	-	-	-	-	U	-	-	-	U	-	-	-	-	-	-	-	**A**	-	-	-
SuGr	-	-	-	-	-	-	-	-	**A**	-	U	-	U	-	-	-	-	-	-	-
Acc	U	U	-	-	-	U	-	U	U	-	-	-	-	-	-	-	U	-	-	-
ChNNCOA	U	U	U	U	-	-	-	-	U	-	-	-	-	-	-	-	-	U	U	U
POA	U	-	U	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	U	U
NPY	U	-	-	-	U	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-
AGr	U	-	-	-	U	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-
ICh	U	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ES	-	-	-	-	-	U	-	U	-	-	-	-	-	-	-	-	-	-	-	-
OC	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TAN	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ChNCOL	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ChFL	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FSc	-	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HR	-	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lvrg	-	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ChCOL	-	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ChPPEIA	-	-	-	-	-	U	-	-	U	-	-	-	-	-	-	-	-	-	-	-
EM	-	-	U	-	-	-	-	-	-	-	-	-	U	-	-	-	-	-	-	-
ChiAT	-	-	-	-	-	-	-	-	U	U	U	U	-	-	-	-	-	-	-	-
NOACh	-	-	-	-	-	-	-	-	U	-	-	-	U	-	-	-	-	-	-	-
TXFIN	-	-	-	-	-	-	-	-	U	-	-	-	-	-	-	-	-	-	-	-
AL	-	-	-	-	-	-	-	-	U	U	-	-	-	-	-	-	-	-	-	-
EC	-	-	-	-	-	-	-	-	-	-	-	-	U	-	-	-	-	-	-	-
IR	-	-	-	-	-	-	-	-	-	-	-	-	U	-	-	-	-	-	-	-
OPtE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U	U	U
ChNNCWC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U	-	U
ICBE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U	U	U
CDI	-	-	-	-	-	-	-	-	-	-	-	U	-	-	-	-	-	-	-	-
HI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U
NDF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	U

1.6 Fundamental Coverage and Expected Returns

We have shown that fundamental coverage in Datastream is related to variables that are themselves related to expected returns on stocks. We will now study if the fundamental coverage is itself related to the expected returns. Negative relationship of the fundamental coverage to size of the stocks suggests that stocks with fundamental coverage have lower returns than those without (see Banz (1981)). The positive relationship to the number of analysts following, however, suggests that it is the other way around (see Elgers et al. (2001)). The predictive power of the fundamental coverage is therefore not immediately obvious.

Table 1.11 shows profitability of a strategy that buys stocks in CRSP that have the fundamental coverage in Datastream and shorts those that do not. We define stocks without the fundamental coverage as those for which we cannot construct book-to-market ratio as defined in Fama and French (1992). The sample spans July 1990 to December 2016. The strategy yields significantly positive returns for both equal-weighted and value-weighted returns. The significance is even higher once the returns are adjusted for the five Fama-French factors (Fama and French (2015)). The increase in significance is due to SMB factor capturing the effect of size that goes in the opposite direction than the effect of the fundamental coverage. The significantly positive mean returns remain even for the all-but-microcaps universe of stocks, defined as stocks with capitalization larger than that of bottom decile of the NYSE. The table also shows minimum and average number of stocks in CRSP without the fundamental coverage in each month. The average number of stocks is over 500 even for the all-but-microcaps universe and the results are therefore based on a large sample of stocks.

The relative underperformance of stocks without fundamental coverage is in line with underperformance of stocks with small number of analysts following. The similarity is hardly surprising since we have previously shown that the number of analysts is one of the criterion for the decision whether to provide the fundamental coverage in Datastream. The similarity can also be strengthened by the fact that Thompson Reuters owns both the database for analysts' forecasts (I/B/E/S) and the fundamental database (Worldscope in Datastream), although it has not been the case historically. The decision whether to cover a given firm can therefore be interconnected in both databases. The theoretical reasoning for no coverage can be very similar across the databases as well. The analysts are less likely to cover stocks that are underperforming and have small growth potential since they have only limited resources at their disposal. They therefore try to channel these resources at firms that attract the most investor's attention. The coverage in Datastream is also likely to prioritize stocks with large investors attention to successfully compete with other data vendors.

The underperformance can also be connected to a backfilling bias in that firms that outperform in the long-term are eventually added to the fundamental database along with the full history of their financial statements. The entries in the database then appear to be there historically even it was not the case at the time. The underperforming firms are more likely to never be added, especially if they are soon delisted. The profitability of the strategy is therefore probably only illusory and cannot be captured in real life.

Table 1.11:
Firms without Fundamental Coverage in Datastream
The table shows returns and alphas with their corresponding t-statistics on long-short portfolios created from stocks in CRSP by buying those that have fundamental coverage in Datastream and shorting those that do not. The portfolios are either equal-weighted (EW) or value-weighted (VW). The full sample includes all available non-financial stocks while the all-but-microcaps universe is restricted to stocks with capitalization larger than that of bottom decile of the NYSE. The sample spans from July 1990 to December 2016. The alpha is estimated with respect to the Fama-French five factor model, and the factor loadings are also provided. The reported returns are in percent per month. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Newey and West (1987), with 12 lags.

	Full Sample		All-but-microcaps	
	EW	VW	EW	VW
Mean Return	1.45	0.44	0.89	0.40
	(8.57)	(3.36)	(6.24)	(3.11)
Alpha FF5	1.20	0.42	0.80	0.39
	(8.15)	(5.18)	(9.60)	(4.67)
Mkt	0.17	-0.03	-0.02	-0.04
	(6.98)	(-0.91)	(-0.91)	(-1.16)
SMB	-0.08	-0.30	-0.22	-0.27
	(-1.65)	(-7.86)	(-5.45)	(-7.11)
HML	0.01	-0.04	0.18	-0.05
	(0.17)	(-0.75)	(3.43)	(-0.83)
RMW	0.41	0.30	0.24	0.29
	(8.28)	(5.45)	(5.57)	(5.08)
CMA	0.06	0.00	0.08	0.01
	(0.71)	(0.06)	(1.24)	(0.22)
Avg # of stocks	1581	1581	551	551
Min # of stocks	342	342	178	178

1.6.1 Low Profitability Firms without Fundamental Coverage in Datastream

We have previously documented large differences in returns on profitability anomalies in Compustat relative to Datastream. The discrepancy is mainly due to stocks in low profitability category and we study them in more detail here. The stocks that are among the least profitable in Compustat and have no fundamental coverage in Datastream have severely underperformed since 2000. This underperformance could be connected to the low interest of the investor since they were not worth following by one of the main data vendors. It could also be because they are difficult to short, which introduces limits to arbitrating and allows only a slow adjustment. We will now study the low profitability stocks without the fundamental coverage in Datastream in more detail.

Table 1.12 presents the average monthly returns on a strategy that buys all stocks without fundamental coverage in Datastream that are in the bottom decile or quintile of operational profitability in Compustat. We measure profitability by operating profits to assets as in Ball et al. (2016). Our sample either includes all non-financial stocks or we further discard all stocks with sizes smaller than bottom decile on the NYSE in the previous June. The portfolios start in 2000 as there are many stocks missing fundamental coverage in Datastream in 1990s and this dilutes the overall effect.[25] The portfolios are either value-weighted (VW) or equal-weighted (EW). A value-weighted strategy in which shorts stocks without fundamental coverage in Datastream that are in the lowest profitability decile in Compustat yields 27% annually over the 2000 to 2016 period. The strategy is also significant for equal-weighted returns. The returns remain significant on a all-but-microcaps universe. Alphas with respect to the Fama-French five factor model are even more significant with t-statistics of approximately 6. There are, on average, 133 stocks in the portfolio for the full sample but fewer for the all-but-microcaps sample. The evidence is therefore based only on few data points. We have tried to look at individual instances of these stocks. The stocks are often facing bankruptcy and have management problems.

There are several possible explanations for this anomaly. First, it could be the case that the fundamental data have been backfilled in Compustat only after some time. The stocks have been in CRSP for 72 months on average, so the late addition of fundamental information on new issues cannot fully explain the difference. It is also possible that the difference is due to the inattention of investors. We can proxy for the attention by the number of analysts following them. Elgers et al. (2001) show that the number of financial analysts covering the stocks can predict the future return. The stocks in the portfolios have, on average, 3.19 analysts covering them, which is lower than the 7.35 analysts for all the other stocks. This is in line with our previous analysis that the stocks would have fundamental coverage if they had more than 4 analysts coverings them. The same caveat applies as for the fundamental coverage anomaly described in the previous section in that the profits are probably only illusory and cannot be captured in real life.

1.7 Robustness

Here, we provide robustness to our findings. Our previous analysis focused on quantile portfolios with return weighting following the original studies. We will now show that our conclusions remain unchanged for a different construction of the portfolios. Table 1.13 presents the differences in the portfolios sorted on anomalies for different constructions of the portfolios. We extend our previous analysis to decile and tercile breakpoints in

[25]The overall inference remains nonetheless unchanged even for 1990 to 2016 period.

Table 1.12:
Low Profitability Firms without Fundamental Coverage in Datastream
The table shows returns and alphas with their corresponding t-statistics on portfolios created from stocks that are within the bottom decile (quintile) of profitability stocks in Compustat but do not have fundamental coverage in Datastream. The portfolios are either equal-weighted (EW) or value-weighted (VW). We measure profitability by operating profits to assets as in Ball et al. (2016). The full sample includes all available non-financial stocks while the all-but-microcaps universe is restricted to stocks with capitalization larger than that of bottom decile of the NYSE. The sample spans from July 2000 to December 2016. The alpha is estimated with respect to the Fama-French five factor model, and the factor loadings are also provided. The reported returns are in percent per month. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Newey and West (1987), with 12 lags.

	Full Sample				All-but-microcaps			
	Decile		Quintile		Decile		Quintile	
	EW	VW	EW	VW	EW	VW	EW	VW
Mean Return	-1.95	-2.21	-1.82	-1.43	-2.15	-2.27	-1.79	-1.33
	(-2.59)	(-2.92)	(-2.67)	(-2.03)	(-2.88)	(-2.65)	(-2.82)	(-1.84)
Alpha FF5	-2.22	-2.68	-2.07	-1.72	-2.61	-2.74	-2.08	-1.58
	(-5.53)	(-7.03)	(-6.05)	(-5.02)	(-6.01)	(-5.35)	(-7.29)	(-4.02)
Mkt	0.73	0.99	0.68	0.99	0.90	1.07	0.85	1.03
	(5.83)	(6.82)	(7.34)	(8.61)	(8.86)	(6.94)	(9.27)	(8.28)
SMB	0.99	1.00	0.91	0.65	1.27	0.96	0.94	0.56
	(6.40)	(4.98)	(7.25)	(5.81)	(8.24)	(4.14)	(10.20)	(4.10)
HML	-0.22	-0.39	-0.21	-0.36	-0.52	-0.30	-0.26	-0.34
	(-0.91)	(-1.67)	(-0.93)	(-1.78)	(-2.60)	(-1.00)	(-1.29)	(-1.54)
RMW	-1.03	-0.72	-0.90	-0.60	-0.72	-0.67	-0.64	-0.57
	(-4.97)	(-3.86)	(-5.96)	(-4.19)	(-4.44)	(-3.36)	(-5.68)	(-3.59)
CMA	0.57	0.58	0.48	0.30	0.55	0.40	0.11	0.18
	(2.25)	(2.25)	(2.16)	(1.34)	(2.37)	(1.27)	(0.62)	(0.80)
Avg # of stocks	133	133	236	236	16.40	16.40	42	42
Min stocks	25	25	50	50	2	2	8	8

portfolio sorts and value-weighting. It is apparent that there is only a slight difference for the various breakpoints on equal-weighted portfolios. Value-weighted portfolios have lower average returns and t-statistics, but some differences among the databases still remain.

Panel D captures the number of significant signals with t-statistics larger than 2 for the various portfolio constructions. The number of significant anomalies is very similar across the two databases, but it is generally smaller in Datastream. The number of signals that are significant across both the databases is always lower than for Compustat alone by at least one fourth. The previous conclusions therefore carry over to other portfolio constructions and are very robust.

Table 1.13:
Robustness - Different Portfolio Construction

The table shows the returns and their corresponding t-statistics among portfolios created from sorts on fundamental anomalies. We compare the portfolios created with CRSP & Compustat or with just Datastream. The portfolios are either value-weighted or equal-weighted with decile, quintile, or tercile breakpoints in sorts. We also show correlation between the two cases. The sample starts in July 1990 and ends in December 2016. The list of anomalies is provided in Appendix A. The portfolios are constructed by buying stocks in the top decile, quintile, or tercile of the signal and shorting stocks in the bottom decile, quintile, or tercile of the signal. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Driscoll and Kraay (1998).

	Equal-weighted Portfolios				Value-weighted Portfolios			
	Corr	CT	DS	Diff	Corr	CT	DS	Diff
Panel A: Decile Portfolios								
Accruals	0.736	0.61 (4.85)	0.61 (4.85)	(1.56)	0.785	0.37 (3.08)	0.37 (3.08)	(-0.24)
Intangibles	0.611	0.43 (3.74)	0.43 (3.74)	(5.41)	0.624	0.12 (2.92)	0.12 (2.92)	(3.57)
Investment	0.776	0.60 (6.48)	0.60 (6.48)	(0.07)	0.757	0.38 (3.00)	0.38 (3.00)	(-1.49)
Profitability	0.761	0.53 (-0.03)	0.53 (-0.03)	(-5.68)	0.827	0.41 (1.42)	0.41 (1.42)	(-2.70)
Value	0.857	0.74 (3.36)	0.74 (3.36)	(0.49)	0.870	0.28 (1.60)	0.28 (1.60)	(0.67)
All	0.747	0.59 (6.81)	0.59 (6.81)	(0.47)	0.772	0.31 (3.75)	0.31 (3.75)	(0.13)
Panel B: Quintile Portfolios								
Accruals	0.762	0.56 (5.05)	0.56 (5.05)	(1.03)	0.834	0.27 (2.45)	0.27 (2.45)	(-0.42)
Intangibles	0.714	0.41 (3.48)	0.41 (3.48)	(3.81)	0.733	0.15 (2.81)	0.15 (2.81)	(3.19)
Investment	0.815	0.49 (6.36)	0.49 (6.36)	(-0.92)	0.835	0.23 (2.78)	0.23 (2.78)	(-0.26)
Profitability	0.827	0.38 (-0.14)	0.38 (-0.14)	(-6.61)	0.868	0.30 (0.92)	0.30 (0.92)	(-3.51)
Value	0.899	0.64 (3.39)	0.64 (3.39)	(-0.24)	0.908	0.25 (1.61)	0.25 (1.61)	(0.64)
All	0.800	0.50 (6.38)	0.50 (6.38)	(-0.82)	0.834	0.24 (3.36)	0.24 (3.36)	(0.07)
Panel C: Tercile Portfolios								
Accruals	0.754	0.47 (5.03)	0.47 (5.03)	(0.74)	0.854	0.17 (1.98)	0.17 (1.98)	(0.65)
Intangibles	0.772	0.35 (3.41)	0.35 (3.41)	(2.51)	0.766	0.12 (2.69)	0.12 (2.69)	(2.63)
Investment	0.832	0.39 (6.32)	0.39 (6.32)	(-0.89)	0.867	0.21 (2.86)	0.21 (2.86)	(-0.68)
Profitability	0.835	0.32 (-0.05)	0.32 (-0.05)	(-7.01)	0.886	0.22 (1.43)	0.22 (1.43)	(-1.69)
Value	0.914	0.54 (3.16)	0.54 (3.16)	(-0.66)	0.935	0.18 (1.26)	0.18 (1.26)	(0.13)
All	0.817	0.42 (5.99)	0.42 (5.99)	(-1.20)	0.860	0.18 (3.03)	0.18 (3.03)	(0.79)

Panel D: Number of significant signals

	Equal-weighted			Value-weighted		
	CT	DS	both	CT	DS	both
Decile Portfolios	44	39	30	14	13	7
Quintile Portfolios	41	39	29	11	12	7
Tercile Portfolios	38	38	26	9	5	3

1.8 Conclusion

We have compared fundamental data from two sources, and we have shown that measurement error in the fundamental data can be large. There are substantial differences in the raw financial statements caused by different methodologies for the construction of statements in the databases. These are less pronounced for portfolios created from sorts on fundamental signals. The findings on the significance of anomalies constructed with Compustat are therefore robust to measurement error. We have documented several problems with Datastream. We have managed to correct some, but others have no clear solution. The strong message of this paper is that Datastream is a good source of data only after approximately 2000, and its use in an earlier period could be connected to a significant bias. This is true for both the US and the international samples. We have also revisited the role of delisting returns and have not found any serious bias introduced by setting missing delisting returns to zero, unlike in the previous studies.

Appendix A

List of Fundamental Anomalies

Table A.1:
List of Published Fundamental Anomalies

Accruals		
Acc	Accruals	Sloan (1996)
ChCE	Change in Common Equity	Richardson et al. (2006)
ChCOA	Change in Current Operating Assets	Richardson et al. (2006)
ChCOL	Change in Current Operating Liabilities	Richardson et al. (2006)
ChFL	Change in Financial Liabilities	Richardson et al. (2006)
ChLTI	Change in Long-Term Investments	Richardson et al. (2006)
ChNFA	Change in Net Financial Assets	Richardson et al. (2006)
ChNNCWC	Change in Net Non-Cash Working Capital	Richardson et al. (2006)
ChNNCOA	Change in Net Non-Current Operating Assets	Richardson et al. (2006)
ChNCOA	Change in Non-Current Operating Assets	Richardson et al. (2006)
ChNCOL	Change in Non-Current Operating Liabilities	Richardson et al. (2006)
GriI	Growth in Inventory	Thomas and Zhang (2002)
ICh	Inventory Change	Thomas and Zhang (2002)
IGr	Inventory Growth	Belo and Lin (2011)
MBaAC	M/B and Accruals	Bartov and Kim (2004)
NWCCh	Net Working Capital Changes	Soliman (2008)
POA	Percent Operating Accrual	Hafzalla et al. (2011)
PTA	Percent Total Accrual	Hafzalla et al. (2011)
TA	Total Accruals	Richardson et al. (2006)
Intangibles		
ChGMChS	$\triangle$ Gross Marging - $\triangle$ Sales	Abarbanell and Bushee (1998)
SmI	$\triangle$ Sales - $\triangle$ Inventory	Abarbanell and Bushee (1998)
AL	Asset Liquidity	Ortiz-Molina and Phillips (2014)
EPr	Earnings Predictability	Francis et al. (2004)
ES	Earnings Smoothness	Francis et al. (2004)
HI	Herfindahl Index	Hou and Robinson (2006)
HR	Hiring rate	Belo et al. (2014)
ICBE	Industry Concentration Book Equity	Hou and Robinson (2006)
IARER	Industry-adjusted Real Estate Ratio	Tuzel (2010)
OC	Org. Capital	Eisfeldt and Papanikolaou (2013)
RDM	RD / Market Equity	Chan et al. (2001)
TAN	Tangibility	Hahn and Lee (2009)
URDI	Unexpected RD Increases	Eberhart et al. (2004)
WWI	Whited-Wu Index	Whited and Wu (2006)
Investment		
CAPEX	$\triangle$ CAPEX - $\triangle$ Industry CAPEX	Abarbanell and Bushee (1998)
AGr	Asset Growth	Cooper et al. (2008)
ChNOA	Change Net Operating Assets	Hirshleifer et al. (2004)
ChPPEIA	Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
CDI	Composite Debt Issuance	Lyandres et al. (2007)
CEI5Y	Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
DI	Debt Issuance	Spiess and Affleck-Graves (1995)
GrLTNOA	Growth in LTNOA	Fairfield et al. (2003)
INV	Investment	Titman et al. (2004)
NDF	Net Debt Finance	Bradshaw et al. (2006)
NEF	Net Equity Finance	Bradshaw et al. (2006)
NOA	Net Operating Assets	Hirshleifer et al. (2004)
NOACh	Noncurrent Operating Assets Changes	Soliman (2008)
SR	Share Repurchases	Ikenberry et al. (1995)
TXFIN	Total XFIN	Bradshaw et al. (2006)
Profitability		
AT	Asset Turnover	Soliman (2008)
CT	Capital Turnover	Haugen and Baker (1996)
CBOP	Cash-based Operating Profitability	Ball et al. (2016)
ChiAT	Change in Asset Turnover	Soliman (2008)
EP	Earnings / Price	Basu (1977)
EC	Earnings Consistency	Alwathainani (2009)
FSc	F-Score	Piotroski (2000)
GP	Gross Profitability	Novy-Marx (2013)
Lvrg	Leverage	Bhandari (1988)
OSc	O-Score (More Financial Distress)	Dichev (1998)
OPtA	Operating Profits to Assets	Ball et al. (2016)
OPtE	Operating Profits to Equity	Fama and French (2015)
Value		
AM	Assets-to-Market	Fama and French (1992)
BM	Book Equity / Market Equity	Fama and French (1992)
CM	Cash Flow / Market Equity	Lakonishok et al. (1994)
DurE	Duration of Equity	Dechow et al. (2004)
ECoBP	Enterprise Component of Book/Price	Penman et al. (2007)
EM	Enterprise Multiple	Loughran and Wellman (2011)
IR	Intangible Return	Daniel and Titman (2006)
LCoBP	Leverage Component of Book/Price	Penman et al. (2007)
NPY	Net Payout Yield	Boudoukh et al. (2007)
OL	Operating Leverage	Novy-Marx (2010)
PY	Payout Yield	Boudoukh et al. (2007)
SaGr	Sales Growth	Lakonishok et al. (1994)
SP	Sales/Price	Barbee Jr et al. (1996)
SuGr	Sustainable Growth	Lockwood and Prombutr (2010)

Appendix B

Construction of the Anomalies

Anomalies are grouped into 5 categories: accruals, profitability, value, investment, and intangibles. Construction of individual anomalies follows Harvey et al. (2016), McLean and Pontiff (2016) and Hou et al. (2017), with the exception of selecting a subset of exchanges and frequency of rebalancing. When these exceptions apply, they are described in the individual anomalies' definitions.

Accruals

Accruals (Acc)

Based on Sloan (1996), accruals are defined as

$$Acc = \frac{(\Delta act_t - \Delta che_t) - (\Delta lct_t - \Delta dlc_t - \Delta tp_t) - dp_t}{(at_t + at_{t-1})/2}$$

where Δact_t is change in current assets, Δche_t is change in cash and cash equivalents, Δlct_t is annual change in current liabilities, Δdlc_t is annual change in debt included in current liabilities, Δtp_t is annual change in income taxes payable and dp is depreciation and amortization expense.

Change in Current Operating Assets (ChCOA)

Based on Richardson et al. (2006), change in current operating assets is defined as

$$ChCOA = \frac{COA_t - COA_{t-1}}{at_{t-1}}$$

where COA_t are current operating assets, $COA_t = act_t - che_t$ in which act_t are current assets, che_t are cash and short-term investment and at_{t-1} are one-year lagged total assets

Change in Current Operating Liabilities (ChCOL)

Based on Richardson et al. (2006), change in current operating liabilities is defined as

$$ChCOL = \frac{COL_t - COL_{t-1}}{at_{t-1}}$$

where COL_t are current operating liabilities, $COL_t = lct_t - dlc_t$ in which lct_t are current liabilities, dlc_t is debt in current liabilities and at_{t-1} are one-year lagged total assets.

Change in Net Non-Cash Working Capital (ChNNCWC)

Based on Richardson et al. (2006), Change in Net Non-Cash Working Capital is defined as

$$ChNNCWC = \frac{WC_t - WC_{t-1}}{at_{t-1}}$$

where WC_t is working capital, $WC_t = COA_t - COL_t$ in which COA_t are current operating assets defined above in Change in Current Operating Assets anomaly and COL_t are current operating liabilities defined above in Change in Current Operating Liabilities anomaly.

Change in Net Non-Current Operating Assets (ChNNCOA)

Based on Richardson et al. (2006), Change in Net Non-Current Operating Assets is defined as

$$ChNNCOA = \frac{NCOA_t - NCOA_{t-1}}{at_{t-1}}$$

where NCO_t are non-current operating asset, $NCOA_t = NCA_t - NCL_t$ in which NCA_t are non-current assets defined in Change in Non-Current Operating Assets anomaly and NCL_t are non-current operating liabilities defined in Change in Non-Current Operating Liabilities anomaly.

Change in Non-Current Operating Assets (ChNCOA)

Based on Richardson et al. (2006), Change in Non-Current Operating Assets is defined as

$$ChNCOA = \frac{NCA_t - NCA_{t-1}}{at_{t-1}}$$

where NCA_t are non-current assets defined as $NCA_t = at_t - act_t - ivao_t$ where at_t are total assets, act_t are current assets, $ivao_t$ is investment and advances (0 if missing).

Change in Non-Current Operating Liabilities (ChNCOL)

Based on Richardson et al. (2006), Change in Non-Current Operating Liabilities is defined as

$$ChNCOL = \frac{NCL_t - NCL_{t-1}}{at_{t-1}}$$

where $NCL_t = lt_t - lct_t - dltt_t$ in which lt_t are total liabilities, lct_t are current liabilities and $dltt_t$ is long-term debt (0 if missing).

Change in Net Financial Assets (ChNFA)

Based on Richardson et al. (2006), Change in Net Financial Assets is defined as

$$ChNFA = \frac{NFNA_t - NFNA_{t-1}}{at_{t-1}}$$

where

$$NFNA_t = FNA_t - FNL_t$$

are net financial assets. FNA_t are financial assets, $FNA_t = ivst_t + ivao_t$. Where $ivst_t$ are short-term investments, $ivao_t$ are long-term investments. FNL_t are financial liabilities, $FNL_t = dltt_t + dlc_t + pstk_t$. Where $dltt_t$ is long-term debt, dlc_t is debt in current liabilities, and $pstk_t$ is preferred stock.

Change in Long-Term Investments (ChLTI)

Based on Richardson et al. (2006), Change in Long-Term Investments is defined as

$$ChLTI = \frac{ivao_t - ivao_{t-1}}{at_{t-1}}$$

where $ivao_t$ are long-term investments and at_{t-1} are one-year lagged total assets.

Change in Common Equity (ChCE)

Based on Richardson et al. (2006), Change in Common Equity is defined as

$$ChCE = \frac{ceq_t - ceq_{t-1}}{at_{t-1}}$$

where ceq_t is common equity and at_{t-1} are one-year lagged total assets.

Change in Financial Liabilities (ChFL)

Based on Richardson et al. (2006), Change in Financial Liabilities is defined as

$$ChFL = \frac{FNL_t - FNL_{t-1}}{at_{t-1}}$$

where FNL_t are net financial liabilities defined in anomaly Change in Net Financial Assets and at_{t-1} are one-year lagged total assets.

Growth in Inventory (GriI)

Based on Thomas and Zhang (2002), Growth in Inventor is defined as

$$GriI = \frac{invt_t - invt_{t-1}}{(at_t + at_{t-1})/2}$$

where $invt_t$ are inventories and at_t are total assets.

Inventory Change (ICh)

Based on Thomas and Zhang (2002), inventory change is defined as

$$ICh = \frac{invt_t - invt_{t-1}}{at_{t-1}}$$

where $invt_t$ are inventories and at_{t-1} are one-year lagged total assets.
Only firms with positive inventories in this or previous year are included.

Inventory Growth (IGr)

Based on Belo and Lin (2011), inventory growth is defined as

$$IGr = \frac{invt_t - invt_{t-1}}{invt_{t-1}}$$

where $invt_t$ are inventories.

M/B and Accruals (MBaAC)

Based on Bartov and Kim (2004), M/B and Accruals is defined as

$$MBaAC = \begin{cases} 1 & \text{if stock is in low book-to-market } (BM_t) \text{ and high accrual } (Accr_t) \text{ quintiles} \\ -1 & \text{if stock is in high book-to-market } (BM_t) \text{ and low accrual } (Accr_t) \text{ quintiles} \\ 0 & \text{otherwise} \end{cases}$$

Accruals (Acc_t) are defined above, and book-to-market (BM_t) - book equity divided by market equity - is defined in category *Value*.

Net Working Capital Changes (NWCCh)

Based on Soliman (2008), net working capital changes are defined as

$$NWCCh = \frac{NWC_t - NWC_{t-1}}{at_{t-1}}$$

$NWC_t = (act_t - che_t) - (lct_t - dlc_t)$ is net working capital, where act_t are current assets, che_t is cash and cash equivalents, clt_t are current liabilities and dlc_t is debt in current liabilities.

Percent Operating Accruals (POA)

Based on Hafzalla et al. (2011), percent operating accruals are defined as

$$POA = \frac{ni_t - oancf_t}{|ni_t|}$$

where ni_t is net income and $oancf_t$ is cash flow from operations.

Percent Total Accruals (PTA)

Based on Hafzalla et al. (2011), percent total accruals are defined as

$$PTA = \frac{ni_t - (-sstk_t + prstkc_t + dv_t + oancf_t + ivncf_t + fincf_t)}{|ni_t|}$$

where ni_t is net income, $sstk_t$ sale of common and preferred stock, $prstkc_t$ is purchase of common and preferred stock, dv_t is total dividends, $oancf_t$ is cash flow from financing, $ivncf_t$ is cash flow from investment and $fincf_t$ is cash from from financing.

Total Accruals (TA)

Based on Richardson et al. (2006), total accruals are defined as

$$TA = \frac{TACCR_t - TACCR_{t-1}}{at_{t-1}}$$

where $TACCR_t = NCO_t + WC_t + NFNA_t$ NCO_t are net non-current operating assets defined in anomaly Change in Net Non-Current Operating Assets, WC_t is working capital defined in anomaly Change in Net Non-Cash Working Capital and $NFNA_t$ are net financial assets defined in anomaly Change in Net Financial Assets.

Intangibles

Asset Liquidity (AL)

Based on Ortiz-Molina and Phillips (2014), asset liquidity is defined as

$$AL = \frac{che_t + 0.75(act_t - che_t) + 0.5(at_t - act_t - gdwl_t - intan_t)}{at_{t-1}}$$

where at_{t-1} are one-year lagged total assets, act_t are current assets, che_t is cash and short-term investments, $gdwl_t$ is goodwill (0 if missing) and $intan_t$ are intangibles (0 if missing).

Asset Liquidity II (AL2)

Based on Ortiz-Molina and Phillips (2014), Asset Liquidity II is defined as

$$AL2 = \frac{che_t + 0.75(act_t - che_t) + 0.5(at_t - act_t - gdwl_t - intan_t)}{ME_{t-1}}$$

where the definition of variables is the same as for AL and market equity ME_{t-1} is price times shares outstanding, $ME_t = prc_{t-1} * shrout_{t-1}$.

∆Sales - ∆Accounts Receivable (ChSChAR)

Based on Abarbanell and Bushee (1998), ∆Sales - ∆Accounts Receivable is defined as

$$ChSChAR = \frac{sale_t - \frac{sale_{t-1}+sale_{t-2}}{2}}{\frac{sale_{t-1}+sale_{t-2}}{2}} - \frac{rect_t - \frac{rect_{t-1}+rect_{t-2}}{2}}{\frac{rect_{t-1}+rect_{t-2}}{2}}$$

where $sale_t$ is net sales and $rect_t$ are total receivables.
Only firms with positive two-year sales and two-year gross margin averages are included.

∆Gross Margin - ∆Sales (ChGMChS)

Based on Abarbanell and Bushee (1998), ∆Gross Margin - ∆Sales is defined as

$$ChSChAR = \frac{GM_t - \frac{GMs_{t-1}+GM_{t-2}}{2}}{\frac{GM_{t-1}+GM_{t-2}}{2}} - \frac{sale_t - \frac{sale_{t-1}+sale_{t-2}}{2}}{\frac{sale_{t-1}+sale_{t-2}}{2}}$$

where $sale_t$ is net sales and GM_t is gross margin, defined as $GM_t = sale_t - cogs_t$, where $cogs_t$ is cost of goods sold.
Only firms with positive two-year sales and two-year gross margin averages are included.

Earnings Conservatism (EC)

Based on Francis et al. (2004),

$$EARN_{it} = \alpha_{i0} + \alpha_{i1} NEG_{it} + \beta_{i1} R_{it} + \beta_{i2} NEG_{it} R_{it} + e_{it}$$

in which $EARN_{it} = \frac{ib_t}{ME_t}$, where ib_t are earnings, ME_t is market equity defined in anomaly book-to-market in Section Value, R_{it} is i's stock 15-month return and NEG_{it} is defined as:

$$NEG_{it} = \begin{cases} 1 & \text{if } R_{it} < 0 \\ 0 & \text{otherwise} \end{cases}$$

Earnings Conservatism is defined as $EC = \frac{\beta_{i1}+\beta_{i2}}{\beta_{i1}}$.

Earnings Persistence (EPe)

Based on Francis et al. (2004), Earnings Persistence is defined as the slope coefficient (beta) from the first-order autoregressive model using the ten-year rolling window for split-adjusted earnings per share. Split-adjusted earnings per share are defined as $EPS_t = \frac{epspx_t}{ajex_t}$.
Only firms with no missing required data over the ten-year rolling window are included.

Earnings Predictability (EPr)

Based on Francis et al. (2004), Earnings Predictability is defined as volatility of residuals from the first-order autoregressive model using the ten-year rolling window for split-adjusted earnings per share. Split-adjusted earnings per share are defined as $EPS_t = \frac{epspx_t}{ajex_t}$.
Only firms with no missing required data over the ten-year rolling window are included.

Earnings Timeliness (ET)

Based on Francis et al. (2004),

$$EARN_it = \alpha_{i0} + \alpha_{i1}NEG_{it} + \beta_{i1}R_{it} + \beta_{i2}NEG_{it}R_{it} + e_{it}$$

in which $EARN_it = \frac{ib_t}{ME_t}$, where ib_t are earnings, ME_t is market equity defined in anomaly book-to-market in Section *Value*, R_{it} is i's stock 15-month return, and NEG_{it} is defined as:

$$NEG_{it} = \begin{cases} 1 & \text{if } R_{it} < 0 \\ 0 & \text{otherwise} \end{cases}$$

Earnings Timeliness is defined as R^2 from the regression.

Earning Smoothness (ES)

Based on Ortiz-Molina and Phillips (2014), earnings smoothness is defined as

$$ES = \frac{std(ELA_t)}{std(CFOA_t)}$$

where the standard deviation is calculated over the ten-year rolling window and only firms with no missing required data over the ten-year history are included. Further

$$ELA_t = \frac{ib_t}{at_{t-1}}$$

and

$$CFOA_t = ib_t - (DCA_t - DCL_t - DCHE_t + DSTD_t - dp_t))$$

where ib_t are earnings and at_{t-1} is lagged total assets. DCA_t is one-year change in current assets, DCL_t is the one-year change in current liabilities, $DCHE_t$ is the one-year change in cash and short-term investments, $DSTD_t$ is the one-year change in debt in current liabilities, and dp_t is depreciation and amortization.

Herfindahl Index (HI)

Based on Hou and Robinson (2006), Herfindahl index as a measure of industry concentration defined as

$$HI = \frac{H_t + H_{t-1} + H_{t-2}}{3}$$

$H_t = \sum_{i=1}^{N_j} sale_{i,j}$, where $sale_{ij}$ is the sale of firm i in industry j and N_j is the total number of firms in the 3-digit SIC code defined industry.

Hiring rate (HR)

Based on Belo et al. (2014), hiring rate is defined as

$$HR = \frac{emp_{t-1} - empt - 2}{0.5emp_{t-1} + 0.5emp_{t-2}}$$

where emp_t is the number of employees. Stocks with $HR = 0$, often a consequence of a stale information, are exluded.

Industry-adjusted Real Estate Ratio (IARER)

Based on Tuzel (2010),industry-adjusted real estate ratio is defined as

$$IARER = RER_t - \frac{\sum_{j=1}^{N_j} RER_{ij}}{N_j}$$

i.e. the real estate ratio minus its, 2-digit SIC code defined, industry average. Real estate ratio is defined as

$$RER_t = (fatb_t + fatl_t)/ppent_t$$

where $fatb_t$ is the sum of buildings at cost, $fatl_t$ is leases at cost and $ppegt_t$ is gross property, plant, and equipment.
Industries with less than five firms are excluded.

Industry-adjusted Organizational Capital-to-Assets (IaOCA)

Based on Eisfeldt and Papanikolaou (2013), Industry-adjusted Organizational Capital-to-Assets is defined as

$$IaOCA = \frac{OCA_t - \frac{\sum_{j=1}^{N_j} OCA_{ij}}{N_j}}{std(OCA_{ij})}$$

where $OCA_t = \frac{OC_t}{at_t}$ is organizational capital-to-assets, in which OC_t is organizational capital defined below in anomaly Org. Capital. Industry-adjusted organizational capital-to-assets is thus firm's org. capital industry demeaned and then divided by the standard deviation of org. capital within its industry.

Industry Concentration Assets (ICA)

Based on Hou and Robinson (2006), Industry Concentration Assets is Herfindahl index (HI), defined above, with total assets at_t as a measure of market share instead of sales $sale_t$.

Industry Concentration Book Equity (ICBE)

Based on Hou and Robinson (2006), Industry Concentration Book Equity is Herfindahl index (HI), defined above, with book equity BE_t defined in anomaly Book Equity / Market Equity.

Org. Capital (OC)

Based on Eisfeldt and Papanikolaou (2013), organizational capital is defined recursively. For the first year of stocks appearance in data, organizational capital is set equal to 4 times selling, general and administrative expense (0 if missing), i.e.

$$OC_{t_0} = 4 * xsga_{t_0}$$

All next years, organizational capital is defined as

$$OC_t = \frac{\frac{0.85*OC_{t-1}+xsga_t}{cpi_t}}{at_t}$$

where cpi_t is and at_t are total assets.

R&D Capital-to-assets (RDCA)

Based on Li (2011), R&D Capital-to-assets is defined as

$$RDCA = \frac{xrd_t + 0.8xrd_{t-1} + 0.6xrd_{t-2} + 0.4xrd_{t-3} + 0.2xrd_{t-4}}{at_t}$$

where xrd_t are R&D expenses and at_t are total assets. Nominator is thus accumulated annual R&D expenses over the past five years with a linear depreciation rate of 20%. Only firms with positive numerator and nonmissing xrd_t are included.

R&D Expenses-to-sales (RDES)

Based on Chan et al. (2001), R&D Expenses-to-sales is defined as

$$RDES = \frac{xrd_t}{sale_t}$$

where xrd_t is research and development expense and $sale_t$ are sales.
Only firms with positive xrd_t are included.

R&D / Market Value of Equity (RDM)

Based on Chan et al. (2001), R&D-to-market value of equity is defined as

$$RDM = \frac{xrd_t}{ME_t}$$

where xrd is research and development expense and $ME_t = prc_t * shrout_t$ is the market equity defined as price times shares outstanding, at the end of the previous year.

ΔSales - ΔInventory (SmI)

Based on Abarbanell and Bushee (1998), change in sales - change in inventory (ΔSales - ΔInventory) is defined as

$$SmI = \frac{sale_t - \frac{sale_{t-1}+sale_{t-2}}{2}}{\frac{sale_{t-1}+sale_{t-2}}{2}} - \frac{invt_t - \frac{invt_{t-1}+invt_{t-2}}{2}}{\frac{invt_{t-1}+invt_{t-2}}{2}}$$

where $sale_t$ is net sales and $invt_t$ is total inventories.
Annual rebalancing frequency.

Tangibility (TAN)

Based on Hahn and Lee (2009), tangibility is defined as

$$TAN = \frac{che_t + 0.715rect_t + 0.547invt_t + 0.535ppegt_t}{at_t}$$

where che_t are cash holdings, $rect_t$ are accounts receivable, $invt_t$ is inventory and $ppegt_t$ is property, plant and equipment.

Unexpected R&D Increases (URDI)

Based on Eberhart et al. (2004), unexpected R&D increases is a binary variable defined as

$$URDI = \begin{cases} 1 & \text{if } (\frac{xrd_t}{revt_t} > 0.05)\ \&\ (\frac{xrd_t}{at_t} > 0.05)\ \&\ (\frac{xrd_t}{xrd_{t-1}} > 1.05)\ \&\ (\frac{\frac{xrd_t}{at_t}}{\frac{xrd_{t-1}}{at_{t-1}}} > 1.05) \\ 0 & \text{otherwise} \end{cases}$$

where xrd_t are R&D expenditures, $revt_t$ is total revenue and at_t is total assets. $URDI = 1$ if R&D scaled by assets and revenue is greater than 5%, the yearly percentage change in R&D expenditures is greater than 5%; and R&D scaled by assets increased by more than 5%.

Whited-Wu Index (WWI)

Based on Whited and Wu (2006), Whited-Wu index is defined as

$$WWI_{it} = -0.091CF_t - 0.062DIVP_t + 0.021LDA_t - 0.044log(at_t) + 0.102ISG_t - 0.035(SG_t)$$

where

$$CF_T = \sqrt[4]{1 + \frac{\frac{ib_t+dp_t}{at_t}}{4}} - 1$$

where ib_t is income before extraordinary items, dp_t is depreciation and amortization, at_t are total assets, $DIVP_t$ is a binary variable equal to one if firm pays cash dividends ($dvpsx_t > 0$) and 0 otherwise, and $LDA_t = \frac{dltt_t}{at_t}$ is the long-term debt to total assets.

$$ISG_t = \frac{(\sum_{i=1}^{N_j} sale_{i,j})_t}{(\sum_{i=1}^{N_j} sale_{i,j})_t}$$

where $sale_{ij}$ is the sale of firm i in industry j and N_j is the total number of firms in the 3-digit SIC code defined industry including at least 3 firms.

$$SG_t = \sqrt[4]{1 + \frac{\frac{sale_t}{sale_{t-1}}}{4}} - 1$$

Investment

Asset Growth (AGr)

Based on Cooper et al. (2008), asset growth is defined as

$$AGr = \frac{at_t}{at_{t-1}}$$

where at_t are total assets.

Change in Net Operating Assets (ChNOA)

Based on Hirshleifer et al. (2004), Change in Net Operating Assets is defined as

$$ChNOA = \frac{NOA_t - NOA_{t-1}}{at_{t-1}}$$

where NOA_t are net operating assets defined below and at_{t-1} are lagged total assets.

Changes in PPE and Inventory-to-Assets (ChPPEIA)

Based on Lyandres et al. (2007), Changes in PPE and Inventory-to-Assets is defined as

$$ChPPEIA_t = \frac{(ppegt_t - ppegt_{t-1}) + (invt_t - invt_{t-1})}{at_{t-1}}$$

wehere $ppegt_t$ is gross property, plant and equipment, $invt_t$ is total inventories and at_{t-1} are lagged total assets.

Composite Debt Issuance (CDI)

Based on Lyandres et al. (2007), Composite Debt Issuance is defined as

$$CDI = log(\frac{dltt_t + dlc_t}{dltt_{t-5} + dlc_{t-5}})$$

where $dltt_t$ is total long-term debt and dlc_t is debt in current liabilities.

ΔCAPEX - ΔIndustry CAPEX (CAPEX)

Based on Abarbanell and Bushee (1998), change in investment minus the change in industry investment (ΔCAPEX - ΔIndustry CAPEX). Where

$$\Delta CAPEX = \frac{capx_{it} - \frac{capx_{i,t-1}+capx_{i,t-2}}{2}}{\frac{capx_{i,t-1}+capx_{i,t-2}}{2}}$$

and ΔIndustry CAPEX is defined analogously for aggregated industry CAPEX. $capx_t$ is capital expenditure.
Stocks in industries with less than 3 firms are excluded.

Debt Issuance (DI)

Based on Spiess and Affleck-Graves (1995), debt issuance is defined as

$$DI = \begin{cases} 1 & \text{if } dltis_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $dltis_t$ is long-term debt/issuance.

Growth in LTNOA (GriLTNOA)

Based on Fairfield et al. (2003), growth in long-term net operating assets is defined as

$$GriLTNOA = NOA_t - NOA_{t-1} - ACCR_t,$$

where NOA_t are net operating assets, defined below and $ACCR_t$ are accruals defined above in category *Accruals*.

Investment (INV)

Based on Titman et al. (2004), investment is defined as

$$INV = \frac{capx_t/revt_t}{avg_{3t}(\frac{capx}{revt})}$$

where $capx_t$ is capital expenditures, $revt_x$ is total revenue and $avg_{3t}()$ is average from the previous three years.
Stocks with revenue ¡ \$10m are excluded.

Net Debt Finance (NDF)

Based on Bradshaw et al. (2006), Net Debt Finance is defined as

$$NDF_t = \frac{dltis_t - dltr_t + dlcch_t}{(at_t + at_{t-1})/2}$$

where $dltis_t$ is long-term debt issuance, $dltr_t$ is long-term debt reduction , $dlcch_t$ are current debt changes and at_t are total assets.

Net Equity Finance (NEF)

Based on Bradshaw et al. (2006), Net Equity Finance is defined as

$$NEF_t = \frac{sstk_t - prstkc_t - dv_t}{(at_t + at_{t-1})/2}$$

where $sstk_t$ is sale of common and preferred stock (0 if missing), $prstkc_t$ is purchase of common and preferred stock (0 if missing) , dv_t are cash dividend, and at_t are total assets.

Net Operating Asset (NOA)

Based on Hirshleifer et al. (2004), net operating assets are defined as

$$NOA = \frac{OA_t - OL_t}{at_{t-1}}$$

OA_t and OL_t are operating assets and operating liabilities defined as $OA_t = at_t - che_t$ and $OL_t = at_t - dlc_t - dltt_t - mib_t - pstkrv_t - ceq_t$, where at_t is total assets, che_t is cash and short-term investment, dlc_t is current portion of long-term debt, $dltt$ is long-term debt, mib_t is minority interest, $pstkrv$ is preferred stock and ceq is common equity.

Noncurrent Operating Assets Changes (NOACh)

Based on Soliman (2008), noncurrent operating assets changes are defined as

$$NOACh = \frac{NCOA_t - NCOA_{t-1}}{at_t}$$

where $NCOA_t$ is noncurrent operating assets. Noncurrent operating assets are defined as

$$NCOA_t = (at_t - act_t - ivaeq_t) - (lt_t - lct_t - dltt_t)$$

, where at_t are total asssets, act_t are current assets, $ivaeq_t$ are investment and advances (0 if missing), lt_t are total liabilities, lct_t are current liabilities and $dltt_t$ is long-term debt.

Share Repurchases (SR)

Based on Ikenberry et al. (1995), share repurchases are defined as binary variable

$$SR = \begin{cases} 1 & \text{if } \; prstkc_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $prstkc_t$ is purchase of common and preferred stock.

Total XFIN (TXFIN)

Based on Bradshaw et al. (2006), total net external financing is defined as

$$TXFIN = \frac{sstk_t - dv_t - prstkc_t + dltis_t - dltr_t}{at_t}$$

where at_t are total assets, $sstk_t$ is sale of common and preferred stock (0 if missing), dv_t are cash dividends, $prstkc_t$ is purchase of common and preferred stock (0 if missing), $dltis_t$ is sale of long-term debt and $dltr_t$ is purchase of long-term debt.

Profitability

Asset Turnover (AT)

Based on Soliman (2008), asset turnover is defined as

$$AT = \frac{sale_t}{avg_{2t}(NOA)}$$

where NOA are net operating assets defined as $NOA = (at_t - che_t) - (lt_t - dltt_t - dlc_t - mib_t)$ and $avg_{2t}(NOA)$ is average NOA from the previous two years. at_t are total assets, che_t is cash and cash equivalents, lt_t are total liabilities, $dltt_t$ is long-term debt, dlc_t is debt in current liabilities, and mib_t is minority interest (0 if missing). Firms with negative NOA and negative operating income ($oiadp$) are exluded.

Capital Turnover (CT)

Based on (Haugen and Baker, 1996), capital turnover is defined as

$$CT = \frac{sale_t}{at_{t-1}}$$

where $sale_t$ is sales and at_{t-1} are one-year lagged total assets.

Cash-based Operating Profitability (CBOP)

Based on Ball et al. (2016), cash-based operating profitability is defined as

$$\begin{aligned} CBOP = \big(&revt_t - cogs_t - xsga_t + xrd_t - (rect_t - rect_{t-1}) - (invt_t - invt_{t-1}) - \\ &-(xpp_t - xpp_{t-1}) + (drc_t + drlt_t - drc_t - drlt_t) + (rect_t - rect_{t-1}) + (ap_t - ap_{t-1}) \\ &+(xacc_t - xacc_{t-1})\big)/at_t \end{aligned}$$

where at_t are total assets, $revt_t$ is total revenue, $cogs_t$ is cost of goods sold, $xsga_t$ are selling, general, and administrative expenses, xrd_t are research and development expenditures (0 if missing), $rect_t$ are accounts receivables, $invt_t$ is inventory, xpp_t are prepaid expenses, drc_t is current deferred revenue, $drlt_t$ is long-term deferred revenue, ap_t are accounts payable and $xacc_t$ are accrued expenses. Changes (in brackets) are all equal to 0 if missing.

Change in Asset Turnover (ChiAT)

Based on Soliman (2008), change in asset turnover is defined as

$$ChiAT = AT_t - AT_{t-1}$$

where AT_t is asset turnover defined above.

Earnings Consistency (EC)

Based on Alwathainani (2009), earnings consistency is defined as

$$EC = \sqrt[5]{\Pi_{i=1}^{5}(1 + eg_i)} - 1$$

where eg_i is earnings growth is defined as

$$eg_t = \frac{epspx_t - epspx_{t-1}}{\frac{|epsx_t|+|epsx_{t-1}|}{2}}$$

where $epspx_t$ are earnings per share excluding extraordinary items. Stocks with $|eg_t| > 6$ are deleted. Also stocks with the previous two earnings growths with opposite signs are excluded $(eg_t * eg_{t-1})$

Earnings / Price (EP)

Based on (Basu, 1977), earnings-to-price is defined as

$$EP = \frac{ib_t}{ME_t}$$

where ib_t is income before extraordinary items and $ME_t = prc_t * shrout_t$ is market equity, i.e. price times shares outstanding.
Firms with $ib_t \leq 0$ are excluded.

F-Score (FSc)

Based on Piotroski (2000), F-score is defined as the sum of nine binary variables (F1-F9) and is further limited only to firms in the highest quintile with respect to book-to-market

$$F = \sum_{i=1}^{9} F_i$$

Binary variables are defined as

$$\begin{aligned}
F1 = &\ \ 1 \text{ if } ni_t > 0;\ \ 0 \text{ otherwise} \\
F2 = &\ \ 1 \text{ if } oancf_t > 0\ ; 0 \text{ otherwise} \\
F3 = &\ \ 1 \text{ if } \tfrac{ni_t}{at_t} > \tfrac{ni_{t-1}}{at_{t-1}}\ ; 0 \text{ otherwise} \\
F4 = &\ \ 1 \text{ if } oancf_t > ni_t\ ; 0 \text{ otherwise} \\
F5 = &\ \ 1 \text{ if } \tfrac{dltt_t}{at_t} < \tfrac{dltt_{t-1}}{at_{t-1}}\ ; 0 \text{ otherwise} \\
F6 = &\ \ 1 \text{ if } \tfrac{act_t}{lct_t} > \tfrac{act_{t-1}}{lct_{t-1}}\ ; 0 \text{ otherwise} \\
F7 = &\ \ 1 \text{ if } sstk_t - (pstk_t - pstk_{t-1}) \leq 0\ ; 0 \text{ otherwise} \\
F8 = &\ \ 1 \text{ if } \tfrac{oiadp_t}{sale_t} > \tfrac{oiadp_{t-1}}{sale_{t-1}}\ ; 0 \text{ otherwise} \\
F9 = &\ \ 1 \text{ if } \tfrac{sale_t}{at_t} > \tfrac{sale_{t-1}}{at_{t-1}}\ ; 0 \text{ otherwise}
\end{aligned}$$

where ni_t is net income, $oancf_t$ is cash-flow from operating activities, at_t are total assets, $dltt_t$ is long term debt, act_t is current assets, lct_t are current liabilities, $sstk_t$ is sale of common and preferred stock, $pstk_t$ is total preferred stock, $oiadp_t$ is operating income after depreciation, and $sale_t$ is net sales.

Gross Profitability (GP)

Based on Novy-Marx (2013), gross profitability is defined as

$$GP = \frac{revt_t - cogs_t}{at_{t-1}}$$

where $revt_t$ is total revenue, $cogs_t$ is cost of goods sold, and at_{t-1} are total assets lagged by one year.

Operating Profits to Assets (OPtA)

Based on Ball et al. (2016), operating profits to assets are defined as

$$OPtA = \frac{revt_t - cogs_t - xsga_t + xrd_t}{at_t}$$

where $revt_t$ is total revenue, $cogs_t$ is cost of goods sold, $xsga_t$ is SG&A, xrd_t are research and development expenditures, and at_t are total assets.

Operating Profits to Assets (OPtE)

Based on Fama and French (2015), operating profits to equity are defined as

$$OPtE = \frac{revt_t - cogs_t - xsga_t + xint_t}{be_t}$$

where $revt_t$ is total revenue, $cogs_t$ is cost of goods sold, $xsga_t$ is SG&A, $xint_t$ is interest and related expense (total), and be_t is book equity defined in Book Equity / Market Equity variable. At least one from xint, cogs, xsga cannot be missing and the missing values are filled with zeros.

Leverage (Lvrg)

Based on Bhandari (1988), leverage is defined as

$$Lvrg = \frac{dltt_t + dlc_t}{ME_t}$$

where $dltt_t$ is long-term debt, dlc_t is debt in current liabilities and $ME_t = prc_t * shrout_t$ is market equity defined in anomaly of earnings/price.

O-Score (OSc)

Based on Dichev (1998), O-score is defined as

$$OSc = -1.32 - 0.4078log(\frac{at_t}{cpi_t}) + 6.03 * (\frac{dltt_t + dlc_t}{at_t}) - 1.43 * (\frac{act_t - lct_t}{at_t}) + 0.076 * (\frac{lct_t}{act_t}) -$$
$$-1.72 * (OENEG_t) - 2.37 * (\frac{ni_t}{at_t}) - 1.83 * (\frac{pi_t}{dp_t}) + 0.285 * (INTWO_t) - 0.521 * (\frac{ni_t - ni_{t-1}}{|ni_t| + |ni_{t-1}|})$$

where at_t are total assets, cpi_t is inflation, $dltt_t$ are long-term liabilities, dlc_t are short-term liabilities, act_t are current assets, lct_t are current liabilities, $OENEG_t$ is binary variable equal to one if $lt_t > at_t$ and 0 otherwise, ni_t is net income, $INTWO_t$ is binary variable equal to one if stock has negative net income in both previous years and 0 otherwise. Only stocks with SIC codes from 1 to 3999 and from 5000 to 5999 are included.

Return on Net Operating Assets (RNOA)

Based on Soliman (2008), return on net operating assets is defined as

$$RNOA = \frac{oiadp_t}{NOA_{t-1}}$$

where NOA are net operating assets defined as $NOA_t = (at_t - che_t) - (lt_t - dltt_t - dlc_t - mib_t)$. at_t are total assets, che_t is cash and cash equivalents, lt_t are total liabilities, $dltt_t$ is long-term debt, and dlc_t is debt in current liabilities and mib_t is minority interest (0 if missing).
Firms with negative NOA and negative operating income ($oiadp$) are exluded.

Value

Assets-to-Market (AM)

Based on Fama and French (1992), assets-to-market is defined as

$$AM = \frac{at_t}{ME_t}$$

where at_t are assets total and ME_t is market equity.

Book Equity / Market Equity (BM)

Based on Fama and French (1992), book-to-market equity is defined as

$$BM = log(\frac{BE_t}{ME_t})$$

Market equity is price times shares outstanding, $ME_t = prc_t * shrout_t$. Book equity is defined conditional on missing items as

$$BE_t = seq_t - PS_t$$

where seq_t is total stockholders' equity, if missing then $seq_t = ceq_t + pstk_t$, or $seq_t = at_t - lt_t$, where ceq_t is tangible common equity, $pstk_t$ is preferred stock using liquidating value, at_t are total assets, lt_t are total liabilities,and PS_t is preferred stock measured using (ordered on availability) redemption, liquidating or par value, i.e. $pstkrv_t, pstkl_t, pstk_t$.

Cash Flow / Market Value of Equity (CM)

Based on Lakonishok et al. (1994), cash flow to market value of equity is defined as

$$CM = \frac{ib_t + dp_t}{ME_t}$$

where ib_t is net income, dp_t is depreciation and amortization and ME_t is market equity defined above in book-to-market equity anomaly.

Duration of Equity (DurE)

Based on Dechow et al. (2004), duration of equity is defined as

$$DurE_t = \frac{58}{3} + \frac{1}{MC_t}\sum_{j=1}^{10}\frac{cd_j j(j - 58/3)}{1.12}$$

where cd_j is defined recursively from the following equations: $g_{j+1} = 0.06 + 0.24g_j$, $be_j = be_0(1 + g_j)$, $roe_{j+1} = 0.12 + 0.57roe_j$, and $cd_j = roe_j be_{j-1}$. The starting values are $be_0 = ceq_t$, $roe_0 = \frac{ib_t}{ceq_{t-1}}$, and $g_0 = \frac{sale_t}{sale_{t-1}} - 1$. be_t is the book equity, ceq_{t-1} is a lag of common equity, ib_t are earnings, and $sale_t$ are net sales.

Enterprise Component of Book/Price (ECoBP)

Based on Penman et al. (2007), enterprise component of book/price is defined as

$$ECoBP = \frac{BE_t + ND_t}{ND_t + ME_t}$$

where BE_t and ME_t are book value of equity and market equiy, defined above in book-to-market equity anomaly. $ND_t = dltt_t + dlc_t + pstk_t + dvpa_t - tstkp_t - che_t$ is net debt, where che_t is cash and short-term investments, $dltt_t$ is long-term debt, dlc_t is debt in current liabilities, $pstk_t$ is preferred stock, $dvpa_t$ is preferred dividends in arrears and $tstkp_t$ is preferred treasury stock.

Enterprise Multiple (EM)

Based on Loughran and Wellman (2011), enterprise multiple is defined as

$$EM = \frac{EV_t}{oibdp_t}$$

where $oibdp_t$ is operating cash flow and EV_t is enterprise value defined as $EV_t = ME_t + dltt_t + dlc_t + pstk_t + dvpa_t - tstkp_t - che_t$. ME_t is market equity defined above in book-to-market equity anomaly, $dltt_t$ is long-term debt, dlc_t is debt in current liabilities, $pstk_t$ is preferred stock, $dvpa_t$ is preferred dividends in arrears, $tstkp_t$ is preferred treasury stock and che_t is cash and short-term investments.

Intangible Return (IR)

Based on Daniel and Titman (2006), intangible return is defined as residual from the following cross-sectional regression

$$log(r_{t-5,t}) = \beta_0 + \beta_1 BM_{t-5} + \beta_2 log(RB_{t-5,t}) + \epsilon_t$$

where $r_{t-5,t}$ is 5- year stock return, BM_{t-5} is 5-year-lagged book-to-market defined in anomaly Book Equity / Market Equity and $RB_{t-5,t} = log(\frac{BE_t}{BE_{t-5}) - \sum_{p=t-5}^{t-1}(r_p - log(\frac{P_p}{P_{p-1}}))})$ in which BE_t is the book equity defined in anomaly Book Equity / Market Equity , r_p is the stock return for year p and P_p is the price at the end of year p.

Leverage Component of Book/Price (LCoBP)

Based on Penman et al. (2007), leverage component of book/price is defined as

$$LCoBP = BE_t - ECoBP_t$$

where BE_t is book value of equity defined above in book-to-market equity anomaly, and $ECoBP_t$ is enterprise component of book/price defined above.

Net Payout Yield (NPY)

Based on Boudoukh et al. (2007), net payout yield is defined as

$$NPY = \frac{dvc_t + prstkc_t - sstk_t}{ME_t}$$

where dvc_t are dividends common/ordinary, $prstkc_t$ is purchase of common and preferred stock, $sstk_t$ is sale of common and preferred stock, and ME_t is market equity.

Operating Leverage (OL)

Based on Novy-Marx (2010), operating leverage is defined

$$OL = \frac{xsga_t + cogs_t}{at_t}$$

where $xsga_t$ is SG&A, $cogs_t$ is cost of goods sold, and at_t are total assets.

Payout Yield (PY)

Based on Boudoukh et al. (2007), payout yield is defined as

$$PY = \frac{dvc_t + prstkc_t - (pstkrv_t + pstkrv_{t-1})}{ME_t}$$

where dvc_t are dividends common/ordinary, $prstkc_t$ is purchase of common and preferred stock, $pstkrv_t$ is preferred stock/redemption, and ME_t is market equity.

Sales Growth (SaGr)

Based on Lakonishok et al. (1994), sales growth is defined as

$$SaGr = \frac{5SGR_t + 4SGR_{t-1} + 3SGR_{t-2} + 2SGR_{t-3} + 1SGR_{t-4}}{15}$$

where SGR_t is the rank of firm in year t based on the simple sales growth defined as $SG = sale_t/sale_{t-1}$.

Sustainable Growth (SuGr)

Based on Lockwood and Prombutr (2010), sustainable growth is defined as $SuGr = BE_t/BE_{t-1}$, where BE_t is book equity defined above in book-to-market equity anomaly.

Sales/Price (SP)

Based on Barbee Jr et al. (1996), sales-to-price is defined as $SP = revt_t/ME_t$, where $revt_t$ is total revenue and ME_t is the market equity defined above in the book-to-market equity anomaly.

Appendix C

Classification of Industries in Datastream

Table C.1:
Industries in the Datastream Level 3 Classification and Corresponding Four-digit SIC

Datastream lvl 3 industry	SIC codes
Automobiles & Parts	3011, 3510, 3714, 3751, 5013
Basic Resources	800, 1000, 1040, 1090, 1220, 1221, 2421, 2600, 2611, 2621, 2631, 3310, 3312, 3317, 3330, 3334, 3350, 3360, 3444, 3460, 3720, 5050, 5051
Chemicals	2810, 2820, 2821, 2833, 2851, 2860, 2870, 2890, 2891, 2990, 3080, 3081, 3341, 5160
Construct. & Material	1400, 1540, 1600, 1623, 1731, 2400, 2430, 2950, 3211, 3231, 3241, 3250, 3270, 3272, 3281, 3290, 3430, 3440, 3442, 3448, 5031, 5070, 5072
Financial Services(3)	6111, 6141, 6153, 6159, 6162, 6163, 6172, 6189, 6200, 6211, 6221, 6282, 6361, 6500, 6510, 6770, 6795, 6798, 6799, 8880, 8888, 9995
Food & Beverage	100, 200, 900, 2000, 2011, 2013, 2015, 2020, 2024, 2030, 2033, 2040, 2050, 2052, 2060, 2070, 2080, 2082, 2086, 2090, 2092
Healthcare	2590, 2800, 2834, 2835, 2836, 3060, 3821, 3826, 3841, 3842, 3843, 3844, 3845, 3851, 4100, 5047, 6324, 8000, 8011, 8050, 8051, 8060, 8062, 8071, 8082, 8090, 8093, 8300, 8731
Ind. Goods & Services	1700, 2390, 2650, 2670, 2673, 2750, 2761, 3050, 3086, 3089, 3221, 3320, 3357, 3390, 3411, 3412, 3443, 3451, 3452, 3470, 3480, 3490, 3523, 3524, 3530, 3531, 3532, 3537, 3540, 3541, 3550, 3555, 3560, 3561, 3562, 3564, 3567, 3569, 3575, 3580, 3585, 3590, 3600, 3612, 3613, 3620, 3621, 3634, 3640, 3669, 3670, 3672, 3677, 3678, 3679, 3690, 3711, 3713, 3715, 3721, 3724, 3728, 3730, 3743, 3760, 3812, 3822, 3823, 3824, 3825, 3827, 3829, 3861, 3910, 4011, 4013, 4210, 4213, 4231, 4400, 4412, 4513, 4700, 4731, 4950, 4953, 4955, 4961, 5000, 5063, 5065, 5080, 5082, 5084, 5090, 5099, 6099, 6794, 7320, 7350, 7359, 7361, 7363, 7374, 7377, 7380, 7381, 7384, 7385, 7389, 7829, 8111, 8200, 8351, 8600, 8700, 8711, 8734, 8741, 8742, 8744, 9721
Insurance	6311, 6321, 6331, 6351, 6411
Media	2711, 2721, 2731, 2732, 2741, 2780, 4832, 4833, 4841, 7310, 7311, 7330, 7331, 7819, 7822, 8900
Oil & Gas	1311, 1381, 1382, 1389, 2911, 3533, 4522, 4610, 4900, 5171, 5172, 6792
Pers & Househld Goods	1531, 2100, 2111, 2200, 2211, 2221, 2250, 2253, 2273, 2300, 2320, 2330, 2340, 2451, 2452, 2510, 2511, 2520, 2522, 2531, 2540, 2771, 2840, 2842, 2844, 3021, 3100, 3220, 3260, 3420, 3433, 3630, 3651, 3716, 3790, 3873, 3911, 3931, 3942, 3944, 3949, 3950, 3960, 5020, 5030, 5064, 5130, 5150, 5190, 6552
Real Estate	6519, 6531
Retail	700, 2790, 3140, 4220, 5094, 5010, 5110, 5122, 5140, 5141, 5180, 5200, 5211, 5271, 5311, 5331, 5399, 5400, 5411, 5412, 5500, 5531, 5600, 5621, 5651, 5661, 5700, 5712, 5731, 5734, 5735, 5912, 5940, 5944, 5945, 5960, 5961, 5990, 6399, 7200, 7340, 7500, 7600, 7841
Technology	3559, 3570, 3571, 3572, 3576, 3577, 3578, 3579, 3661, 3663, 3674, 3695, 4899, 5040, 5045, 7370, 7371, 7372, 7373
Telecommunications	4812, 4813, 4822
Travel & Leisure	1520, 3652, 3990, 4512, 4581, 5810, 5812, 6512, 6513, 6532, 7000, 7011, 7510, 7812, 7830, 7900, 7948, 7990, 7997
Utilities	4911, 4922, 4923, 4924, 4931, 4932, 4941, 4991, 5900
Banks	6021, 6022, 6029, 6035, 6036, 6199

Appendix D

Detailed Results for Individual Anomalies

The following tables are constructed as described in the corresponding aggregated tables in the main text of this study. The significance is determined with t-test of mean returns. The standard errors in t-statistics are adjusted for autocorrelation and heteroskedasticity, as in Newey and West (1987), with 12 lags. The significance of difference of means of two return time-series is determined in a t-test of time-series of differences of returns of the two series.

Table D.1:
Impact of Delisting in Compustat - Detailed

	Our Delisting vs No Delisting				Our Delisting vs Shumway (1997)			
	Corr	Our	No delisting	Diff	Corr	Our	Shumway	Diff
Accruals								
Acc	1.000	0.29 (2.26)	0.31 (2.42)	(6.33)	1.000	0.29 (2.26)	0.29 (2.24)	(-3.60)
ChCE	0.999	0.38 (2.17)	0.42 (2.41)	(5.45)	1.000	0.38 (2.17)	0.36 (2.08)	(-2.81)
ChCOA	0.998	0.53 (5.04)	0.55 (5.17)	(2.48)	0.999	0.53 (5.04)	0.53 (4.96)	(-1.77)
ChCOL	0.999	0.36 (4.00)	0.37 (4.07)	(2.51)	1.000	0.36 (4.00)	0.36 (3.98)	(-1.60)
ChFL	0.999	0.56 (7.88)	0.55 (7.83)	(-2.21)	1.000	0.56 (7.88)	0.56 (7.91)	(2.59)
ChLTI	0.992	0.13 (2.91)	0.14 (3.03)	(1.06)	0.995	0.13 (2.91)	0.12 (2.73)	(-1.39)
ChNCOA	1.000	0.68 (5.46)	0.70 (5.60)	(4.20)	1.000	0.68 (5.46)	0.68 (5.45)	(-1.95)
ChNCOL	0.997	0.17 (2.01)	0.18 (2.15)	(2.04)	0.998	0.17 (2.01)	0.17 (1.91)	(-1.48)
ChNFA	0.997	0.42 (6.10)	0.41 (5.99)	(-2.71)	1.000	0.42 (6.10)	0.42 (6.13)	(2.82)
ChNNCOA	0.999	0.71 (5.97)	0.72 (6.04)	(1.99)	0.999	0.71 (5.97)	0.71 (6.00)	(0.57)
ChNNCWC	0.997	0.35 (4.19)	0.36 (4.33)	(2.00)	0.998	0.35 (4.19)	0.34 (4.09)	(-1.49)
GriI	0.998	0.48 (5.34)	0.50 (5.49)	(2.23)	0.998	0.48 (5.34)	0.48 (5.27)	(-1.45)
ICh	0.998	0.50 (5.26)	0.52 (5.43)	(2.56)	0.999	0.50 (5.26)	0.50 (5.18)	(-1.61)
IGr	0.998	0.55 (5.22)	0.56 (5.38)	(2.47)	0.999	0.55 (5.22)	0.54 (5.15)	(-1.50)
MBaAC	0.996	1.43 (7.02)	1.43 (6.97)	(-0.01)	1.000	1.43 (7.02)	1.42 (7.00)	(-1.62)
NWCCh	0.994	0.49 (6.76)	0.48 (6.77)	(-0.42)	0.997	0.49 (6.76)	0.48 (6.66)	(-0.84)
POA	0.999	0.70 (5.51)	0.69 (5.50)	(-1.58)	1.000	0.70 (5.51)	0.70 (5.52)	(1.73)
PTA	0.999	0.33 (3.20)	0.35 (3.37)	(4.26)	1.000	0.33 (3.20)	0.33 (3.19)	(-0.41)
TA	0.999	0.44 (3.56)	0.48 (3.84)	(5.15)	0.999	0.44 (3.56)	0.44 (3.46)	(-2.17)
Intangibles								
AL	0.999	0.44 (2.74)	0.45 (2.81)	(1.33)	0.999	0.44 (2.74)	0.44 (2.69)	(-1.22)
ChGMChS	0.999	0.24 (3.55)	0.22 (3.34)	(-5.01)	1.000	0.24 (3.55)	0.24 (3.56)	(2.13)
EPr	0.998	0.66 (4.60)	0.65 (4.42)	(-2.13)	1.000	0.66 (4.60)	0.67 (4.61)	(2.29)
ES	1.000	0.21 (1.00)	0.23 (1.10)	(4.40)	1.000	0.21 (1.00)	0.21 (0.99)	(-0.77)
HI	0.999	0.06 (0.50)	0.07 (0.52)	(0.14)	0.999	0.06 (0.50)	0.06 (0.47)	(-0.88)
HR	0.999	0.42 (3.75)	0.43 (3.88)	(3.47)	1.000	0.42 (3.75)	0.41 (3.72)	(-3.42)
IARER	0.999	0.31 (2.43)	0.31 (2.47)	(0.45)	1.000	0.31 (2.43)	0.31 (2.45)	(1.02)
ICBE	0.999	0.12 (0.86)	0.13 (0.90)	(0.84)	0.999	0.12 (0.86)	0.12 (0.83)	(-1.00)
OC	1.000	0.45 (2.57)	0.46 (2.65)	(3.66)	1.000	0.45 (2.57)	0.45 (2.56)	(-1.38)
RDM	1.000	1.19 (4.16)	1.21 (4.21)	(3.82)	1.000	1.19 (4.16)	1.19 (4.14)	(-2.64)
SmI	0.996	0.37 (5.39)	0.37 (5.42)	(0.10)	0.997	0.37 (5.39)	0.36 (5.35)	(-0.56)
TAN	0.999	0.29 (1.81)	0.28 (1.77)	(-1.37)	1.000	0.29 (1.81)	0.28 (1.79)	(-0.50)
URDI	1.000	0.47 (2.42)	0.46 (2.37)	(-2.54)	1.000	0.47 (2.42)	0.47 (2.41)	(-1.33)
WWI	0.999	0.36 (1.76)	0.41 (1.98)	(4.51)	1.000	0.36 (1.76)	0.35 (1.71)	(-1.67)
Investment								
AGr	0.999	0.63 (3.90)	0.66 (4.05)	(3.97)	1.000	0.63 (3.90)	0.62 (3.83)	(-2.36)
CAPEX	0.998	0.37 (4.67)	0.38 (4.74)	(2.15)	1.000	0.37 (4.67)	0.37 (4.65)	(-1.96)
CDI	1.000	0.21 (2.22)	0.21 (2.18)	(-1.11)	1.000	0.21 (2.22)	0.21 (2.24)	(1.92)
CEI5Y	0.999	0.28 (2.30)	0.27 (2.30)	(-0.33)	1.000	0.28 (2.30)	0.28 (2.30)	(0.80)
ChNOA	0.997	0.27 (4.09)	0.27 (4.10)	(0.86)	1.000	0.27 (4.09)	0.27 (4.08)	(-1.49)
ChPPEIA	1.000	0.63 (5.32)	0.65 (5.44)	(3.51)	1.000	0.63 (5.32)	0.63 (5.31)	(-2.89)
DI	0.998	0.25 (3.78)	0.25 (3.74)	(-1.62)	0.999	0.25 (3.78)	0.25 (3.75)	(0.01)
GrLTNOA	0.999	0.61 (4.41)	0.62 (4.49)	(2.94)	1.000	0.61 (4.41)	0.60 (4.39)	(-3.44)
INV	0.996	0.27 (3.80)	0.28 (3.85)	(0.53)	0.997	0.27 (3.80)	0.28 (3.82)	(0.94)
NDF	0.998	0.34 (4.29)	0.34 (4.31)	(-0.04)	1.000	0.34 (4.29)	0.34 (4.28)	(0.02)
NEF	1.000	0.72 (3.16)	0.69 (3.07)	(-6.04)	1.000	0.72 (3.16)	0.72 (3.17)	(2.81)
NOA	0.998	0.53 (4.99)	0.54 (5.12)	(2.52)	0.999	0.53 (4.99)	0.52 (4.91)	(-1.60)
NOACh	1.000	0.55 (4.03)	0.55 (4.03)	(-0.05)	1.000	0.55 (4.03)	0.55 (4.03)	(-0.50)
SR	0.997	0.20 (2.66)	0.17 (2.31)	(-4.80)	0.999	0.20 (2.66)	0.20 (2.76)	(2.31)
TXFIN	1.000	0.89 (4.80)	0.88 (4.71)	(-3.86)	1.000	0.89 (4.80)	0.90 (4.81)	(1.49)
Profitability								
AT	1.000	0.26 (2.25)	0.26 (2.25)	(-0.26)	1.000	0.26 (2.25)	0.26 (2.25)	(0.34)
CBOP	1.000	0.53 (3.30)	0.53 (3.30)	(-0.19)	1.000	0.53 (3.30)	0.53 (3.31)	(1.79)
CT	1.000	0.28 (1.97)	0.27 (1.96)	(-0.77)	1.000	0.28 (1.97)	0.28 (1.98)	(1.41)
ChiAT	0.999	0.21 (3.60)	0.21 (3.54)	(-1.46)	1.000	0.21 (3.60)	0.21 (3.63)	(1.13)
EC	1.000	0.20 (2.69)	0.20 (2.65)	(-1.63)	1.000	0.20 (2.69)	0.20 (2.70)	(1.01)
EP	0.998	0.72 (5.32)	0.71 (5.28)	(-1.42)	0.999	0.72 (5.32)	0.72 (5.38)	(1.15)
FSc	0.999	0.45 (2.99)	0.41 (2.74)	(-5.17)	1.000	0.45 (2.99)	0.45 (3.05)	(3.22)
GP	0.999	0.34 (2.49)	0.32 (2.32)	(-4.46)	0.999	0.34 (2.49)	0.34 (2.55)	(1.93)
Lvrg	1.000	0.30 (1.81)	0.32 (1.95)	(6.01)	1.000	0.30 (1.81)	0.29 (1.80)	(-2.69)
OPtA	0.999	0.56 (2.93)	0.51 (2.70)	(-5.88)	0.999	0.56 (2.93)	0.58 (3.00)	(3.13)
OPtE	1.000	0.34 (1.78)	0.31 (1.63)	(-4.68)	1.000	0.34 (1.78)	0.35 (1.83)	(2.64)
OSc	1.000	0.08 (0.39)	0.03 (0.15)	(-5.71)	1.000	0.08 (0.39)	0.09 (0.44)	(4.25)
Value								
AM	1.000	0.88 (4.61)	0.91 (4.71)	(4.60)	1.000	0.88 (4.61)	0.88 (4.60)	(-2.60)
BM	1.000	0.98 (5.75)	0.99 (5.77)	(1.36)	1.000	0.98 (5.75)	0.98 (5.74)	(-1.78)
CM	1.000	0.87 (4.05)	0.83 (3.85)	(-5.10)	1.000	0.87 (4.05)	0.88 (4.11)	(3.11)
DurE	1.000	0.94 (4.47)	0.94 (4.45)	(-1.24)	1.000	0.94 (4.47)	0.94 (4.47)	(0.40)
ECoBP	1.000	0.79 (4.36)	0.79 (4.37)	(0.12)	1.000	0.79 (4.36)	0.79 (4.36)	(-0.25)
EM	0.999	0.26 (1.74)	0.28 (1.84)	(2.31)	1.000	0.26 (1.74)	0.26 (1.72)	(-0.45)
IR	1.000	0.49 (2.67)	0.52 (2.84)	(5.70)	1.000	0.49 (2.67)	0.48 (2.64)	(-3.20)
LCoBP	1.000	0.39 (3.14)	0.39 (3.14)	(-0.32)	1.000	0.39 (3.14)	0.39 (3.14)	(0.67)
NPY	1.000	0.90 (4.15)	0.88 (4.05)	(-4.85)	1.000	0.90 (4.15)	0.91 (4.18)	(3.23)
OL	1.000	0.46 (2.90)	0.47 (2.97)	(3.29)	1.000	0.46 (2.90)	0.46 (2.90)	(-0.27)
PY	1.000	0.34 (1.94)	0.33 (1.85)	(-4.75)	1.000	0.34 (1.94)	0.35 (1.99)	(3.83)
SP	1.000	1.01 (4.42)	1.03 (4.49)	(3.08)	1.000	1.01 (4.42)	1.01 (4.42)	(-0.92)
SaGr	0.998	0.24 (2.30)	0.25 (2.39)	(2.06)	1.000	0.24 (2.30)	0.24 (2.28)	(-1.77)
SuGr	0.999	0.17 (1.26)	0.20 (1.52)	(5.09)	0.999	0.17 (1.26)	0.16 (1.19)	(-2.07)

Table D.2:
Datastream vs Compustat in the Common Sample - Panel A - Detailed

	Signals		Portfolios			
	Pears Corr	Spear Corr	Corr	CS	DS	Diff
Accruals						
Acc	0.993	0.987	0.997	0.55 (2.63)	0.54 (2.61)	(-0.19)
ChCE	0.966	0.981	0.995	1.16 (3.88)	1.14 (3.86)	(-1.30)
ChCOA	0.951	0.981	0.989	0.68 (3.69)	0.66 (3.61)	(-0.92)
ChCOL	0.943	0.972	0.979	0.49 (2.87)	0.46 (2.75)	(-1.13)
ChFL	0.932	0.957	0.939	0.26 (2.98)	0.27 (3.21)	(0.13)
ChLTI	0.787	0.640	0.659	0.24 (2.48)	0.18 (1.88)	(-1.02)
ChNCOA	0.966	0.946	0.987	0.93 (4.12)	0.94 (4.36)	(0.17)
ChNCOL	0.847	0.843	0.864	0.33 (2.72)	0.26 (2.39)	(-1.06)
ChNFA	0.882	0.884	0.883	-0.05 (-0.47)	0.02 (0.23)	(1.73)
ChNNCOA	0.967	0.960	0.988	0.86 (4.21)	0.89 (4.44)	(0.65)
ChNNCWC	0.944	0.969	0.967	0.45 (3.52)	0.49 (3.77)	(1.63)
GriI	0.917	0.960	0.979	0.42 (2.87)	0.38 (2.69)	(-1.75)
ICh	0.935	0.970	0.981	0.51 (3.15)	0.48 (3.04)	(-1.21)
IGr	0.888	0.966	0.978	0.52 (3.04)	0.44 (2.70)	(-2.45)
MBaAC	0.895	0.895	0.972	1.85 (4.87)	1.81 (4.69)	(-0.79)
NWCCh	0.938	0.969	0.968	0.31 (3.65)	0.33 (3.67)	(1.03)
POA	0.868	0.981	0.978	0.42 (3.30)	0.48 (3.88)	(2.65)
PTA	0.883	0.965	0.971	0.52 (4.62)	0.48 (4.27)	(-1.58)
TA	0.913	0.924	0.979	1.02 (3.55)	0.97 (3.58)	(-1.39)
Intangibles						
AL	0.820	0.874	0.969	0.45 (1.56)	0.64 (1.72)	(1.51)
ChGMChS	0.318	0.804	0.841	-0.15 (-1.59)	-0.16 (-1.62)	(-0.07)
EPr	0.934	0.961	0.976	0.21 (1.12)	0.23 (1.13)	(0.27)
ES	0.994	0.997	0.999	0.83 (2.64)	0.82 (2.64)	(-0.99)
HI	0.721	0.747	0.718	-0.13 (-0.69)	-0.11 (-0.68)	(0.22)
HR	0.858	0.925	0.972	0.73 (3.63)	0.72 (3.79)	(-0.33)
IARER	0.660	0.633	0.681	0.75 (2.33)	0.45 (1.52)	(-1.43)
ICBE	0.615	0.644	0.631	-0.20 (-0.52)	-0.07 (-0.34)	(0.45)
OC	0.738	0.963	0.986	0.62 (2.40)	0.58 (2.33)	(-0.82)
RDM	0.889	0.977	0.998	1.87 (3.09)	1.86 (3.14)	(-0.15)
SmI	0.931	0.958	0.976	0.09 (0.91)	0.04 (0.40)	(-2.54)
TAN	0.964	0.976	0.997	0.40 (1.17)	0.39 (1.15)	(-0.29)
URDI	0.882	0.882	0.991	0.54 (1.77)	0.62 (1.89)	(1.94)
WWI	0.998	0.998	0.998	1.11 (3.12)	1.13 (3.24)	(0.62)
Investment						
AGr	0.986	0.985	0.995	1.16 (3.84)	1.12 (3.80)	(-1.12)
CAPEX	0.723	0.943	0.969	0.59 (4.14)	0.58 (4.42)	(-0.08)
CDI	0.980	0.982	0.963	-0.08 (-0.57)	-0.11 (-0.74)	(-0.75)
CEI5Y	0.893	0.954	0.995	0.19 (0.94)	0.17 (0.86)	(-1.01)
ChNOA	0.976	0.957	0.954	0.32 (2.70)	0.33 (2.87)	(0.25)
ChPPEIA	0.894	0.969	0.990	0.64 (3.52)	0.62 (3.53)	(-0.88)
DI	0.920	0.920	0.981	0.18 (2.29)	0.18 (2.36)	(0.40)
GrLTNOA	0.784	0.958	0.995	0.77 (4.29)	0.78 (4.29)	(0.74)
INV	0.636	0.895	0.918	0.44 (4.02)	0.41 (3.73)	(-0.65)
NDF	0.939	0.963	0.923	0.12 (1.35)	0.18 (2.31)	(1.42)
NEF	0.975	0.979	0.999	-0.05 (-0.13)	-0.06 (-0.17)	(-0.59)
NOA	0.972	0.980	0.995	0.73 (2.58)	0.72 (2.60)	(-0.21)
NOACh	0.981	0.990	0.995	0.46 (2.32)	0.46 (2.27)	(-0.27)
SR	0.956	0.956	0.993	-0.17 (-1.74)	-0.19 (-1.88)	(-1.22)
TXFIN	0.954	0.973	0.994	0.22 (0.78)	0.18 (0.68)	(-1.20)
Profitability						
AT	0.931	0.991	0.993	0.20 (1.35)	0.21 (1.44)	(0.35)
CBOP	0.822	0.899	0.809	0.50 (1.75)	0.32 (1.23)	(-1.29)
CT	0.988	0.994	0.998	0.05 (0.22)	0.06 (0.22)	(0.11)
ChiAT	0.859	0.961	0.955	0.10 (0.87)	0.14 (1.24)	(1.30)
EC	0.952	0.961	0.952	0.03 (0.25)	0.03 (0.28)	(0.19)
EP	0.696	0.972	0.994	0.51 (2.12)	0.50 (2.12)	(-0.27)
FSc	0.962	0.960	0.951	-0.36 (-1.29)	-0.36 (-1.30)	(0.05)
GP	0.818	0.904	0.962	-0.02 (-0.10)	-0.03 (-0.15)	(-0.17)
Lvrg	0.627	0.991	0.998	0.49 (1.29)	0.48 (1.25)	(-0.55)
OPtA	0.854	0.934	0.975	-0.03 (-0.10)	-0.00 (-0.01)	(0.53)
OPtE	0.769	0.870	0.990	-0.33 (-0.78)	-0.31 (-0.73)	(0.36)
OSc	0.770	0.978	0.997	-1.04 (-2.99)	-1.01 (-2.91)	(1.32)
Value						
AM	0.641	0.993	0.998	1.28 (3.03)	1.26 (3.01)	(-0.99)
BM	0.962	0.985	0.996	1.18 (3.46)	1.14 (3.28)	(-2.49)
CM	0.655	0.968	0.999	-0.09 (-0.17)	-0.06 (-0.12)	(0.89)
DurE	0.653	0.982	0.997	0.75 (2.25)	0.74 (2.26)	(-0.55)
ECoBP	0.598	0.980	0.997	0.80 (2.10)	0.74 (1.96)	(-2.58)
EM	0.347	0.907	0.991	0.79 (2.64)	0.76 (2.52)	(-0.89)
IR	0.985	0.989	0.997	1.12 (3.68)	1.12 (3.72)	(0.16)
LCoBP	0.394	0.956	0.994	0.41 (1.36)	0.35 (1.25)	(-1.51)
NPY	0.683	0.965	0.996	-0.01 (-0.03)	-0.02 (-0.06)	(-0.32)
OL	0.987	0.988	0.977	0.64 (3.06)	0.59 (2.72)	(-1.52)
PY	0.627	0.947	0.987	-0.34 (-1.50)	-0.34 (-1.57)	(0.13)
SP	0.663	0.992	0.999	1.20 (2.56)	1.18 (2.52)	(-1.36)
SaGr	0.991	0.991	0.995	0.26 (1.48)	0.26 (1.53)	(0.37)
SuGr	0.716	0.970	0.992	0.96 (3.61)	0.96 (3.73)	(-0.04)

Table D.3:
Datastream vs Compustat in the Common Sample - Panel B - Detailed

	CRSP Returns				Compustat Signals			
	Corr	CS	DS	CS - DS	Corr	CS	DS	CS - DS
Accruals								
Acc	0.999	0.55 (2.63)	0.54 (2.58)	(-0.96)	0.997	0.55 (2.65)	0.55 (2.65)	(0.05)
ChCE	0.998	1.16 (3.88)	1.13 (3.84)	(-2.07)	0.997	1.17 (3.88)	1.17 (3.90)	(-0.08)
ChCOA	0.994	0.68 (3.69)	0.68 (3.65)	(0.04)	0.994	0.68 (3.69)	0.66 (3.65)	(-1.41)
ChCOL	0.991	0.49 (2.87)	0.49 (2.84)	(0.10)	0.991	0.49 (2.87)	0.44 (2.70)	(-2.38)
ChFL	0.952	0.26 (2.98)	0.29 (3.35)	(0.95)	0.980	0.26 (2.91)	0.25 (2.90)	(-0.26)
ChLTI	0.658	0.24 (2.48)	0.17 (1.78)	(-1.18)	0.984	0.24 (2.47)	0.25 (2.59)	(0.97)
ChNCOA	0.989	0.93 (4.12)	0.95 (4.28)	(0.62)	0.996	0.93 (4.11)	0.91 (4.16)	(-0.96)
ChNCOL	0.879	0.33 (2.72)	0.27 (2.31)	(-1.08)	0.988	0.33 (2.73)	0.35 (2.88)	(1.09)
ChNFA	0.893	-0.05 (-0.47)	0.03 (0.31)	(1.88)	0.972	-0.04 (-0.43)	-0.06 (-0.54)	(-0.70)
ChNNCOA	0.991	0.86 (4.21)	0.90 (4.42)	(1.33)	0.996	0.86 (4.19)	0.86 (4.24)	(-0.21)
ChNNCWC	0.977	0.45 (3.52)	0.48 (3.67)	(1.19)	0.988	0.45 (3.55)	0.46 (3.62)	(0.37)
GriI	0.984	0.42 (2.87)	0.39 (2.75)	(-1.80)	0.994	0.43 (2.90)	0.42 (2.87)	(-0.82)
ICh	0.988	0.51 (3.15)	0.49 (3.07)	(-0.94)	0.994	0.51 (3.20)	0.51 (3.19)	(-0.20)
IGr	0.988	0.52 (3.04)	0.47 (2.82)	(-2.18)	0.989	0.52 (2.99)	0.49 (2.90)	(-0.97)
MBaAC	0.977	1.85 (4.87)	1.84 (4.74)	(-0.25)	0.986	1.85 (4.87)	1.77 (4.78)	(-1.69)
NWCCh	0.972	0.31 (3.65)	0.33 (3.81)	(1.07)	0.985	0.32 (3.67)	0.34 (3.62)	(1.00)
POA	0.981	0.42 (3.30)	0.46 (3.62)	(1.41)	0.993	0.42 (3.28)	0.44 (3.38)	(1.31)
PTA	0.981	0.52 (4.62)	0.50 (4.31)	(-1.04)	0.988	0.52 (4.64)	0.49 (4.56)	(-1.56)
TA	0.986	1.02 (3.55)	0.97 (3.50)	(-1.55)	0.996	1.02 (3.55)	1.02 (3.60)	(-0.19)
Intangibles								
AL	0.969	0.45 (1.56)	0.62 (1.60)	(1.22)	0.997	0.45 (1.58)	0.43 (1.58)	(-0.65)
ChGMChS	0.857	-0.15 (-1.59)	-0.16 (-1.77)	(-0.15)	0.991	-0.14 (-1.54)	-0.15 (-1.53)	(-0.25)
EPr	0.982	0.21 (1.12)	0.24 (1.19)	(0.57)	0.994	0.22 (1.16)	0.21 (1.08)	(-0.88)
ES	0.999	0.83 (2.64)	0.81 (2.63)	(-1.20)	1.000	0.81 (2.62)	0.82 (2.63)	(1.42)
HI	0.718	-0.13 (-0.69)	-0.10 (-0.63)	(0.30)	0.999	-0.14 (-0.74)	-0.16 (-0.82)	(-1.97)
HR	0.980	0.73 (3.63)	0.74 (3.86)	(0.16)	0.994	0.74 (3.72)	0.73 (3.69)	(-0.31)
IARER	0.680	0.75 (2.33)	0.45 (1.53)	(-1.43)	0.999	0.77 (2.37)	0.78 (2.41)	(0.37)
ICBE	0.630	-0.20 (-0.52)	-0.07 (-0.32)	(0.46)	1.000	-0.20 (-0.52)	-0.21 (-0.54)	(-0.86)
OC	0.989	0.62 (2.40)	0.58 (2.28)	(-1.03)	0.998	0.63 (2.42)	0.62 (2.42)	(-0.44)
RDM	0.998	1.87 (3.09)	1.87 (3.14)	(0.09)	0.999	1.86 (3.09)	1.85 (3.08)	(-0.47)
SmI	0.978	0.09 (0.91)	0.04 (0.45)	(-2.31)	0.998	0.09 (0.97)	0.10 (1.02)	(0.61)
TAN	0.999	0.40 (1.17)	0.38 (1.14)	(-0.95)	0.999	0.39 (1.18)	0.39 (1.16)	(-0.15)
URDI	0.992	0.54 (1.77)	0.62 (1.89)	(2.05)	1.000	0.54 (1.76)	0.53 (1.72)	(-1.71)
WWI	1.000	1.11 (3.12)	1.11 (3.15)	(-0.15)	0.999	1.11 (3.12)	1.13 (3.21)	(0.84)
Investment								
AGr	0.998	1.16 (3.84)	1.14 (3.82)	(-0.66)	0.996	1.15 (3.86)	1.14 (3.84)	(-0.74)
CAPEX	0.971	0.59 (4.14)	0.60 (4.44)	(0.28)	0.996	0.59 (4.22)	0.59 (4.29)	(-0.16)
CDI	0.971	-0.08 (-0.57)	-0.13 (-0.83)	(-1.29)	0.991	-0.12 (-0.80)	-0.10 (-0.74)	(0.64)
CEI5Y	0.995	0.19 (0.94)	0.16 (0.82)	(-1.33)	1.000	0.19 (0.97)	0.19 (0.96)	(-0.38)
ChNOA	0.964	0.32 (2.70)	0.34 (2.98)	(0.62)	0.991	0.32 (2.64)	0.32 (2.61)	(-0.01)
ChPPEIA	0.993	0.64 (3.52)	0.62 (3.53)	(-0.77)	0.997	0.64 (3.53)	0.63 (3.55)	(-0.51)
DI	0.983	0.18 (2.29)	0.18 (2.34)	(0.34)	0.998	0.18 (2.29)	0.18 (2.28)	(0.41)
GrLTNOA	0.996	0.77 (4.29)	0.78 (4.25)	(0.61)	0.999	0.77 (4.33)	0.78 (4.37)	(0.91)
INV	0.926	0.44 (4.02)	0.42 (3.87)	(-0.43)	0.985	0.43 (4.04)	0.41 (3.90)	(-1.23)
NDF	0.953	0.12 (1.35)	0.15 (2.02)	(0.87)	0.965	0.14 (1.60)	0.16 (1.88)	(1.20)
NEF	0.999	-0.05 (-0.13)	-0.06 (-0.17)	(-0.65)	1.000	-0.05 (-0.14)	-0.05 (-0.16)	(-1.15)
NOA	0.997	0.73 (2.58)	0.72 (2.56)	(-0.58)	0.998	0.72 (2.60)	0.73 (2.64)	(0.43)
NOACh	0.997	0.46 (2.32)	0.46 (2.28)	(-0.18)	0.998	0.47 (2.43)	0.47 (2.44)	(-0.54)
SR	0.994	-0.17 (-1.74)	-0.18 (-1.81)	(-0.64)	0.999	-0.17 (-1.74)	-0.18 (-1.81)	(-1.49)
TXFIN	0.995	0.22 (0.78)	0.18 (0.65)	(-1.55)	0.998	0.21 (0.77)	0.21 (0.78)	(0.21)
Profitability								
AT	0.997	0.20 (1.35)	0.20 (1.37)	(-0.13)	0.995	0.21 (1.45)	0.21 (1.46)	(-0.15)
CBOP	0.808	0.50 (1.75)	0.29 (1.13)	(-1.55)	0.997	0.50 (1.75)	0.50 (1.76)	(0.12)
CT	0.999	0.05 (0.22)	0.06 (0.24)	(0.46)	0.999	0.06 (0.24)	0.05 (0.19)	(-0.82)
ChiAT	0.979	0.10 (0.87)	0.15 (1.27)	(2.35)	0.975	0.12 (0.98)	0.10 (0.91)	(-0.92)
EC	0.953	0.03 (0.25)	0.02 (0.19)	(-0.24)	0.999	-0.01 (-0.05)	0.00 (0.00)	(1.99)
EP	0.994	0.51 (2.12)	0.50 (2.09)	(-0.63)	1.000	0.51 (2.10)	0.52 (2.14)	(2.21)
FSc	0.968	-0.36 (-1.29)	-0.32 (-1.16)	(0.66)	0.983	-0.35 (-1.29)	-0.40 (-1.45)	(-1.05)
GP	0.962	-0.02 (-0.10)	-0.03 (-0.12)	(-0.07)	0.997	-0.02 (-0.08)	-0.02 (-0.09)	(-0.13)
Lvrg	0.999	0.49 (1.29)	0.46 (1.21)	(-1.72)	0.999	0.49 (1.29)	0.51 (1.35)	(2.12)
OPtA	0.977	-0.03 (-0.10)	0.02 (0.06)	(1.02)	0.994	-0.03 (-0.10)	-0.08 (-0.25)	(-2.00)
OPtE	0.991	-0.33 (-0.78)	-0.31 (-0.75)	(0.28)	1.000	-0.33 (-0.77)	-0.35 (-0.81)	(-1.36)
OSc	0.998	-1.04 (-2.99)	-1.00 (-2.86)	(1.88)	0.998	-1.04 (-2.98)	-1.06 (-3.02)	(-1.36)
Value								
AM	0.999	1.28 (3.03)	1.27 (3.01)	(-1.03)	0.999	1.28 (3.01)	1.27 (3.06)	(-0.32)
BM	0.997	1.18 (3.46)	1.16 (3.36)	(-1.85)	0.998	1.18 (3.46)	1.16 (3.37)	(-1.80)
CM	0.999	-0.09 (-0.17)	-0.07 (-0.14)	(0.57)	1.000	-0.09 (-0.17)	-0.09 (-0.19)	(-0.64)
DurE	0.998	0.75 (2.25)	0.73 (2.20)	(-1.74)	0.999	0.74 (2.22)	0.76 (2.26)	(1.06)
ECoBP	0.999	0.80 (2.10)	0.76 (2.02)	(-1.99)	0.998	0.80 (2.09)	0.77 (2.04)	(-1.49)
EM	0.993	0.79 (2.64)	0.75 (2.48)	(-1.12)	0.998	0.79 (2.67)	0.79 (2.75)	(-0.19)
IR	0.997	1.12 (3.68)	1.11 (3.72)	(-0.15)	0.999	1.10 (3.66)	1.11 (3.67)	(0.74)
LCoBP	0.997	0.41 (1.36)	0.36 (1.27)	(-1.55)	0.998	0.41 (1.37)	0.39 (1.34)	(-1.07)
NPY	0.997	-0.01 (-0.03)	-0.01 (-0.03)	(0.03)	0.999	-0.02 (-0.05)	-0.03 (-0.11)	(-1.26)
OL	0.981	0.64 (3.06)	0.61 (2.85)	(-0.96)	0.983	0.64 (3.08)	0.64 (2.96)	(-0.34)
PY	0.988	-0.34 (-1.50)	-0.33 (-1.50)	(0.52)	0.998	-0.34 (-1.51)	-0.35 (-1.56)	(-0.41)
SP	1.000	1.20 (2.56)	1.18 (2.51)	(-2.48)	0.999	1.20 (2.55)	1.21 (2.59)	(0.49)
SaGr	0.996	0.26 (1.48)	0.26 (1.52)	(0.45)	0.997	0.26 (1.47)	0.27 (1.56)	(1.08)
SuGr	0.995	0.96 (3.61)	0.98 (3.72)	(0.72)	0.997	0.96 (3.62)	0.94 (3.57)	(-1.64)

Table D.4:
Datastream vs Compustat in Their Own Full Samples - Detailed

	Full Samples				Cap Over $100 million & No Financial & 2001+			
	Corr	CS	DS	CS - DS	Corr	CS	DS	CS - DS
Accruals								
Acc	0.953	0.31 (1.39)	0.58 (2.56)	(2.65)	0.956	-0.05 (-0.34)	0.04 (0.27)	(2.82)
ChCE	0.896	0.64 (2.27)	1.40 (4.16)	(3.87)	0.910	0.11 (0.53)	0.26 (1.16)	(2.13)
ChCOA	0.692	0.64 (3.28)	0.73 (3.77)	(0.59)	0.933	0.09 (0.40)	0.04 (0.18)	(-0.85)
ChCOL	0.710	0.55 (3.49)	0.43 (2.07)	(-0.82)	0.946	0.18 (0.90)	0.21 (0.95)	(0.49)
ChFL	0.612	0.50 (5.46)	0.29 (3.41)	(-3.24)	0.909	0.04 (0.28)	0.09 (0.80)	(0.60)
ChLTI	0.400	0.21 (2.71)	0.19 (1.83)	(-0.16)	0.597	0.21 (1.54)	0.26 (1.28)	(0.42)
ChNCOA	0.876	0.82 (3.55)	1.12 (4.70)	(2.11)	0.911	0.24 (1.13)	0.31 (1.61)	(1.49)
ChNCOL	0.802	0.20 (1.74)	0.30 (2.73)	(1.24)	0.874	-0.02 (-0.08)	0.04 (0.23)	(0.56)
ChNFA	0.503	0.29 (2.72)	-0.04 (-0.39)	(-3.04)	0.667	0.01 (0.05)	0.02 (0.17)	(0.15)
ChNNCOA	0.870	0.84 (3.86)	1.05 (4.87)	(1.42)	0.891	0.31 (1.54)	0.36 (2.46)	(0.71)
ChNNCWC	0.777	0.29 (1.97)	0.44 (3.10)	(2.00)	0.867	-0.05 (-0.43)	-0.09 (-0.99)	(-0.65)
GriI	0.771	0.49 (3.26)	0.42 (3.19)	(-0.70)	0.945	0.20 (0.84)	0.07 (0.36)	(-2.27)
ICh	0.760	0.52 (3.09)	0.54 (3.63)	(0.16)	0.913	0.28 (1.12)	0.17 (0.73)	(-2.05)
IGr	0.776	0.50 (2.72)	0.62 (3.85)	(1.06)	0.921	0.02 (0.10)	0.04 (0.18)	(0.31)
MBaAC	0.842	1.67 (4.96)	1.47 (3.69)	(-1.13)	0.939	-0.02 (-0.04)	-0.23 (-0.41)	(-1.98)
NWCCh	0.778	0.43 (4.44)	0.30 (3.16)	(-1.98)	0.841	-0.01 (-0.05)	-0.01 (-0.07)	(-0.02)
POA	0.825	0.68 (5.18)	0.26 (1.96)	(-4.57)	0.941	0.15 (0.65)	0.08 (0.46)	(-0.85)
PTA	0.783	0.35 (3.30)	0.61 (4.06)	(2.35)	0.861	0.13 (1.20)	0.13 (1.21)	(0.03)
TA	0.851	0.64 (2.67)	1.07 (3.87)	(2.61)	0.785	0.28 (1.61)	0.31 (1.86)	(0.38)
Intangibles								
AL	0.927	0.38 (1.35)	0.73 (1.88)	(1.75)	0.808	0.36 (1.02)	0.09 (0.27)	(-2.82)
ChGMChS	0.482	0.16 (1.57)	-0.19 (-1.59)	(-2.90)	0.841	0.01 (0.03)	-0.05 (-0.31)	(-0.51)
EPr	0.835	0.72 (3.82)	0.14 (0.70)	(-4.20)	0.971	0.43 (0.99)	0.34 (0.94)	(-0.91)
ES	0.904	0.07 (0.20)	0.81 (2.83)	(3.67)	0.982	0.06 (0.22)	0.12 (0.43)	(1.08)
HI	0.702	0.00 (0.03)	-0.19 (-0.62)	(-0.88)	0.890	0.26 (1.34)	0.23 (0.82)	(-0.27)
HR	0.821	0.59 (3.09)	0.99 (4.50)	(2.68)	0.942	0.39 (1.69)	0.36 (1.73)	(-0.45)
IARER	0.148	0.31 (2.11)	0.56 (2.41)	(0.89)	0.084	-0.17 (-0.49)	-0.14 (-0.81)	(0.08)
ICBE	0.087	-0.03 (-0.14)	0.07 (0.31)	(0.34)	0.682	0.13 (0.72)	0.25 (1.12)	(0.72)
OC	0.913	0.46 (1.66)	0.97 (3.05)	(3.04)	0.930	0.32 (0.95)	0.30 (0.90)	(-0.23)
RDM	0.978	1.37 (2.31)	1.99 (3.11)	(4.05)	0.970	0.54 (1.61)	0.47 (1.29)	(-0.79)
SmI	0.345	0.33 (2.78)	0.07 (0.74)	(-2.05)	0.805	-0.36 (-4.41)	-0.36 (-4.65)	(0.08)
TAN	0.972	0.45 (1.34)	0.37 (1.03)	(-0.73)	0.981	-0.33 (-1.63)	-0.20 (-1.16)	(2.73)
URDI	0.944	0.52 (1.56)	0.48 (1.45)	(-0.39)	0.955	0.22 (0.90)	0.22 (0.90)	(-0.06)
WWI	0.929	0.44 (1.19)	1.26 (3.59)	(5.49)	0.986	-0.15 (-0.83)	-0.02 (-0.11)	(2.62)
Investment								
AGr	0.880	0.96 (3.01)	1.42 (4.30)	(2.13)	0.951	0.25 (0.97)	0.25 (0.86)	(-0.02)
CAPEX	0.660	0.41 (2.77)	0.60 (4.83)	(1.78)	0.890	0.18 (1.12)	0.15 (0.96)	(-0.60)
CDI	0.822	0.17 (1.05)	-0.06 (-0.37)	(-2.40)	0.893	0.18 (1.06)	0.11 (0.78)	(-0.77)
CEI5Y	0.984	0.41 (2.10)	0.22 (1.13)	(-3.84)	0.978	0.42 (1.73)	0.48 (2.04)	(0.89)
ChNOA	0.586	0.28 (2.66)	0.51 (3.99)	(2.11)	0.838	-0.11 (-0.63)	-0.12 (-0.83)	(-0.32)
ChPPEIA	0.849	0.65 (3.38)	0.81 (4.38)	(1.29)	0.903	0.41 (2.22)	0.38 (2.46)	(-0.44)
DI	0.782	0.35 (3.53)	0.25 (2.25)	(-1.23)	0.955	-0.02 (-0.13)	-0.02 (-0.24)	(-0.20)
GrLTNOA	0.930	0.42 (2.56)	0.86 (4.61)	(6.77)	0.959	0.22 (1.15)	0.33 (2.14)	(2.00)
INV	0.790	0.21 (1.77)	0.45 (3.91)	(3.51)	0.875	0.04 (0.27)	0.10 (0.73)	(0.82)
NDF	0.553	0.42 (4.90)	0.30 (3.34)	(-1.71)	0.630	0.02 (0.12)	0.15 (1.33)	(0.93)
NEF	0.968	0.68 (2.03)	0.01 (0.02)	(-6.08)	0.990	0.33 (1.31)	0.28 (1.05)	(-0.97)
NOA	0.709	0.70 (3.19)	0.91 (3.22)	(1.07)	0.950	0.19 (1.07)	0.23 (1.51)	(0.63)
NOACh	0.936	0.55 (2.59)	0.59 (3.29)	(0.69)	0.936	0.38 (2.48)	0.43 (2.73)	(0.83)
SR	0.892	0.13 (1.16)	-0.27 (-1.76)	(-4.97)	0.976	0.31 (2.51)	0.23 (2.01)	(-2.90)
TXFIN	0.884	1.04 (3.86)	0.29 (1.04)	(-6.44)	0.949	0.54 (1.73)	0.32 (1.08)	(-1.68)
Profitability								
AT	0.946	0.16 (1.12)	0.30 (1.95)	(2.20)	0.967	0.39 (2.03)	0.55 (2.75)	(1.80)
CBOP	0.815	0.80 (2.56)	0.43 (1.61)	(-2.65)	0.826	0.42 (0.83)	0.60 (1.67)	(0.74)
CT	0.726	0.09 (0.42)	0.01 (0.04)	(-0.29)	0.975	0.52 (2.48)	0.52 (2.81)	(0.10)
ChiAT	0.743	0.18 (1.86)	0.12 (1.14)	(-0.96)	0.890	-0.19 (-1.64)	-0.21 (-1.26)	(-0.25)
EC	0.725	0.09 (0.85)	0.11 (1.03)	(0.33)	0.839	-0.04 (-0.41)	0.14 (1.42)	(2.37)
EP	0.914	0.56 (2.63)	0.64 (2.84)	(1.16)	0.950	-0.01 (-0.07)	0.10 (0.51)	(1.33)
FSc	0.749	0.29 (1.03)	-0.71 (-1.83)	(-4.14)	0.739	0.22 (0.59)	0.43 (1.07)	(1.16)
GP	0.835	0.28 (1.17)	-0.01 (-0.07)	(-2.05)	0.923	0.37 (1.32)	0.39 (1.33)	(0.13)
Lvrg	0.967	0.25 (0.72)	0.55 (1.44)	(3.00)	0.973	0.44 (1.28)	0.46 (1.36)	(0.35)
OPtA	0.772	0.93 (2.69)	-0.10 (-0.29)	(-5.84)	0.890	0.51 (1.33)	0.44 (1.56)	(-0.47)
OPtE	0.954	0.44 (1.08)	-0.38 (-0.83)	(-4.75)	0.974	0.39 (1.62)	0.38 (1.88)	(-0.11)
OSc	0.942	0.28 (0.75)	-1.05 (-2.83)	(-8.17)	0.982	0.10 (0.26)	-0.09 (-0.25)	(-2.55)
Value								
AM	0.946	1.09 (2.86)	1.20 (2.89)	(0.71)	0.988	0.43 (0.94)	0.46 (1.03)	(0.62)
BM	0.932	1.20 (3.79)	1.19 (3.37)	(-0.10)	0.978	0.14 (0.40)	0.13 (0.38)	(-0.32)
CM	0.961	0.71 (1.53)	-0.26 (-0.51)	(-6.79)	0.983	0.58 (2.34)	0.44 (1.88)	(-2.54)
DurE	0.914	0.90 (2.65)	0.81 (2.35)	(-0.60)	0.979	0.04 (0.11)	0.07 (0.19)	(0.55)
ECoBP	0.951	0.82 (2.22)	0.52 (1.36)	(-2.29)	0.985	0.16 (0.41)	0.20 (0.50)	(0.98)
EM	0.942	-0.05 (-0.15)	0.76 (2.45)	(5.19)	0.946	-0.31 (-1.44)	-0.09 (-0.45)	(3.15)
IR	0.902	0.58 (1.85)	1.22 (3.83)	(4.16)	0.959	-0.00 (-0.01)	0.08 (0.27)	(0.90)
LCoBP	0.949	0.39 (1.48)	0.22 (0.82)	(-1.78)	0.977	-0.20 (-1.09)	-0.23 (-1.27)	(-0.41)
NPY	0.950	0.87 (2.81)	-0.03 (-0.08)	(-7.80)	0.980	0.49 (1.87)	0.37 (1.32)	(-1.42)
OL	0.761	0.49 (2.73)	0.78 (3.92)	(2.36)	0.908	0.47 (1.99)	0.48 (2.39)	(0.19)
PY	0.864	0.25 (1.16)	-0.40 (-1.42)	(-4.23)	0.934	0.14 (0.59)	0.23 (0.99)	(1.59)
SP	0.973	1.13 (2.55)	1.39 (2.92)	(1.91)	0.987	0.42 (1.15)	0.53 (1.47)	(1.90)
SaGr	0.682	0.21 (1.45)	0.36 (1.91)	(1.05)	0.939	0.18 (1.00)	0.24 (1.30)	(1.09)
SuGr	0.863	0.40 (1.68)	1.04 (3.94)	(4.37)	0.891	0.12 (0.64)	0.23 (1.19)	(1.19)

Table D.5:
Portfolio Constructions Reducing Discrepancy Between Databases - Detailed

	All-but-microcaps VW				Breakpoints from 1000 Largest Stocks VW			
	Corr	CS	DS	CS - DS	Corr	CS	DS	CS - DS
Accruals								
Acc	0.940	0.09 (0.96)	0.17 (1.54)	(2.47)	0.953	0.07 (0.56)	0.16 (1.20)	(3.51)
ChCE	0.957	0.21 (0.94)	0.29 (1.26)	(1.57)	0.963	0.27 (1.27)	0.31 (1.39)	(0.82)
ChCOA	0.928	0.12 (0.63)	0.15 (0.66)	(0.42)	0.953	0.10 (0.56)	0.16 (0.75)	(0.96)
ChCOL	0.958	-0.02 (-0.11)	-0.00 (-0.02)	(0.29)	0.973	0.01 (0.03)	0.01 (0.06)	(0.11)
ChFL	0.760	0.29 (2.68)	0.16 (1.73)	(-1.55)	0.809	0.30 (2.72)	0.19 (1.90)	(-1.64)
ChLTI	0.678	-0.02 (-0.17)	-0.04 (-0.23)	(-0.19)	0.731	-0.03 (-0.19)	-0.06 (-0.33)	(-0.40)
ChNCOA	0.819	0.32 (2.25)	0.34 (2.07)	(0.36)	0.748	0.27 (1.89)	0.33 (2.03)	(0.72)
ChNCOL	0.733	-0.08 (-0.56)	-0.07 (-0.60)	(0.05)	0.765	-0.00 (-0.03)	-0.04 (-0.35)	(-0.46)
ChNFA	0.820	0.27 (1.55)	0.20 (0.98)	(-0.76)	0.844	0.25 (1.50)	0.16 (0.79)	(-1.01)
ChNNCOA	0.829	0.42 (3.07)	0.34 (2.23)	(-1.10)	0.789	0.33 (2.52)	0.35 (2.27)	(0.25)
ChNNCWC	0.911	0.29 (1.78)	0.33 (2.08)	(0.52)	0.884	0.16 (1.11)	0.26 (1.86)	(1.64)
GriI	0.896	0.38 (2.52)	0.22 (1.55)	(-2.45)	0.855	0.36 (2.46)	0.23 (1.69)	(-1.56)
ICh	0.828	0.41 (2.50)	0.26 (1.62)	(-1.66)	0.870	0.44 (2.76)	0.32 (2.13)	(-1.84)
IGr	0.911	0.06 (0.32)	0.05 (0.29)	(-0.12)	0.926	0.07 (0.45)	0.08 (0.49)	(0.12)
MBaAC	0.805	0.75 (1.46)	0.72 (1.28)	(-0.07)	0.829	0.90 (1.85)	0.89 (1.60)	(-0.02)
NWCCh	0.895	0.24 (1.71)	0.23 (1.65)	(-0.17)	0.898	0.14 (1.04)	0.20 (1.49)	(0.97)
POA	0.914	0.24 (2.05)	0.36 (2.96)	(1.77)	0.921	0.18 (1.21)	0.27 (2.23)	(1.62)
PTA	0.888	0.16 (0.93)	0.21 (1.18)	(0.60)	0.914	0.09 (0.55)	0.17 (0.98)	(1.38)
TA	0.825	0.17 (0.92)	0.23 (1.11)	(0.66)	0.876	0.15 (0.81)	0.19 (0.90)	(0.44)
Intangibles								
AL	0.664	0.10 (0.56)	0.48 (2.82)	(2.58)	0.766	0.24 (1.35)	0.32 (1.85)	(0.67)
ChGMChS	0.670	0.02 (0.20)	0.04 (0.25)	(0.10)	0.640	0.06 (0.73)	-0.05 (-0.38)	(-0.96)
EPr	0.940	0.42 (2.04)	0.30 (1.63)	(-1.40)	0.931	0.44 (2.20)	0.30 (1.62)	(-1.67)
ES	0.955	0.05 (0.19)	0.26 (1.04)	(2.00)	0.961	0.13 (0.67)	0.25 (1.05)	(1.32)
HI	0.528	-0.05 (-0.40)	0.09 (0.55)	(0.89)	0.558	0.07 (0.46)	0.08 (0.42)	(0.01)
HR	0.959	0.04 (0.17)	-0.01 (-0.05)	(-0.77)	0.965	-0.05 (-0.22)	-0.07 (-0.29)	(-0.38)
IARER	0.067	0.31 (1.34)	0.44 (1.20)	(0.34)	0.079	0.32 (1.49)	0.13 (0.34)	(-0.34)
ICBE	0.537	0.14 (0.94)	0.13 (0.92)	(-0.08)	0.525	0.27 (1.26)	0.20 (1.32)	(-0.36)
OC	0.916	0.28 (1.55)	0.27 (1.38)	(-0.23)	0.931	0.28 (1.53)	0.28 (1.46)	(0.11)
RDM	0.946	0.54 (1.93)	0.71 (2.48)	(1.89)	0.967	0.52 (2.21)	0.69 (2.59)	(2.54)
SmI	0.713	0.13 (0.92)	0.12 (0.72)	(-0.07)	0.764	0.12 (0.85)	0.11 (0.63)	(-0.10)
TAN	0.942	0.02 (0.17)	0.04 (0.23)	(0.19)	0.957	0.06 (0.44)	0.11 (0.72)	(1.07)
URDI	0.865	0.34 (1.76)	0.45 (1.71)	(0.90)	0.868	0.34 (1.76)	0.45 (1.72)	(0.93)
WWI	0.968	0.02 (0.07)	0.35 (0.95)	(2.69)	0.980	0.13 (0.48)	0.37 (1.27)	(3.06)
Investment								
AGr	0.960	0.31 (1.43)	0.32 (1.39)	(0.15)	0.965	0.25 (1.08)	0.27 (1.16)	(0.53)
CAPEX	0.731	0.22 (1.29)	0.11 (0.61)	(-0.93)	0.786	0.13 (0.81)	0.07 (0.46)	(-0.60)
CDI	0.918	0.08 (0.51)	-0.05 (-0.31)	(-1.44)	0.898	0.08 (0.56)	-0.04 (-0.24)	(-1.12)
CEI5Y	0.967	0.28 (1.76)	0.32 (1.83)	(0.68)	0.963	0.23 (1.44)	0.21 (1.23)	(-0.34)
ChNOA	0.825	0.28 (2.12)	0.40 (2.41)	(1.00)	0.660	0.32 (2.45)	0.24 (2.16)	(-0.63)
ChPPEIA	0.909	0.25 (1.60)	0.27 (1.87)	(0.26)	0.920	0.26 (1.80)	0.27 (1.86)	(0.03)
DI	0.884	0.26 (2.59)	0.21 (1.52)	(-0.78)	0.887	0.26 (2.61)	0.22 (1.59)	(-0.66)
GrLTNOA	0.809	0.19 (1.84)	0.20 (2.15)	(0.11)	0.809	0.21 (1.68)	0.20 (1.88)	(-0.21)
INV	0.886	0.17 (1.29)	0.15 (0.94)	(-0.45)	0.914	0.18 (1.72)	0.16 (1.29)	(-0.32)
NDF	0.726	0.18 (1.70)	0.20 (2.23)	(0.23)	0.787	0.16 (1.29)	0.19 (2.01)	(0.43)
NEF	0.950	0.22 (0.77)	0.05 (0.17)	(-2.21)	0.962	0.18 (0.71)	0.13 (0.48)	(-0.77)
NOA	0.614	0.45 (3.56)	0.44 (2.83)	(-0.09)	0.598	0.41 (3.33)	0.46 (3.51)	(0.35)
NOACh	0.816	0.40 (2.58)	0.56 (3.73)	(1.79)	0.865	0.38 (3.21)	0.46 (3.75)	(1.53)
SR	0.914	0.03 (0.24)	0.05 (0.32)	(0.33)	0.918	0.03 (0.26)	0.04 (0.24)	(0.11)
TXFIN	0.862	0.54 (2.17)	0.31 (1.25)	(-2.39)	0.878	0.26 (1.30)	0.22 (1.04)	(-0.50)
Profitability								
AT	0.981	0.21 (1.30)	0.31 (1.65)	(1.80)	0.975	0.25 (1.70)	0.34 (1.92)	(1.45)
CBOP	0.778	0.64 (2.54)	0.49 (2.05)	(-1.13)	0.801	0.57 (2.56)	0.48 (2.27)	(-0.88)
CT	0.794	0.06 (0.29)	0.16 (1.01)	(0.73)	0.830	0.05 (0.23)	0.16 (1.03)	(0.92)
ChiAT	0.860	0.24 (1.49)	0.22 (1.47)	(-0.28)	0.895	0.15 (1.01)	0.11 (0.75)	(-0.84)
EC	0.905	0.24 (1.88)	0.12 (1.08)	(-2.20)	0.925	0.21 (1.60)	0.15 (1.18)	(-1.17)
EP	0.965	0.32 (1.15)	0.36 (1.29)	(0.76)	0.978	0.37 (1.30)	0.37 (1.36)	(0.03)
FSc	0.620	-0.02 (-0.07)	-0.30 (-0.81)	(-0.83)	0.635	0.09 (0.34)	-0.39 (-1.07)	(-1.52)
GP	0.928	0.13 (0.64)	0.22 (0.96)	(1.15)	0.927	0.14 (0.71)	0.15 (0.66)	(0.05)
Lvrg	0.983	0.03 (0.07)	0.01 (0.03)	(-0.24)	0.984	0.10 (0.29)	0.08 (0.26)	(-0.32)
OPtA	0.820	0.48 (1.85)	0.41 (1.68)	(-0.56)	0.790	0.43 (1.82)	0.28 (1.41)	(-1.33)
OPtE	0.864	0.39 (1.36)	0.12 (0.36)	(-1.67)	0.829	0.31 (1.44)	0.24 (0.84)	(-0.43)
OSc	0.931	0.11 (0.44)	-0.10 (-0.41)	(-2.59)	0.935	0.01 (0.04)	-0.16 (-0.83)	(-2.36)
Value								
AM	0.986	0.15 (0.50)	0.19 (0.63)	(1.00)	0.985	0.15 (0.49)	0.22 (0.71)	(2.00)
BM	0.970	0.20 (0.73)	0.29 (1.06)	(1.66)	0.975	0.12 (0.44)	0.19 (0.73)	(1.57)
CM	0.968	0.39 (1.12)	0.30 (0.78)	(-1.06)	0.976	0.38 (1.17)	0.24 (0.71)	(-2.10)
DurE	0.903	0.23 (0.78)	0.13 (0.48)	(-0.97)	0.918	0.21 (0.73)	0.24 (0.83)	(0.33)
ECoBP	0.969	0.11 (0.34)	0.14 (0.41)	(0.41)	0.984	0.08 (0.24)	0.10 (0.30)	(0.44)
EM	0.923	0.10 (0.44)	0.27 (1.16)	(1.52)	0.948	0.20 (1.00)	0.23 (1.30)	(0.61)
IR	0.965	0.15 (0.54)	0.16 (0.59)	(0.10)	0.967	0.17 (0.60)	0.25 (0.94)	(1.78)
LCoBP	0.964	0.27 (1.02)	0.46 (1.73)	(2.00)	0.983	0.25 (0.96)	0.36 (1.26)	(1.61)
NPY	0.950	0.37 (1.11)	0.23 (0.77)	(-1.84)	0.938	0.17 (0.61)	0.12 (0.40)	(-0.51)
OL	0.904	0.21 (1.07)	0.27 (1.47)	(0.93)	0.860	0.23 (1.36)	0.29 (1.90)	(0.77)
PY	0.900	-0.01 (-0.03)	0.06 (0.15)	(0.42)	0.751	0.13 (0.41)	-0.09 (-0.29)	(-1.24)
SP	0.973	0.26 (0.81)	0.42 (1.28)	(3.39)	0.985	0.28 (0.90)	0.40 (1.19)	(2.23)
SaGr	0.960	0.09 (0.40)	0.02 (0.08)	(-0.95)	0.961	0.06 (0.27)	0.01 (0.04)	(-0.73)
SuGr	0.938	0.16 (0.90)	0.23 (1.15)	(0.92)	0.952	0.20 (1.04)	0.24 (1.18)	(0.81)

Chapter 2

Omitted Strategy Bias in Anomalies Research

In this paper, we study the statistical significance of 93 fundamental anomalies published in academic journals in a multiple hypothesis setting. Harvey et al. (2016) documented the importance of adopting an appropriate approach to testing when considering many possibly significant signals. For every published anomaly there are potentially many others that were tried but not published. If one considers 20 possible signals sequentially in a single hypothesis test of their significance, she would on average find one significant signal even if, in reality, none of them is significant.[1] The multiple hypothesis testing framework then corrects for this error rate in specific and controlled ways. One problem is that all the explored signals cannot be observed as the insignificant findings were not published; it is thus impossible to account for them. Harvey et al. (2016) attempted to overcome the issue by making strong structural assumptions and simulating a hypothetical sample of t-statistics on these unpublished signals. However, this can also be very problematic as the results depend on assumptions that cannot be fully tested. We take a different approach by revisiting the data mining approach to fundamental signals of Yan and Zheng (2017) in order to generate the universe of potential strategies. The potential strategies can then be studied directly with well-established testing methods.

The fundamental anomalies in this study describe characteristics related to individual stocks that can predict their future returns. No distinction is being made between characteristics that are related to risk premia and variables that are related to mispricing due to frictions or other market imperfections. The studied anomalies are, for example, accruals of Sloan (1996), earnings over price of Basu (1977), composite equity issuance of Daniel and Titman (2006), and R&D over Market Equity of Chan et al. (2001).

The analysis considers 48,387 data-mined fundamental signals in an international setting generated from a mixture of items on balance sheet, income, and cash flow statements. The main body of the analysis focuses only on 1,497 data-mined signals that are the clos-

[1]See Lo and MacKinlay (1990) and MacKinlay (1995) for early warning that data snooping can become a serious problem in empirical finance.

est to the published anomalies. The generated fundamental signals are very close to 25 of the published anomalies and loosely to another 6; this should provide a realistic setting for the universe of potential strategies. The 25 anomalies can be considered a subset of the universe of data-mined signals and the data-mined signals are therefore a good approximation of the universe of possibly tried strategies the 25 anomalies were selected from. Construction of some of the 93 anomalies is, however, not close to construction of any of the data-mined signals. The data-mined signals are nevertheless closely correlated with the anomalies regardless of the differences in construction. There are 34 anomalies whose equal-weighted decile long-short portfolio returns in the US have at least 90% correlation with one of the 1,497 data-mined signals. This number goes up to 61 for a minimum 75% correlation and up to 88 for a minimum 50% correlation. The data-mined signals are therefore a good approximation of the universe of tried signals from which the anomalies were selected.[2]

All portfolios and predictions in our analysis are updated at an annual frequency. The fundamental signals are dominant drivers of returns at the annual frequency and our universe of strategies should, therefore, be an ideal testing ground for the selection of significant annual signals. Omitted published anomalies, that are dominant at higher frequencies, should only have limited influence on our analysis.

We apply the formal multiple hypothesis test of Storey (2002) to deal with the possible false positive signals.[3] The test controls for the false discovery rate of signals at a 5% level. We focus on a universe of stocks with price over $1 and size larger than the bottom decile of the New York Stock Exchange (NYSE) stocks since any annual signal could be profitably traded there. This is motivated by Hou et al. (2017) and Green et al. (2017) who show that micro-caps account for only 3% of overall capitalization of stock market in the US but can have large impact on the number of discovered anomalies. The analysis covers stocks in the majority of developed markets, which are grouped into following regions: Asia Pacific, Europe, Japan, and the US. The number of significant signals detected is highly dependent on the precise specification of the tests. There are fewer significant signals for value-weighted returns and for factor models with a larger number of factors. Using the Fama and French (2015) five factor model (FF5) decreases the number of significant signals in comparison with Capital Asset Pricing Model (CAPM). The findings of Harvey et al. (2016) where the test statistics are taken from different studies with various methodologies can therefore suffer from large biases.

The number of significant signals varies greatly across the regions. There are notably

[2]The main analysis in this study is always conducted on both the full set of 93 anomalies and the reduced set of 25 anomalies to provide a robustness check for the assumption.

[3]Storey (2002) was first introduced in finance context in Barras et al. (2010) to test performance of mutual funds and in Bajgrowicz and Scaillet (2012) to test performance of technical trading rules. Further papers dealing with performance of mutual funds in multiple hypothesis framework are Kosowski et al. (2006) and Kosowski et al. (2007). For early finance literature that attempts to correct for data mining biases, see Sullivan et al. (1999), Sullivan et al. (2001), and White (2000). Foster et al. (1997), Cooper and Gulen (2006), Green et al. (2017) discuss data mining during variable selection in regression setting.

no significant signals at all found in Japan. The number of significant signals is lower in Europe and Asia Pacific compared to the US, especially for value-weighted returns. The critical values of t-statistics for 5% significance level, after accounting for multiple hypothesis setting, are higher than standard value of 1.96 in single hypothesis tests but they are generally lower than 3 as suggested in Harvey et al. (2016). Equal-weighted returns require lower cut-off of about 2 to 2.5 since there are many more significant signals in this case. The critical value for value-weighted return is close to 3. The critical value also generally increases with number of risk factors for which we adjust the returns. This is in line with findings of Fama and French (2017) that their five factor model dissects more anomalies that their three factor model. The results are similar for the full set of 93 anomalies and the reduced set of 25 anomalies which implies that the data-mined signals are a good approximation for the universe of signals for the 93 anomalies.

We show that the number of significant fundamental anomalies strongly depends on (a) the adjustment of standard errors on portfolio returns for Heteroskedasticity and Autocorrelation (HAC) and (b) method to obtain p-values from t-statistics. Most authors apply the Newey-West (1987) adjustment which requires specification of lags.[4] We show that number of significant signals can drop to one half depending on the specification of the lag length in HAC adjustment. There is no prior evidence on this issue, to the best of our knowledge, and different authors choose the number of lags apparently completely arbitrarily. The frequent choice is fewer than 6 lags.[5] The framework with many signals is an ideal testing ground to see the impact of this choice. Another problem with HAC robust standard errors is that they tend to understate confidence intervals and reject too many signals. The over-rejection rate is a well-documented phenomenon in the testing literature and there are now many remedies available.[6] We tackle the over-rejection problem by relying on the "naive" block bootstrap of Goncalves and Vogelsang (2011) with a block length of 3 or 12. Bootstrapping p-values leads to fewer significant signals compared with the standard approach which implies that relying on quantiles of the normal distribution for critical values of the t-statistics can be very misleading. None of the reviewed anomalies studies uses the bootstrap or non-normal critical values and the p-values reported there are therefore inflated.

We propose a new simulation approach to study power and size of the significance tests in a controlled environment. We randomly generate fundamental signals and create portfolios based on them. Returns on the portfolios inherit properties of the data-mined signals while having zero expected returns by definition which allows us to study how the number of significant signals changes with varying expected returns. The simulation

[4]Note that even estimators with automatic selection of lags, such as Newey and West (1994) and Andrews and Monahan (1992), suffer from similar problems as the procedures tend to select standard errors that understate confidence intervals.

[5]The issue is so neglected that most of the authors do not even mention any adjustment. See, for example, Fama and French (2015), Fama and French (2016), Fama and French (2017), and Ang et al. (2006b). Eisfeldt and Papanikolaou (2013) adjust for one lag and Ang et al. (2009) for four lags.

[6]See, for example, Andrews and Monahan (1992), Newey and West (1994), Kiefer et al. (2000), Kiefer and Vogelsang (2005), and many others.

exercise reveals that the size of individual tests can be heavily distorted for small numbers of lags in the HAC adjustment on annually rebalanced portfolio returns. The distortion is the largest for equal-weighted annually rebalanced portfolios where the bootstrapped empirical size of the tests is almost double the intended size. Multiple hypothesis tests inherit false discovery rate distortions due to their dependence on tests of individual signals. We next examine power of the tests depending on expected annual returns of the signals. Equal-weighted portfolios lead to much larger power relative to value-weighted portfolios which explains why many of the anomalies disappear for value-weighting. Number of risk factors used to adjust the raw returns also plays a large role in the tests' power. The power of the tests decreases with the number of risk factors and FF5 leads to tests with the smallest power. This implies that some of the anomalies in Fama and French (2017) could have been explained because of the poor power of the tests and not because of the higher explanatory power of FF5 model.

The number of significant fundamental signals increases proportionally with the number of data-mined signals in the multiple hypothesis tests applied to long-short portfolios based on the signals. There are therefore about 20,000 statistically significant strategies on the extended universe of data-mined signals. It is, however, hard to believe that there are so many profitable independent annual signals as it would point to severely inefficient markets. The analysis so far has disregarded correlation structure between the signals. Discarding closely correlated signals heavily reduces the number of data-mined signals and leads to a decrease in proportion of significant signals when more data-mined signals are added. The portfolio setting therefore offers only limited insight when it comes to the number of independently significant signals and it is possible that there are only few truly significant signals that are then mirrored in the other significant signals.

We next examine the impact of missing unpublished signals for the selection of independently significant signals in regression setting of Lewellen et al. (2015) and Green et al. (2017). That is, we try to find signals that significantly predict returns on individual stocks. Green et al. (2017) found that there are only 12 such signals in the US on their set of published anomalies. We document that the omission of tried but unpublished signals leads to the same biases as at the portfolio level analysis and the standard multiple hypothesis methods are not conservative enough. The number of independently significant signals does not increase with larger number of data-mined signals, as was the case in the portfolio setting. Most of the data-mined signals in the portfolio setting were therefore closely related to few common risk premiums.

We then select signals with least absolute shrinkage and selection operator (LASSO) that are both economically (size of coefficients) and statistically significant. The selected published anomalies are very similar to the selected data-mined signals which documents that our data mining process leads to similar selection procedure as publishing process for the academic research.

We next study out-of-sample performance of the data mined signals versus published anomalies. Anomalies have to undergo a vetting procedure by referees in order to get published and in principle this should in turn lead to better performance out-of-sample. We demonstrate that this is indeed the case for a simple strategy that equally invests in historically significant signals. In particular, the anomalies identified in the US are profitable in all the other regions under study while data mined signals are only profitable to a smaller extent. The academic publishing process is therefore able to identify important risk factors that are valid everywhere. We next try to create an optimal combination of predictive fundamental signals with LASSO regressions of individual stock returns on the fundamental signals. LASSO leads to a significant improvement in the out-of-sample performance of data mining to the point that it is not significantly different from published anomalies. The more advanced methods of supervised machine learning and data mining can therefore lead to comparable predictive capability as academic research.

Our study is the most closely related in methodology to Yan and Zheng (2017) and Chordia et al. (2017) but there are some stark contrasts. Yan and Zheng (2017) focused on full universe of stocks including micro-caps but failed to account for multiple hypothesis setting in the choice of individually significant signals. Yan and Zheng (2017) used the generated signals as a "fishing license" to introduce hundreds of new signals.[7] Chordia et al. (2017) then introduced proper multiple hypothesis tests on their 2.1 million signals. The enormous number of signals led them to conclude that it is not possible to select economically meaningful new signals and critical values for t-statistics stop playing any role in their setting. We show that these conclusions are mainly caused by the use of methods that are unfit for the purpose. The number of significant signals increases uncontrollably with the number of signals only if the correlation structure between the signals is disregarded. Controlling for the correlation structure and focusing on independently significant signals leads to sensible critical values for t-statistics and number of discoveries of the significant signals.

In terms of substance, our paper is the closest to Harvey et al. (2016) who applied the multiple hypothesis framework to the findings of many journal articles. The analysis in Harvey et al. (2016) is limited by the fact that it is not based on panel data of returns and is rather relying on a simulation framework with strong assumptions. Our analysis overcomes these difficulties by generating an universe of potential strategies. There is now a large literature on the choice of independently significant signals.[8] This paper touches the topic of selection of independently significant signals but it mainly focuses on the impact of considering the full universe of tried signals rather than on which factors are significant per se. The use of formal statistical methods to isolate a predictive fundamental signal is similar to Bartram and Grinblatt (2018a) but the methodology on how to do it

[7]Note that Yan and Zheng (2017) also provide international results using Compustat in their online appendix . The international results here are expected to be very different since we rely on Datastream which has much better international coverage historically. See Chapter 1 for description of problems connected to using fundamental database with imperfect coverage.

[8]See, for example, Lewellen et al. (2015), Green et al. (2017), Feng et al. (2017), and Freyberger et al. (2017) for the US evidence and Jacobs and Müller (2017a) for international evidence.

is very different here.

Our paper contributes to the international finance literature through its global focus. It broadly belongs to a class of studies investigating cross-sectional predictability of individual signals outside the US. See, for example, Chui et al. (2010), Barber et al. (2013), McLean et al. (2009), Rouwenhorst (1998), Lam and Wei (2011), Titman et al. (2013), and Watanabe et al. (2013). Some similarity in scope is also shared with papers investigating factor structure of international returns. See, for example, Fama and French (2012), Fama and French (2017), Rouwenhorst (1999), Griffin (2002), Griffin et al. (2010), Hou et al. (2011b), and Bartram and Grinblatt (2018b).

This paper contributes in a number of ways: firstly, it studies the significance of anomalies in multiple hypothesis context in different statistical settings. It shows that the choice of statistical setting can have a large impact on the critical values required for significance and number of significant signals. It also shows that the choice of adjustment of standard errors can have a large impact on number of significant anomalies detected. In particular, it proposes a new simulation approach to study power and size of statistical tests under an empirically realistic setting. The key takeaway is that t-statistics testing significance of annually rebalanced strategies are not well approximated by critical values of standard normal distribution and should be bootstrapped with block length 12, corresponding to the frequency of updates of the annual fundamental signals. Finally, the chapter revisits the value of data mining fundamental signals and documents that its performance can be heavily improved with proper tools.

2.1 Data and Methodology

2.1.1 Data

The source of accounting variables and trade data for US stocks is annual Merged CRSP/Compustat database. The US sample spans 1963 to 2016 period and contains all common stocks (CRSP share code 10 or 11). We adjust the returns for delisting following guidance in Hou et al. (2017).[9] We use three risk factor models in this study: CAPM, Fama and French (2015) five factor model (FF5), and Fama and French (1993) three factor models (FF3), which are taken from Kenneth French's website.[10] The source of data for global stocks is Reuters Datastream. The international sample includes 22 developed countries. We use the classification of Fama and French (2017) to sorting developed countries into 4 groups: (1) North America (United States and Canada); (2) Europe (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom);

[9] Specifically, we use return over the month if the delisting is on the last day of the month. Relevant delisting return is then added as a return over the next month. Then we use delisting return (DLRET) from monthly file if it is not missing. If it is missing then we use $(1 + ret_{cum}) * (1 + DLRET_d) - 1$, where ret_{cum} is cumulative return in the month of delisting and $DLRET_d$ is delisting return from the daily file. Lastly, we fill the gaps with $(1 + ret_{cum}) * (1 + DLRET_{avg}) - 1$, where $DLRET_{avg}$ is average delisting return for stocks with the same first digit of delisting code (DLSTCD).

[10] *http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html*

(3) Japan; and (4) Asia Pacific (Australia, New Zealand, Hong Kong, and Singapore).

The coverage of fundamental data in Datastream in individual counties is provided in Table 2.1. The coverage was weak in the beginning of 1980s but has progressively improved. The international sample starts in 1990 where there was large enough coverage for the USA, Europe, and Japan. There are only a few large cap stocks in Asia Pacific region. The stocks in individual countries are from the largest exchange in the given country with the exception of the US (NYSE, NASDAQ, and Amex) and Japan (Tokyo and Osaka).

We manually filter stocks following Ince and Porter (2006), Lee (2011), and Griffin et al. (2010). The procedure comprises manually checking names of the shares in the database for over 100 expressions describing their share class. Only primary quotes of ordinary shares of companies are left. We closely follow the description in Griffin et al. (2010) on what shares are not common. All REITs are also excluded. This selection procedure is not very important in the current work as stocks with fundamental coverage in Datastream are not plagued by as many errors and missing categorization compared to those without. The price of stocks at the time of portfolio formation, at the end of June, is required to be larger than \$1 with the exception of developed countries in Asia Pacific group where the cut-off is \$0.1.[11] The sample is restricted to industrial firms (WC06010 < 4) as Datastream constructs items in the financial statements differently for financial firms, banks, or insurance companies and they are thus not directly comparable. Adjustments of raw returns to improve their quality are described in the Chapter 1. Chapter 1 provides detailed coverage of adjustments to improve quality of data in Datastream.

The focus of our study is on universe of stocks that excludes micro-caps with size smaller than the smallest decile of stocks in the NYSE. Only a universe of stocks that can be traded in quantitative strategies without extreme transaction costs is therefore considered. More fundamental reason to focus on larger cap stocks is that Datastream has limited historical coverage of stocks with capitalization lower than 100 million USD. This can have a huge impact on the measurement of performance on individual signals as detailed in Chapter 1.

[11] We have selected lower required price for Asia Pacific as the minimum tick size is only \$0.001 there. Setting the threshold to \$1 would mean that almost 90% of the sample is discarded.

Table 2.1:

Number of Firms in Datastream with Accounting Information

The table shows number of industrial firms with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June that have fundamental coverage in Datastream or Compustat (just for USA) for at least 2 years.

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	···	2015
Australia	73	80	78	94	92	106	97	113	128	95	116	118	94	95	109	114	136	148	180	214	156	170		170
Austria	23	27	33	31	26	34	31	31	27	28	33	35	31	26	24	20	22	24	28	33	31	28		27
Belgium	24	31	31	31	25	31	30	25	26	35	47	49	37	30	33	29	39	46	44	52	46	44		44
Denmark	23	36	40	47	41	53	57	52	55	52	56	59	45	34	33	32	35	35	38	46	39	35		37
Finland	18	23	23	20	25	40	38	45	46	51	62	67	53	43	39	38	49	48	50	56	55	55		46
France	159	211	229	232	196	222	224	210	211	253	252	330	269	226	205	187	207	214	219	230	222	197		201
Germany	117	156	178	179	167	189	200	184	185	203	282	424	261	185	167	153	150	165	170	193	190	174		158
Greece	3	8	10	9	11	14	19	16	15	26	92	114	53	56	50	35	37	45	54	60	55	33		
Hong Kong	12	16	14	22	24	27	24	28	41	23	28	33	29	32	32	40	50	59	94	75	64	93		143
Ireland	16	19	19	19	15	17	18	16	15	15	16	17	16	15	10	13	14	13	17	12	10	10		13
Italy	101	116	110	98	73	82	62	61	54	76	76	111	113	84	84	75	90	103	121	124	100	84		85
Japan	936	993	1101	1368	1668	1742	1698	1751	1527	1004	1350	1617	1292	1042	1029	1080	1156	1132	924	1123	1281	1064		1045
Luxembourg				4	4	5	5	5	5	6	6	6	3	4	3	2	3	3	5	6	4	3		2
Netherlands	52	66	70	73	61	71	80	80	88	107	109	113	77	64	61	59	59	64	66	63	57	47		47
New Zealand	1	6	6	9	10	12	12	11	18	12	16	18	15	17	19	15	19	17	24	23	20	19		33
Norway	29	40	39	35	33	39	47	49	51	63	59	72	56	38	32	37	50	63	81	101	61	65		55
Portugal	10	21	23	22	16	21	20	18	23	23	22	29	17	18	18	17	17	19	22	26	25	20		16
Singapore	23	27	31	32	44	50	49	57	61	29	57	42	30	26	22	28	38	38	70	55	38	46		44
Spain	43	59	60	62	56	68	64	63	69	82	81	84	78	71	69	63	70	70	77	77	78	64		56
Sweden	42	49	41	46	54	69	70	73	77	97	87	114	88	72	65	61	66	78	81	87	82	79		93
Switzerland	70	90	87	80	73	84	94	85	87	102	105	120	109	98	81	79	88	91	102	120	110	105		104
UK	470	538	513	521	485	535	542	571	542	565	548	571	462	386	352	340	373	385	412	414	330	329		393
USA	1777	1864	1962	1980	2099	2217	2367	2477	2458	2514	2552	2624	2311	2095	2029	1942	1925	1890	1835	1851	1926	1837		1762

2.1.2 Construction of Data-mined Fundamental Signals

This section describes how we create the data-mined fundamental signals for potential anomalies. One of the concerns addressed in this paper is multiple hypothesis problem of selecting signals that are truly significant among many different alternatives. We deal with this problem by considering a large universe of signals so that the analysis does not suffer from sample selection bias from only considering published signals. The signals are constructed following Yan and Zheng (2017) from balance sheet, income, and cash flow statement items in either Datastream or Compustat. Yan and Zheng (2017) consider over 17,000 and Chordia et al. (2017) even 2.1 million signals but we restrict our main analysis to only 1,497. The reason for this is to work with a sample of signals that are as close to the universe of the published anomalies as possible. The chosen signals are very close to 25 anomalies. Examples include value, investment, and profitability anomalies in (Fama and French, 2015) and R&D anomalies in Eisfeldt and Papanikolaou (2013). We further consider a reduced set of 772 data-mined signals that are a subset of the 1,497 signals and an extended set of 48,387 signals in order to study the potential benefits of considering fewer or more signals.

Each signal is constructed by a transformation of numerator and denominator. We use 49 variables for numerator and list them in Table E.1 in the Appendix E. We have chosen all the fundamental variables that have large coverage (at least 1,000 stocks in the US every year since 1990) and have been used for the construction of signals for published anomalies in the next section. The denominators are: total asset (AT, WC02999); total liabilities (LT, WC03351); total common equity (CEQ, WC03501); stockholders' equity (SEQ, WC03501 + WC03451); total sale (SALE, WC01001); and market size (MKT-CAP). We apply the following 6 transformations relating numerator (X) and denominator (Y):

1. X_t/Y_t
2. $\triangle\ (X_t/Y_t)$
3. $\%\ \triangle\ (X_t/Y_t)$
4. $\triangle\ X_t/Y_{t-1}$
5. $(X_t + 0.8X_{t-1} + 0.6X_{t-2} + 0.4X_{t-3} + 0.2X_{t-4})/Y_t$
6. $\triangle\ X_t/X_{t-1}$

This together makes 1,519 signals out of which we exclude 20 signals where denominator is the same as numerator and further 2 that are completely identical to some the anomalies. The fifth transformation is motivated by Li (2011) and Eisfeldt and Papanikolaou (2013) who have discovered an anomaly based on accumulation of past R&D expenses. The motive behind the transformation is that it will capture a trailing average of the given variable.

The reduced set of 772 signals is constructed in the same way as the 1,497 signals but the set of denominators is constrained to total sale, market size, and stockholders' equity. The extended set of 48,387 signals also shares the same construction as the 1,497 signals but the set of numerators is extended by combining individual items on financial statements in a structured way. The numerator includes the following combinations of items on the financial statements:

1. Assets; any combination of CHE, RECT, INVT, ACO, IVAO, IVAEQ, INTAN, and PPENT.
2. Liabilities; any combination of AP, DLC, LCO, DLTT, and LO.
3. Net current assets; any combination of CHE, RECT, INVT, ACO minus any combination of AP, DLC, and LCO.
4. Net long term assets; any combination of IVAO, IVAEQ, INTAN, PPENT minus any combination of DLTT and LO.
5. Cash flow statement; any combination of OANCF, CAPX, IVNCF, PRSTKC, SSTK, DV, DLTIS, DLTR, DLCCH, and FINCF.
6. Income statement; SALE minus any combination of COGS, XRD, DP, XINT, TXT, XSGA minus XRD, and Accruals.

All of the items above are referenced in Table G.1 except for Accruals which is defined as a change in RECT plus a change in INVT plus a change in XPP minus a change in AP. The Accruals correspond to cash outflows not reflected in the income statement and are inspired by cash based operating profitability of Ball et al. (2016). The extended set of numerators results in 49,079 signals when combined with the 1,497 signals but only 48,387 of the 49,079 signals are unique.

2.1.3 Published Fundamental Anomalies

Further 93 anomalies published in academic journals are studied. The full list is provided in Appendix G. All of the anomalies have been described in McLean and Pontiff (2016), Hou et al. (2017), or Harvey et al. (2016). The sample includes all the fundamental anomalies that can be replicated outside the US and from which portfolios can be constructed via cross-sectional sorts of stocks.[12] The sole focus of this study is thus on cross-sectional characteristics of the stocks. The restriction of sample of stocks in construction of portfolios is the same as for data-mined signals.[13] That is, all financial firms

[12]This includes anomalies: based on quarterly fundamental data since there is only short coverage internationally; connected to hand collected data in the US such as IPOs, SPOs, and mergers; requiring segment information and NBER data; and that are institutionally specific such as share turnover or effective tax rate. Some fundamental anomalies could not be implemented in Datastream as the required items are missing there.

[13]Some anomalies also require the classification of industries such as Hou and Robinson (2006). The choice in the original papers is mostly with respect to SIC industry classification. We use third level

are excluded and the same restrictions on price and size at the end of previous June are applied.[14]

Anomalies that are based on short-term investments, and therefore have to be rebalanced more frequently than annually, are excluded from the analysis. All anomalies that are not constructed from fundamental data are also omitted. Most of the analysis is based on annually rebalanced portfolios and there are only few non-fundamental anomalies that are relevant. The most obvious ones are size, price, firm age, liquidity, and long-term reversals. Only long-term reversals are robustly significant on the universe of stocks excluding micro-caps but they are likely subsumed by the other signals. Omitted variable bias is therefore not a cause of worry.

2.1.4 Construction of Portfolios

Finally, we describe how we construct portfolios from the fundamental signals. The portfolios sorted on fundamental signals are rebalanced annually at the end of June every year, based on signals from business year ending in the previous calendar year. They are either value- or equal-weighted and are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. Portfolios based on published anomalies are always constructed to have positive returns in line with the findings in the original papers.[15] The portfolios are zero cost and returns correspond to monetary payoff each month. They are thus different from what an investor would get if he tried to invest in the signals as he would have to hold some collateral. Reason for this choice is that the value of collateral would often drop below zero within the 12 months before annual rebalancing period. The only solution would be to introduce leverage constraints and more frequent rebalancing, which would unnecessarily complicate the analysis.

2.2 Multiple Hypothesis Tests - Bootstrap Methods

When testing the statistical significance of new anomalies it is important to take into account the full universe of potential anomalies and try to include those that have not been published. The justification is simple: the value of t-statistic required for significance will be higher if 20 strategies are tested compared to testing only one. The difference is due to the fact that there is, on average, one false positive discovery among the 20 tested strategies. The false discovery appears to be significant in its individual test, while in fact it is not. It is important to control for these false discoveries in order to maintain the same rate of type I errors in the statistical tests.

Datastream classification which sorts industries into 19 groups instead. This has one main reason. The industry classification in Datastream is available only from the static file which means that only the latest value is available. Variation over time for individual firms between closely related SIC codes would thus cause problems.

[14]Constructing the portfolios on large cap universe but with the same restrictions as in the original studies has no effect on the main results of this study.

[15]That is, upper deciles of the (signed) signals are always used to produce long legs of the portfolios.

Every test in a classical statistical framework is framed in terms of type I (size) and type II (power) errors.

	Null hypothesis	
Decision	*True*	*False*
Reject	Type I error	OK
Not Reject	OK	Type II error

The goal is to select a test that will have the required size, typically 5%, and the largest possible power. There is always some trade-off between power and size unless the sample size in increasing. Tests that have smaller size tend to under-reject truly significant, and thus profitable, signals. In the present study this means that fewer fundamental signals are deemed significant. It is therefore important to apply appropriate methods with the largest possible power.

Harvey et al. (2016) studied the problem of identifying significant anomalies in a multiple hypothesis setting. They collected p-values reported in original studies and generated a hypothetical sample of p-values on all tried signals, thereby recreated the original sample of p-values before most of the tried strategies were discarded. However, the sample of p-values depends on strong underlying assumptions about structure of correlation among the anomalies. We take a more structured approach in this study by generating a universe of possible data-mined fundamental signals instead. This allows us to study the relation between individual anomalies in much greater detail. Specifically, it allows us to study the role of cross-sectional dependence between the signals. There are 93 published and 772, 1,497, or 48,387 data-mined signals in our sample, or about a 1:8, 1:16 ratio, or 1:520 ratio. This should provide very reasonable setting for the multiple hypothesis tests. Harvey et al. (2016) estimated that 71.1% of the tried signals were not published which translates to about 322 overall signals in our case with 93 published anomalies. This is fewer than 865 but we will show that the main results do not depend on the number of data-mined signals and the larger number is more reasonable due to the number of active researchers in the area over the years.

Harvey et al. (2016) reported that: "*We find that the difference in rejections rates produced by single and multiple hypothesis testing is such that most rejections of the null of no out-performance under single hypothesis testing are likely false.*" They then propose that the proper cut-off for t-statistics should be three. We will show in the rest of this section that this conclusion greatly depends on the precise specification of the tests. 63% of anomalies is significant under most favourable setting and the cut-off t-statistic is close to two, whereas, none of the anomalies is significant in the most conservative setting.

There are many simple correction methods for individual p-values to make them valid in multiple testing framework but these usually lead to poor power.[16] Harvey et al. (2016) had to rely on these methods since they did not have a ready access to the original data. We present three of the most frequently used methods. The simplest method is Bonferroni

[16] Good overview of the methods is provided in Harvey et al. (2016) and Chordia et al. (2017).

where p-value on individual tests are multiplied by a number of tests (M). The individual p-values then have to be M times smaller than the required size in single hypothesis tests. Holm (1979) provided a refinement by introducing a stepwise method where all the p-values are ordered from smallest to largest and the penalty is decreasing with their size. Specifically, the method rejects any hypothesis where $p_i(M+1-i) < \alpha$ for $1 \leq i \leq M$ and size α. This method is a refinement of Bonferroni. It tends to reject additional true positive hypothesis and is less strict for larger p-values. Benjamini and Yekutieli (2001) provide further refinement. The test proceeds again by first sorting p-values from the smallest to the largest so that $p_1 \leq p_2 \ldots \leq p_i \ldots \leq p_M$. False discovery rate (FDR) adjusted p-values are determined with backward induction where $p_M^{FDR} = p_M \sum_{1\leq j\leq M} \frac{1}{j}$ and

$$p_i^{FDR} = \min\left\{p_{i+1}^{FDR}, p_i \frac{M}{i} \sum_{1\leq j\leq M} \frac{1}{j}\right\} \tag{2.1}$$

The individual hypothesis are rejected with FDR of 5% if their adjusted p-values p_i^{FDR} are smaller than 5%.

The methods presented so far have focused on standard testing framework that controls for probability of at least one false positive discovery (type I error), but in practice this rapidly becomes too strict. The approach where we try to correct for probability of even one false positive discovery is denoted family-wise error rate (FWER). This assumption becomes too restrictive when there are many signals, as is the case here, and it is advantageous to allow for some false discoveries if it leads to acceptance of many positive discoveries. In our case of trading strategies, this means that several unprofitable strategies are accepted in order to select many more truly profitable strategies. The increase in number of profitable strategies should lead to a more profitable meta-strategy. This approach to the testing is defined by the maximum FDR, which is the proportion of false positive discoveries among all signals that were deemed significant. The rest of this section then discusses FDR methods that require bootstrap but should lead to greater power in the tests.

2.2.1 Cross-sectional Bootstrap

There are two types of bootstrap that we use in this study. A simpler block bootstrap resamples fixed blocks of returns on individual portfolios and its main purpose is estimation of p-values. A more complex block cross-sectional bootstrap then resamples blocks of the whole cross-section of returns on portfolios created from the fundamental signals. There are several ways how to introduce time dependence into the block bootstraps such as the bootstrap of Politis and Romano (1994) where the length of the block is assumed to follow exponential distribution with expected value of l. We will rely on a simpler version of circular bootstrap which was proposed in Politis and Romano (1992) and which resamples blocks of fixed length l. The benefit of this later bootstrap is that it produces more stable results due to lower uncertainty when we study the impact of the block length.

We use "naive" block bootstrap to estimate p-values on individual fundamental signals. The naive bootstrap consists of applying the same adjustment to the standard errors on both observed and bootstrapped returns. In particular, we apply Newey and West (1987) HAC robust estimator with a number of lags equal to the length of the blocks.[17] An alternative approach is to estimate t-statistics on original series with HAC robust estimator and to opt for a "natural" estimator of standard errors on the bootstrapped returns.[18] The "natural" estimator leads to almost identical findings and the choice of naive bootstrap thus does not influence our conclusions.

The block cross-section bootstrap that controls for time-series dependence in a financial setting was implemented in Fama and French (2010) and Kosowski et al. (2006). The main idea behind the bootstrap is to draw the complete cross-section of returns on all portfolios at the same time so that the correlation structure between them is preserved. The blocks of returns then allow for arbitrary structure for any time dependence. This is important for multiple hypothesis testing if returns on the portfolios tend to move together since alphas and their t-statistics then tend to be correlated. Resampling in blocks is especially important in an international setting where returns on a signal in one country may be related to returns on the same signal in other countries at leads or lags.

The null hypothesis implemented in the bootstrap corresponds to the "least favorable" conditions, that is, all the true alphas are equal to zero. This zero hypothesis is least favorable because it puts the largest hurdles for any potentially significant signal. Another approach introduced below will relax the assumption.

The bootstrap can be used for any statistic such as p-value, t-statistic, or alpha. Yan and Zheng (2017) give their preference to evidence in t-statistics, since they are pivotal statistics and should be less prone to outliers. The use of t-statistics instead of alphas is also recommended in Romano et al. (2008) and we make the same choice here to make our analysis more robust.

The bootstrap proceeds in the following steps:

1. Create portfolios based on fundamental signals.

2. Estimate alpha with respect to factor models and any statistic of interest.

3. Remove alpha from the portfolios, i.e. implement the null hypothesis of no alphas. Bootstrap proceeds with these adjusted returns.

4. Draw a sample of time periods of the same length as the original portfolios. Suppose that the index for the original sample period is 1 to T. The time-periods are drawn with replacement as a sequence by drawing start period t from 1 to T with equal probability and length s. The first piece of sequence is then adjusted to t, t+1,

[17] See Goncalves and Vogelsang (2011) for asymptotic theory and detail on the method. The method was originally covered in Götze et al. (1996).

[18] See Götze et al. (1996) for asymptotic theory and Romano and Wolf (2006) for automatic selection of block length.

..., t+s-1 mod T to stay in the original sample period of 1 to T. The rest of the sequence is drawn until it includes T elements.

5. Create a new sample of returns by drawing the whole cross-section of returns with the sequence of time periods as specified in the previous step.

6. Estimate alpha with respect to factor models the statistic of interest from resampled returns.

7. Repeat steps 4 to 6 10,000 times.

8. Compute required statistics from values from step 2 and 6.

2.2.2 Storey (2002)

The simple cross-sectional bootstrap is suitable for the estimation of the probability of generating the returns on all strategies by pure chance. However, it does not offer any guidance on number of significant signals. The following method captures the number. The method was developed by Storey (2002) and introduced into finance in Barras et al. (2010) and Bajgrowicz and Scaillet (2012).

The method is based on a simple idea that we can infer proportion of true null hypothesis from distribution of p-values on single hypothesis tests of significance of individual strategies. It would be expected that the p-values would be uniformly distributed on $[0, 1]$ interval if there are no truly significant strategies. This corresponds to the case where the strategies contain only noise. The p-values would cluster around 0 if there is some proportion of profitable strategies.

Figure 2.1 plots the distribution of p-values on value-weighted portfolios of the data-mined and published strategies. Their returns are adjusted with the CAPM model. It is indeed the case that the p-values cluster around zero. The method is based on a simple idea that the number of deep-in-the-null strategies should well approximate the true proportion of strategies with zero expected return. The solid horizontal red line shows expected density of strategies that are truly insignificant. The number of strategies can be estimated with

$$\frac{\sum_i \{p_i \geq \lambda\}}{1-\lambda} = \pi_0. \tag{2.2}$$

Bajgrowicz and Scaillet (2012) suggest setting $\lambda = .6$ and we follow the suggestion. Values between .4 and .7 do not have a large impact on the findings. The dashed horizontal red line then shows what the density of p-values has to be in order for the signals to be significant with FDR of 10% at a given p-value. Number of significant strategies can then be estimated with

$$\max_i \left\{ \frac{\pi_0 p_i}{i} \leq \gamma \right\} \tag{2.3}$$

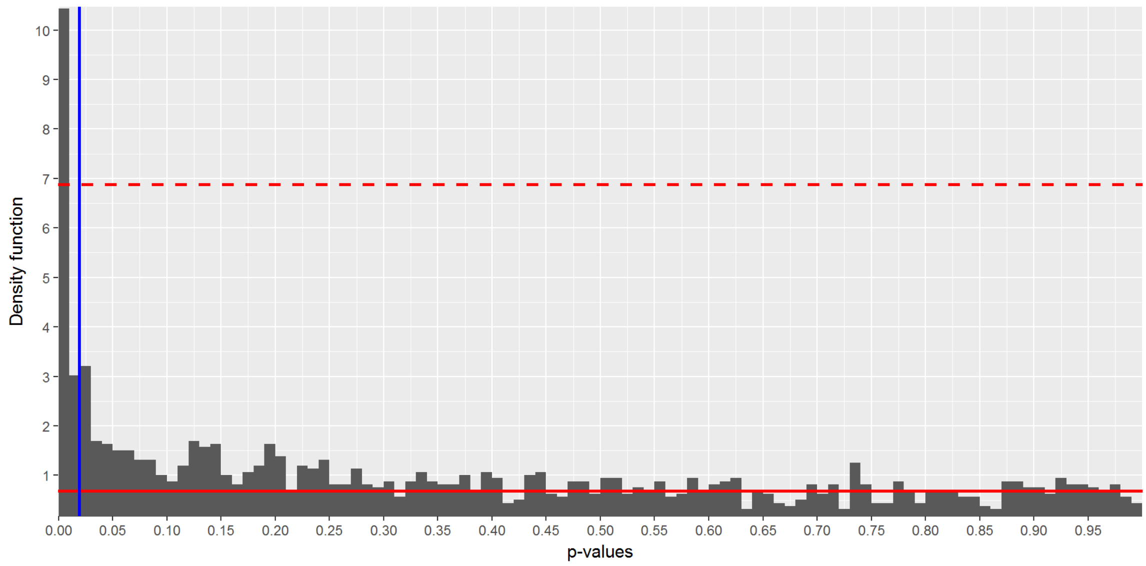


Figure 2.1: **Distribution of p-values.** The figure shows density of bootstrapped p-values on 1,590 fundamental signals, 1,497 data-mined and 93 from published studies. The sample is restricted to industrial stocks with price over \$1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016. The value-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with CAPM. Standard errors in t-statistics are HAC adjusted, as in Newey and West (1987) with 3 lags.

where γ stands for the user selected FDR. The vertical solid blue line shows critical value of p_i^*. It is apparent that the method also accepts strategies with marginal FDR for a given p-value of less than γ. This is because the test accepts exactly γ of false discoveries and thus continues to accept hypotheses that are under the line to compensate for a larger number of significant signals with very small p-values.[19]

The method is very simple but has its limitation. Barras et al. (2010) originally applied it to returns from mutual funds which are not heavily correlated with each other. They thus relied on independence between the funds. This claim is harder to maintain for portfolios created from the same dataset. Bajgrowicz and Scaillet (2012) relaxed this strong assumption and showed that it is also consistent under weak dependence and block dependence. Weak dependence occurs when the signals are asymptotically independent. Block dependence means that returns on portfolios can be correlated in blocks of signals but number of these blocks tends to infinity as number of signals increases.

Bootstrap is not required for the method, strictly speaking, but Bajgrowicz and Scaillet (2012) use it for estimation of the p-values. It is related to the cross-sectional method described earlier in that they should both lead to similar inference for $\pi_0 = M$. The relationship is easy to show when p-values are bootstrapped instead of the usual t-statistics in the cross-sectional bootstrap. Implementation of the null hypothesis of no outperforming strategies should translate into a distribution of bootstrapped p-values that are roughly

[19]Note that this could lead to sub-optimal selection of strategies for out-of-sample tests but can be simply remedied by setting stricter desired FDR.

uniform on the $[0, 1]$ interval. We are then comparing the realized distribution of p-values with a uniform distribution. Setting $\pi_0 < M$ leads to conditions that are less strict than the least favorable conditions, which leads to a larger power to reject the outperforming strategies.

The two methods thus differ in their null hypotheses. Barras et al. (2010) assume that there is some fraction of strategies with alpha equal to zero but there is also a fraction that genuinely outperforms. The method then tries to separate these two sets to keep the error rate at the specified level. The cross-sectional bootstrap, on the other hand, assumes that there are no outperforming strategies.

2.2.3 Empirical Results

Cross-sectional Bootstrap

Table 2.2 provides bootstrap evidence on likelihood of the observed t-statistics being generated by pure chance from signals with zero true alpha. The likelihood is estimated with cross-sectional bootstrap with 93 published anomalies included along with the 1,497 data-mined signals. The columns with p-values provide the proportion of simulation runs where a given quantile of absolute value of t-statistics on alphas was higher than the quantile observed in the original sample. The reasoning behind quantiles of absolute values is that the signal would be deemed significant if it was either significantly positive or negative and skewness in samples would distort this evidence for positive or negative values.[20] We focus on the US in Panel A, Europe in Panel B, Japan in Panel C, and on Asia Pacific in Panel D.

Standard errors on t-statistics in the table are HAC robust and adjusted with Newey and West (1987) procedure with 3 lags. Large number of lags leads to large p-values and this effect can be significant. All the p-values when adjusting for FF5 model are generally larger than 5% for 24 lags and would thus lead to no anomalies under $FDR = 5\%$. We have selected 3 lags as the bootstrap of t-statistics is then very similar to bootstrap of alphas. The increase in p-values hints on loss of power for more lags.

The table documents that neither value-weighted nor equal-weighted returns can be plausibly randomly generated in the US. p-values on equally-weighted portfolios are generally smaller and thus lead to possibility of additional significant signals. Even value-weighted returns do not, however, have large p-values. None of the factor model is able to plausibly explain returns on the signals.

The European sample leads to very similar conclusions as the US sample. Japanese sample is very different as there is no sign of any violation of market efficiency and all the fundamental signals could be generated by pure chance with p-values around 50%. This will later translate into no statistically significant signals there in multiple hypothesis tests since it is not possible to statistically distinguish between true or false positive signals.

[20] It is indeed the case that alphas tend to be more clustered on either positive or negative side in the US and Japan. This is possibly due to a latent risk factor that has not been properly accounted for.

Table 2.2:
Simulated p-values of Quantiles of t-statistics

The table shows quantiles of t-statistics on alphas on long-short portfolios created from the sorts on fundamental signals in the US, Japan, Europe, and Asia Pacific with their bootstrapped p-values from 10,000 runs. We employ cross-sectional time dependent bootstrap described in Section 2.2.1 to determine the p-values. The bootstrap is conducted on 1,590 fundamental signals, 1,497 data-mined and 93 from published studies. The 1,497 data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. The sample is restricted to industrial stocks with price over \$1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 for the US and July 1990 to December 2016 elsewhere. The value-weighted or equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with regional versions of CAPM, Fama-French three, and five factor models. Standard errors in t-statistics are HAC adjusted, as in Newey and West (1987) with 3 lags.

	Equal-weighted portfolios						Value-weighted portfolios					
	CAPM		FF3		FF5		CAPM		FF3		FF5	
	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value	t-stat	p-value
Panel A: Absolute value of t-statistic in the US												
100	6.800	0.000	8.574	0.000	7.848	0.000	5.269	0.001	6.473	0.000	5.385	0.003
99.9	6.392	0.000	7.082	0.000	6.276	0.000	4.890	0.000	4.486	0.003	4.761	0.003
99	5.602	0.000	4.948	0.000	4.433	0.006	4.173	0.000	3.793	0.001	3.892	0.001
98	5.356	0.000	4.536	0.000	4.163	0.004	3.861	0.000	3.530	0.001	3.562	0.001
97	5.066	0.000	4.271	0.000	4.008	0.003	3.637	0.000	3.336	0.001	3.395	0.001
96	4.876	0.000	4.155	0.000	3.833	0.003	3.425	0.000	3.132	0.001	3.245	0.001
95	4.746	0.000	3.998	0.000	3.715	0.003	3.262	0.000	3.013	0.001	3.177	0.000
90	3.862	0.000	3.375	0.000	3.223	0.002	2.664	0.000	2.613	0.000	2.762	0.000
Panel B: Absolute value of t-statistic in Europe												
100	6.118	0.000	6.543	0.000	5.517	0.004	4.204	0.062	4.994	0.015	4.509	0.066
99.9	5.345	0.000	5.919	0.000	4.690	0.012	4.095	0.021	4.575	0.013	4.024	0.076
99	4.690	0.000	5.000	0.000	3.892	0.007	3.388	0.015	3.367	0.039	3.550	0.021
98	4.333	0.000	4.628	0.000	3.468	0.011	3.003	0.025	3.067	0.038	3.183	0.025
97	4.159	0.000	4.388	0.000	3.232	0.014	2.765	0.032	2.899	0.034	3.005	0.022
96	3.969	0.000	4.114	0.000	3.126	0.011	2.655	0.030	2.695	0.046	2.742	0.037
95	3.807	0.000	3.958	0.000	2.959	0.013	2.567	0.027	2.565	0.050	2.677	0.028
90	3.203	0.000	3.228	0.001	2.505	0.016	2.133	0.041	2.204	0.041	2.116	0.071
Panel C: Absolute value of t-statistic in Japan												
100	4.411	0.068	4.063	0.205	3.697	0.377	3.093	0.636	3.409	0.497	3.252	0.631
99.9	3.432	0.264	3.278	0.421	3.148	0.530	2.752	0.707	3.090	0.497	2.901	0.692
99	2.575	0.422	2.843	0.272	2.789	0.307	2.099	0.822	2.537	0.436	2.505	0.496
98	2.400	0.372	2.683	0.214	2.600	0.255	1.880	0.837	2.267	0.473	2.300	0.468
97	2.253	0.368	2.581	0.179	2.467	0.235	1.751	0.838	2.141	0.459	2.182	0.438
96	2.139	0.368	2.462	0.172	2.387	0.204	1.617	0.872	2.046	0.443	2.048	0.460
95	2.028	0.390	2.382	0.159	2.293	0.198	1.538	0.872	1.968	0.435	1.979	0.439
90	1.716	0.395	2.023	0.158	1.923	0.213	1.277	0.878	1.700	0.409	1.690	0.430
Panel D: Absolute value of t-statistic in Asia Pacific												
100	5.832	0.001	5.630	0.003	5.459	0.005	4.365	0.054	4.377	0.079	5.054	0.053
99.9	5.522	0.001	5.152	0.002	4.364	0.025	3.857	0.076	4.234	0.042	3.952	0.154
99	4.465	0.000	4.118	0.000	3.650	0.005	3.374	0.018	3.378	0.043	3.080	0.167
98	3.960	0.000	3.815	0.000	3.390	0.003	3.025	0.022	3.174	0.028	2.687	0.207
97	3.729	0.000	3.683	0.000	3.193	0.003	2.853	0.019	2.964	0.028	2.462	0.233
96	3.583	0.000	3.473	0.000	2.994	0.004	2.699	0.021	2.798	0.030	2.293	0.261
95	3.432	0.000	3.371	0.000	2.886	0.003	2.485	0.039	2.720	0.026	2.229	0.237
90	2.981	0.000	2.922	0.000	2.381	0.006	2.086	0.047	2.364	0.018	1.869	0.248

Proportion of Significant Signals

The cross-sectional bootstrap has shown that the fundamental signals can be plausibly explained by pure chance only in Japan. We will now study number of significant signals in the multiple hypothesis tests. Table 2.3 presents number of significant data-mined signal and anomalies with corresponding critical values. The significance of signals is determined from the multiple hypothesis tests applied to returns on portfolios sorted on the individual signals. The tests are performed on the 865 (Reduced), 1,590 (Base), and 48,480 (Extended) portfolios of data-mined signals and anomalies. The proportion of total significant signals (N) and proportion of significant anomalies (N_A) are provided. The proportion of significant signals in the multiple hypothesis tests is determined with the Storey (2002) test. False discovery rate is set at 5% for all the specifications. Panel A is based on all the 93 anomalies while Panel B is based the reduced set of 25 anomalies that are the closest to the data-mined signals.

Number of significant signals depends on settings of the tests. There tend to be many more significant signals for equal-weighted returns on portfolios. Additional risk factors, that we adjust the performance for, also tend to depress the number of significant signals. The proportion of significant signals is the smallest with FF5 model. This support the evidence in Fama and French (2017) that their five factor model is useful in explanation of anomalies. The table also documents that a larger proportion of anomalies is significant than a proportion of data-mined signals. There are, for example, about 40% significant data-minded signals with CAPM and equal-weighted returns in the US but about 60% anomalies. This is as expected since published anomalies had to overcome significance level hurdles when they were published and are thus not generated by pure chance.

There is a large difference in the number of significant signals across the regions. There are generally no significant signals in Japan, as would be expected from the previous results with the cross-sectional bootstrap.[21] The number of significant signals in Europe and Asia Pacific is also lower with respect to the US. This is partly due to the shorter sample there. The drop is much more apparent for value-weighted portfolios that are significant only for few signals outside the US.

The results are very similar between Panel A and Panel B which confirms that the data-mined signals are a good approximation of the universe of potential fundamental signals for all the anomalies. The 25 anomalies in Panel B are all very close in construction to the 1,497 data-mined signals in the base case. The proximity guarantees that the multiple hypothesis tests are factually completely correct and the anomalies could be picked from the universe of data-mined signals using a mechanical rule. Some of the 93 anomalies in Panel A do not have a close construction to the data-mined signals. The approximation of universe of tried strategies therefore has to rely on the close correlation of portfolio returns on anomalies and data-mined signals rather than the fundamental signals themselves. The

[21]Note that rebalancing of portfolios at the end of June is a very unfavorable assumption in Japan where 73% of the firms have their accounting year ending in March. Items on financial statements then take 15 months to appear in the fundamental signals unlike the usual 6 months in the US. Rebalancing portfolios at the end of October leads to the same conclusion in that there are no significant signals in the multiple hypothesis tests. Distribution of t-statistics on portfolio returns created based on the individual signals, however, has heavier tails suggesting that October rebalancing ostensibly leads to more profitable signals.

Table 2.3:
Multiple Hypothesis Tests

The table shows proportion of significant fundamental signals (N) in percentage points and corresponding critical p-values (p-val) under Storey (2002) multiple hypothesis framework controlling for 5% FDR. We also show proportion of significant anomalies among the signals (N_A). The p-values and distribution of t-statistics is approximated with circular block bootstrap with 1,000 runs and block size equal to lags in HAC adjustment of errors. The bootstrap is conducted on 865 (Reduced), 1,590 (Base), and 48,480 (Extended) fundamental signals in Panel A, 93 signals from published studies and the rest data-mined. Panel B further restricts the number of published anomalies to 25 that are closely tied to the data-mined signals. The data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 for the US and July 1990 to December 2016 elsewhere. We run the multiple hypothesis tests independently on equal-weighted and value-weighted portfolios. The value-weighted or equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with regional versions of CAPM, Fama-French three, and five factor models. Standard errors in t-statistics are HAC adjusted, as in Newey and West (1987) with 3 lags.

	Equal-weighted Portfolios									Value-weighted Portfolios								
	CAPM			FF3			FF5			CAPM			FF3			FF5		
	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val
Panel A: 93 Anomalies																		
USA 1963-2016																		
Reduced	47	63	6.6	32	46	3.5	22	38	1.9	11	26	0.9	12	29	1.1	12	26	1.0
Base	37	59	4.2	27	48	2.8	18	33	1.6	8.1	23	0.6	6.9	22	0.6	8.0	24	0.7
Extended	39	62	4.6	28	46	2.6	26	44	2.7	6.2	24	0.5	7.1	20	0.6	4.7	17	0.4
Japan 1990-2016																		
Reduced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	26	43	2.8	23	34	2.2	2.7	5.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	19	37	1.8	18	30	1.6	2.6	5.4	0.2	0.0	0.0	0.0	0.8	3.2	0.1	0.0	0.0	0.0
Extended	16	34	1.4	14	26	1.1	1.0	4.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	21	30	1.9	21	30	2.0	4.2	4.3	0.3	1.2	3.2	0.1	2.3	2.2	0.2	0.0	0.0	0.0
Base	20	31	1.7	20	32	2.0	2.5	4.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	5.0	16	0.4	1.3	4.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Panel B: 25 Anomalies																		
USA 1963-2016																		
Reduced	43	64	5.7	28	40	2.8	18	32	1.4	9.2	24	0.7	7.3	16	0.7	9.2	24	0.8
Base	36	64	3.9	25	36	2.4	17	28	1.4	7.2	24	0.5	3.1	0.0	0.3	5.7	16	0.5
Extended	39	68	4.6	28	40	2.6	26	40	2.7	6.2	24	0.5	7.1	4.0	0.6	4.7	16	0.4
Japan 1990-2016																		
Reduced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	23	52	2.2	21	44	1.8	1.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	18	48	1.6	17	32	1.5	1.5	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	16	44	1.4	14	28	1.1	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	20	36	1.5	18	36	1.7	3.3	0.0	0.2	0.0	0.0	0.0	2.3	0.0	0.2	0.0	0.0	0.0
Base	19	36	1.6	19	40	1.8	2.3	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	5.0	16	0.4	1.3	4.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

proximity of results between Panel A and Panel B means that the conclusions from the multiple hypothesis tests are not heavily influenced by the assumption. The results are also similar across the three sets of data-mined signals. The proportion of significant signals tends to be the smallest for the extended set of data-mined signals which is as expected as there should be a larger proportion of signals with pure noise among them if the set of anomalies is finite. The assumption on number of tried unpublished signals therefore does not overly influence the analysis.

Harvey et al. (2016) showed that t-statistic of two is far too low for single hypothesis tests of anomalies. Their advice is that a threshold of three is much more realistic given the number of tried signals and, therefore, should be required instead of two. We show that this threshold depends on settings of the tests of significance. The critical value generally increases with number of factors that we adjust returns on anomalies with. The critical p-value for equal-weighted portfolio in the US is 4.2% for CAPM but 1.6% for FF5 for the base number of data-mined signals. The p-values correspond to t-statistics of 2.03 and 2.41, respectively, under an assumption of normal distribution.This is higher than 1.96 normally required in individual tests but it is much lower than three proposed in Harvey et al. (2016). The critical value also depends on weighting of the returns. Value-weighted returns tend to require much higher threshold since many more signals can be explained by pure chance. To conclude, the critical value of t-statistic for value-weighted returns tends to be close to three as suggested in Harvey et al. (2016) but this threshold is much lower for equal-weighted returns since there are many more outperforming signals present.

The table is based on standard errors that are HAC robust per Newey and West (1987) with 3 lags. The choice of adjustment of covariance matrix for heteroskedasticity and autocorrelation can have a large impact on the number of significant signals as it influence both power and size of the tests. The impact of increasing the number of lags is discussed in Section 2.3.

Proportion of Significant Signals after Omitting Closely Correlated Signals

The analysis has so far focused on signals without any regard for correlation structure between them. The anomalies are expected not to be heavily correlated as that is a prerequisite for their publishing. The same does not, however, apply to the data-mined signals. It is possible that there are many data-mined signals related to just one true anomaly. Accruals or leverage can, for example, have many forms. We will now consider only strategies that are not heavily correlated and we will study its impact on the multiple hypothesis tests.

Table 2.4 is generated in the same way as the Table 2.3 but the set of signals is reduced by discarding signals that are heavily correlated to any other signal. The strategies are discarded iteratively so that the correlation between two equal-weighted portfolios in the US is at most 80%. The selection process starts with the 93 anomalies and discards any other closely correlated signal so that as many anomalies as possible are preserved. Only

459 (Reduced), 578 (Base), or 5,455 (Extended) signals survive which implies that many of the data-mined signals were indeed closely connected.

The results in Table 2.4 are very similar to Table 2.3 with one prominent exception. The number of data-mined signals now has powerful impact on proportion of the significant signals. Setting with the extended number of signals leads to the smallest proportion of the significant signals and the smallest critical threshold for p-values. This can be interpreted as an increase in proportion of signals that are just noise when signals closely related to the original 1497 signals are discarded. The correlation structure among the signals is therefore of first order importance and the portfolio setting offers only limited insights when it comes to independently significant signals controlling for the other signals.[22]

2.3 The Role of Estimator of Standard Errors of Portfolio Returns

Previous section has noted that the adjustment of standard errors on returns on anomalies can have a vast impact on a number of significant signals. We will now demonstrate why it is the case. We will first discuss problems with asymptotic distribution of t-statistics with HAC adjusted standard errors. The common methods predict that the distribution is normal. We will, however, show that it is not normally distributed in practice. The asymptotic distribution is derived under an assumption that the number of lags in the HAC adjustment divided by the length of the time series approaches zero, which is never the case in practice. We will then focus on the selection of appropriate number of lags in the adjustment.

There is no prior published evidence or consensus in the literature on the choice of the appropriate estimator of standard errors, to the best of our knowledge. Different authors usually choose the adjustment arbitrarily without any justification. The issue is so neglected that most of the authors do not even mention any adjustment. See, for example, Fama and French (2015), Fama and French (2016), Fama and French (2017), Sloan (1996), and Ang et al. (2006b). Other studies report HAC robust standard errors but the choice of number of lags in the HAC adjustment is again arbitrary. Eisfeldt and Papanikolaou (2013) and Cooper et al. (2008), for example, adjust for one lag, Ang et al. (2009) adjust for four lags, while Green et al. (2017) choose twelve lags.

2.3.1 Fixed-b Asymptotic Distribution for HAC Robust Standard Errors

Commonly used HAC robust standard errors estimators, such as Newey and West (1987), operate based on the asymptotic theory that predicts that their resulting t-statistics are normal. It is then possible to derive appropriate p-values from quantiles of normal

[22]Note that there are now only 12 anomalies that are closely connected to the data-mined signals in Panel B which makes the results there anecdotal and hard to interpret.

Table 2.4:
Multiple Hypothesis Tests Omitting Closely Correlated Signals

The table shows proportion of significant fundamental signals (N) in percentage points and corresponding critical p-values (p-val) under Storey (2002) multiple hypothesis framework controlling for 5% FDR. The table is constructed identically to Table 2.3 except that the set of signals is restricted so that the correlation between two equal-weighted portfolios in the US is at most 80%. The bootstrap is conducted on 459 (Reduced), 578 (Base), and 5,455 (Extended) fundamental signals in Panel A, 69 signals from published anomalies and the rest data-mined. Panel B further restricts the number of published signals to 12 that are closely tied to the data-mined signals.

	Equal-weighted Portfolios									Value-weighted Portfolios								
	CAPM			FF3			FF5			CAPM			FF3			FF5		
	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val
Panel A: 69 Anomalies																		
USA 1963-2016																		
Reduced	37	59	4.4	39	57	4.4	26	46	2.6	14	28	1.1	9.6	23	0.8	4.8	13	0.4
Base	34	57	3.6	31	52	3.0	24	45	2.4	10	28	0.8	6.9	19	0.6	4.2	14	0.4
Extended	22	51	2.0	22	48	2.1	15	35	1.3	2.9	20	0.2	2.6	13	0.2	1.4	10	0.1
Japan 1990-2016																		
Reduced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	19	36	1.8	15	25	1.3	3.9	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	13	29	1.2	13	26	1.1	4.0	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	9.0	28	0.7	8.7	23	0.7	1.2	2.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	11	22	0.9	15	23	1.5	1.7	2.9	0.2	0.9	2.9	0.1	0.9	1.4	0.1	0.0	0.0	0.0
Base	12	22	0.9	16	25	1.4	2.4	4.3	0.2	0.9	1.4	0.1	0.9	1.4	0.1	0.0	0.0	0.0
Extended	1.4	7.2	0.1	1.3	4.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Panel B: 12 Anomalies																		
USA 1963-2016																		
Reduced	34	67	3.4	32	50	3.1	18	33	1.6	10	33	0.7	2.0	0.0	0.2	3.0	17	0.2
Base	30	67	2.8	27	50	2.3	19	42	1.7	6.1	33	0.4	1.7	0.0	0.2	1.3	8.3	0.1
Extended	22	67	2.0	21	50	2.0	15	33	1.3	2.7	25	0.2	2.5	0.0	0.2	1.3	8.3	0.1
Japan 1990-2016																		
Reduced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	13	33	1.1	12	25	0.9	2.7	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	10.0	33	0.8	11	33	0.8	3.6	8.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	8.8	50	0.7	7.9	25	0.6	1.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	6.2	25	0.5	10	17	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	10	33	0.7	12	17	1.0	2.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	1.3	8.3	0.1	1.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

distribution. This asymptotic theory is derived under an assumption that as length of time series (T) goes to infinity, number of lags of autocorrelations (M) divided by T goes to zero. In other words, the number of lags that we adjust for grows at a slower rate than length of the time series. The assumption tacitly implies that M/T (b) should be very close to zero. This is, however, never the case in practice and b is always positive. Authors of these estimators of covariance matrix were well aware of this fact and Andrews and Monahan (1992) report "*As shown in the Monte Carlo results of Andrews (1991), however, the kernel estimators considered in the above papers all perform quite poorly in certain contexts. In particular, kernel HAC covariance matrix estimators often yield confidence intervals whose coverage probabilities are too low (equivalently, test statistics* *that reject too often) and this phenomenon is not attributable to a particular choice of kernel or bandwidth parameter.*"

Aware of these shortcomings, Kiefer et al. (2000) developed a new asymptotic theory that does not suffer from these over-rejections. They labeled it fixed-b asymptotics since it assumes that $b = 1$. They then extended their results to $b \in [0, 1]$ in Kiefer and Vogelsang (2005). They show that under Bartlett kernel, as in Newey and West (1987), the distribution of test t-statistic is not normal but depends on b. The critical value for two sided test of a single hypothesis at 5% level is 2.02 assuming that the number of lags is twelve and time period is 1963 to 2016, as in our US sample. The critical value is higher than the usual 1.96 for normal distribution and fewer signals are therefore rejected. The critical value, however, goes up to 4.771 when $b = 1$. The density of the new standardized distribution is close to normal but it has larger tails. The problem is thus not only related to t-statistics having larger variance as the whole distribution is distorted. Choice of lags is a problem that does not disappear with this approach. The new asymptotic theory gives correct size to the tests. Choosing number of lags that is far from its true value, however, leads to poor power of the tests and lowers number of rejections of truly out-performing signals. We will discuss the power in more detail below.

So why does asymptotic theory matter in our case? First, we cannot rely on critical values of standard normal distribution. It is therefore advantageous to rely on a bootstrap that takes care of this problem. Second, and more importantly, HAC adjustment can cause large problems when mixing signals of various lengths. This is exactly our case where some anomalies and data-mined signals cannot be constructed for the whole sample. The result is a mix of short and long time series with different critical values for significance of their t-statistics. This can then lead to under-rejection of truly significant signals in both cross-sectional bootstrap and Storey (2002) test. Some of the signals will tend to provide more extreme values under the null and this will create noise in the tests. The problem could be further exaggerated by resampling, as it is often the case that bootstrap selects very short time-series in some samples. The t-statistics are therefore not unconditionally pivotal, which destroys their main benefit.

Figure 2.2 plots bootstrapped density of .95 quantile of absolute value of t-statistics on 1590 data-mined signals and anomalies. We use naive bootstrap and the same setting as described before. The portfolios are value-weighted.[23] Standard asymptotic theory predicts that the density should be centered around 1.96. This is obviously not the case and appropriate critical value is closer to two. It is also worth noting that the problems with distribution of critical values are exacerbated by more complicated factor models. The critical value for CAPM is significantly lower than for FF5 model. The distribution is significantly shifted to the right for more lags in the adjustment.

Is there any easy solution to non-standard distribution of t-statistics? At many places of this study, it is required to estimate p-values. This would be very complicated if we

[23]Equal-weighting leads to even larger distortions.

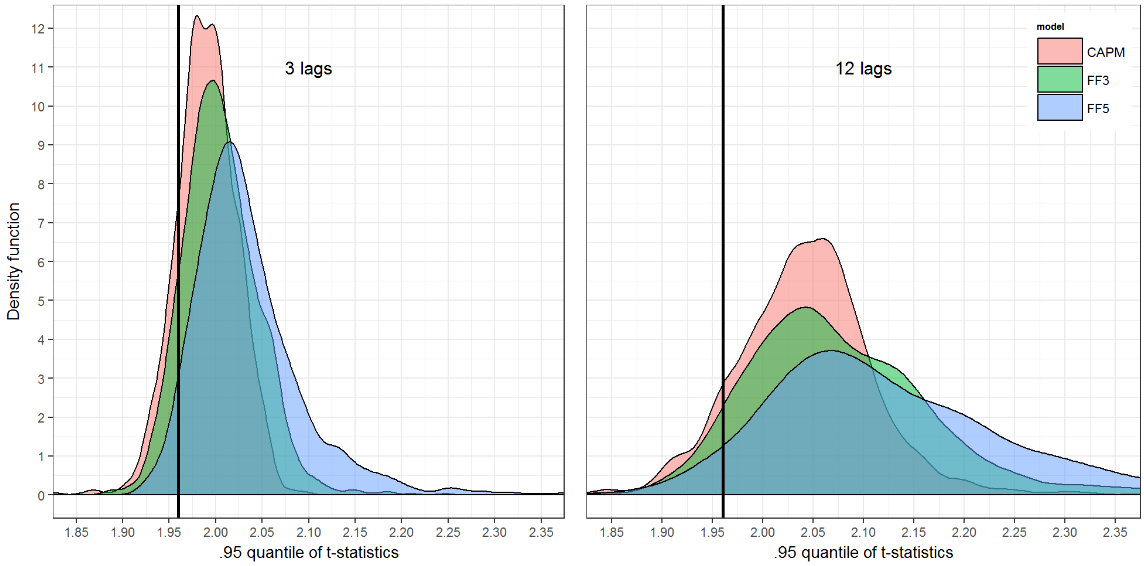


Figure 2.2: **Distribution of 95th percentile of bootstrapped t-statistics.** The figure shows bootstrapped density of 95th percentile of t-statistics on 1,590 fundamental signals, 1,497 data-mined and 93 from published studies. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016. The value-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with CAPM and Fama-French three and five factor models. Standard errors in t-statistics are HAC adjusted, as in Newey and West (1987) with either 3 lags or 12 lags.

had to rely on non-standard distribution to derive them. Fortunately, Goncalves and Vogelsang (2011) showed that the naive bootstrap shares many characteristics with fixed-b asymptotic theory and it correctly adjusts the p-values.

2.3.2 Impact of Number of Lags in HAC Adjustment on Number of Individually Significant Signals

The previous sections discussed the problems related to standard HAC robust covariance matrix in Newey and West (1987) and showed that it is possible to overcome them with the correct asymptotic theory. This section discusses the last required ingredient - number of lags; or more precisely kernel bandwidth.

Andrews and Monahan (1992) and Newey and West (1994) have proposed automatic rules to select the bandwidth. Problem with the rules is that they were optimized to select appropriate standard error but not for optimal confidence interval coverage in tests. Sun et al. (2008) explain the problem: "*For typical economic time series, the optimal bandwidth that minimizes a weighted average of type I and type II errors is larger by an order of magnitude than the bandwidth that minimizes the asymptotic mean squared error of the corresponding long-run variance estimator.*" The automatic selection rules thus tend to provide lower number of the required lags which results in confidence intervals under-coverage and rejection of too many hypothesis. Sun et al. (2008) also provide a new rule for automatic selection that overcomes these problems. It is, however, not suited for the

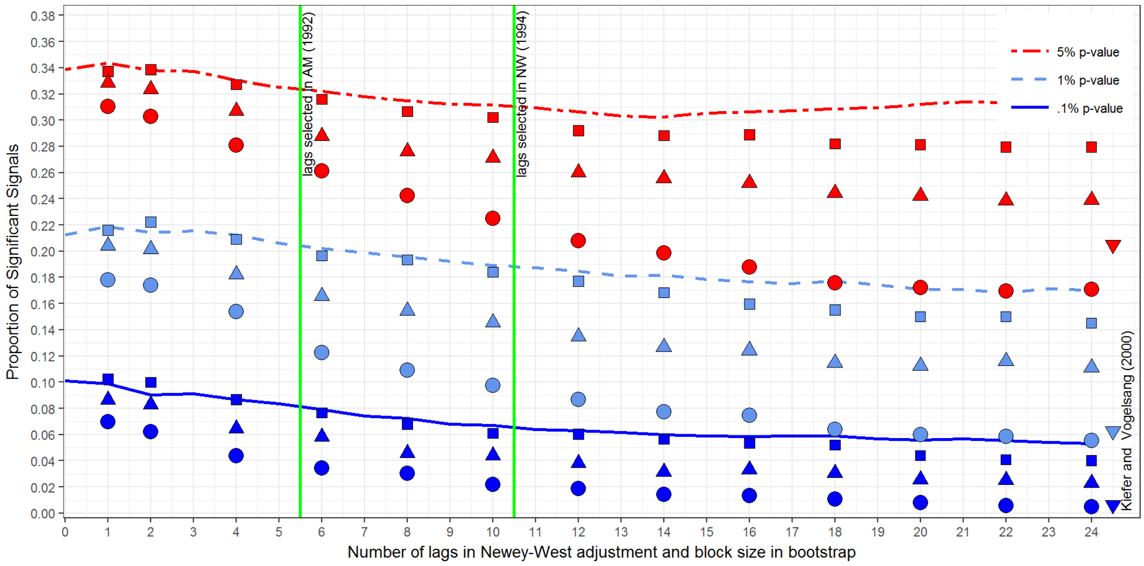


Figure 2.3: **Proportion of Significant Signals Depending on Adjustment of Standard Errors.** The figure shows the proportion of significant signals at 5%, 1%, and .1% level as a function of adjustment of standard errors and block length in bootstrap. Lines show the proportion as a function of lags in Newey and West (1987) adjustment for autocorrelation and heteroskedasticity. Squares, triangles, and circles stand for bootstrapped values using alphas, unadjusted t-statistics, and HAC adjusted t-statistics as in Newey and West (1987), respectively. The figure is based on 1,590 fundamental signals, 1,497 data-mined and 93 from published studies. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016. The equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with Fama-French five factor model.

type of bootstrap that we use here and is again based on an asymptotic behaviour. The problem of selection of bandwidth does not disappear for the bootstrap but it translates into selection of block length.[24]

We next turn to assessment of impact of choice of the bandwidth on the number of significant signals. Figure 2.3 shows how the proportion of signals significant at 5% (red), 1% (light blue), and 0.1% (dark blue) level evolves depending on the number of lags and estimator of covariance matrix. We consider only equal-weighted portfolios of data-mined signals and anomalies here as equal-weighting is an overwhelming choice in the literature (McLean and Pontiff (2016)). The performance of anomalies is adjusted for five Fama-French factors. The horizontal lines correspond to Newey and West (1987) estimator with critical values from normal distribution. We also provide results for three specifications of naive bootstrap. Critical values for alphas without any standardization are depicted with squares. The upward-facing triangles show naive bootstrap for t-statistic with standard errors without any adjustment for heteroskedasticity or autocorrelation. The circles then describe proportion of significant signals with Newey and West (1987)

[24]See, for example, Romano and Wolf (2006) who provide rule for automatic selection of block length.

HAC robust estimator and bootstrapped critical values.

One apparent feature is that a larger number of lags leads to fewer significant signals. The number generally drops uniformly for the first 24 lags. The number of significant signals levels off after 24 lags and reaches the minimum at about 60. It starts increasing after that because the block length starts causing problems with randomness of the sample. The increase is mainly in tail of the distribution. The decline in the number of significant signals can be substantial. 10% of all signals are significant at 0.1% level with Newey and West (1987) adjustment and critical values from normal distribution. This drops to 0.3% for bootstrapped critical values and 24 lags. Notably, the critical values from normal distribution seem to over-reject null hypothesis for any number of lags. This is in line with evidence in the previous subsection. The green vertical lines correspond to mean number of lags selected by Andrews and Monahan (1992) and Newey and West (1994). The optimal number of lags is 5.5 and 10.4, respectively.[25] The proportion of significant signals is then intercept of these vertical lines and lines for Newey and West (1987) adjustment.

There is also a large discrepancy between different versions of the bootstrap. The simplest version, without studentization, tends to reject the most signals possibly due to distorted size of the tests.[26] It can also get heavily distorted with the existence of large outliers. Studentization should improve properties of the bootstrap and its importance is emphasized in Davison and Hall (1993), Götze et al. (1996), and Romano and Wolf (2006). The drop in number of significant signals with the number of lags in HAC robust adjustment does not have to imply improper size of tests with the small number of the lags. This is because there is a trade-off between type I and type II error rate. Type I error rate decreases with the number of lags but type II rate increases. The test then has poor power and rejects fewer truly significant signals. Bootstrap without HAC robust adjustment should capture role of block size in the bootstrap since that is the only thing that is changing. This should in turn capture the effect of autocorrelation on standard errors without any distortion in power as in the case of Newey and West (1987) adjustment. The optimal number of lags for naive bootstrap with HAC adjustment therefore appears to be around six where number of rejected hypothesis is similar to minimum number of rejected hypothesis without the HAC adjustment. Larger number of lags then probably leads to poor power of the tests. We also provide proportion of signals significant with Kiefer et al. (2000) estimator denoted by inverse triangles. It is higher than for naive bootstrap with Newey and West (1987) adjustment and 24 lags. This hints that the optimal number of lags is lower than that. We will next turn to simulations to study power and size of the tests in a controlled environment.

[25]Note that the number of lags is lower for value-weighted portfolios at 3 and 9 lags, respectively.

[26]Shao and Politis (2013) showed that this version of bootstrap does not have a normal distribution in finite sample analogously to fixed-b asymptotics for HAC errors.

2.3.3 Simulations

Figure 2.3 showed that adjustment of standard errors on portfolio returns can have a large impact on proportion of significant signals. It is, however, not obvious if the change in proportion of significant signals is due to decreasing power the tests or improper size of the tests for the small number of lags in HAC adjustment. This section proposes a simulation framework that inherits all the properties of the empirical data and enables us to study the size and power of the tests in the controlled environment.

Section 2.1.2 has described construction of a large universe of fundamental signals. The fundamental signals are of two types: noise signals that are not connected to any excess return on stocks and true predictive signals that are connected to excess return on stocks. The multiple hypothesis test is trying to distinguish between the two types of the signals. It is easy to generate any number of the noise signals since randomly generated fundamental signals are not connected to any excess returns by definition. In particular, we randomly draw simulated fundamental signals from uniform distribution for each company-year (GVKEY-year in Compustat). The simulated fundamental signals are updated annually at the end of June and portfolios are constructed from them in exactly the same way as for the data-mined fundamental signals as described in Section 2.1.4. The noise signals share the same properties as the data-mined fundamental signals due to having the same construction. The simulations focus solely on the US where the quality of the data is the highest.[27]

The power and size of the statistical tests can be studied on these simulated signals. The proportion of significant signals should be the same as size of the individual tests if the tests have correct size. The multiple hypothesis tests should therefore lead to no significant signal if there is truly no significant signal. It is also possible to test power of the tests by adding positive monthly return to the portfolios created based on the simulated signals. The correct functioning of Storey (2002) test requires that some noise signals remain. The monthly return is therefore added only to 20% of the simulated signals when power and size of the multiple hypothesis test is studied. The empirical false discovery rate of multiple hypothesis test can be estimated by increasing the monthly returns up to a point where all of the 20% of truly significant signals are rejected. The simulated fundamental signals are unrelated to the risk factors by definition. The added annual return is therefore equal to excess return after adjusting for the risk factors. The simulated signals are also useful to assess whether number of the risk factors has any role on power and size of the statistical tests.

Figure 2.4 plots proportion of significant simulated signals as a function of true annual returns on the portfolios. The upper two subplots show the proportions for individual tests of significance at 5% confidence level based on bootstrapped p-values. The lower two subplots depict the proportions based on multiple hypothesis test (MHT) of Storey

[27]Note that the authors have also experimented with generating correlated signals but there was no impact on the overall results. Storey (2002) method is therefore robust to correlation among the signals.

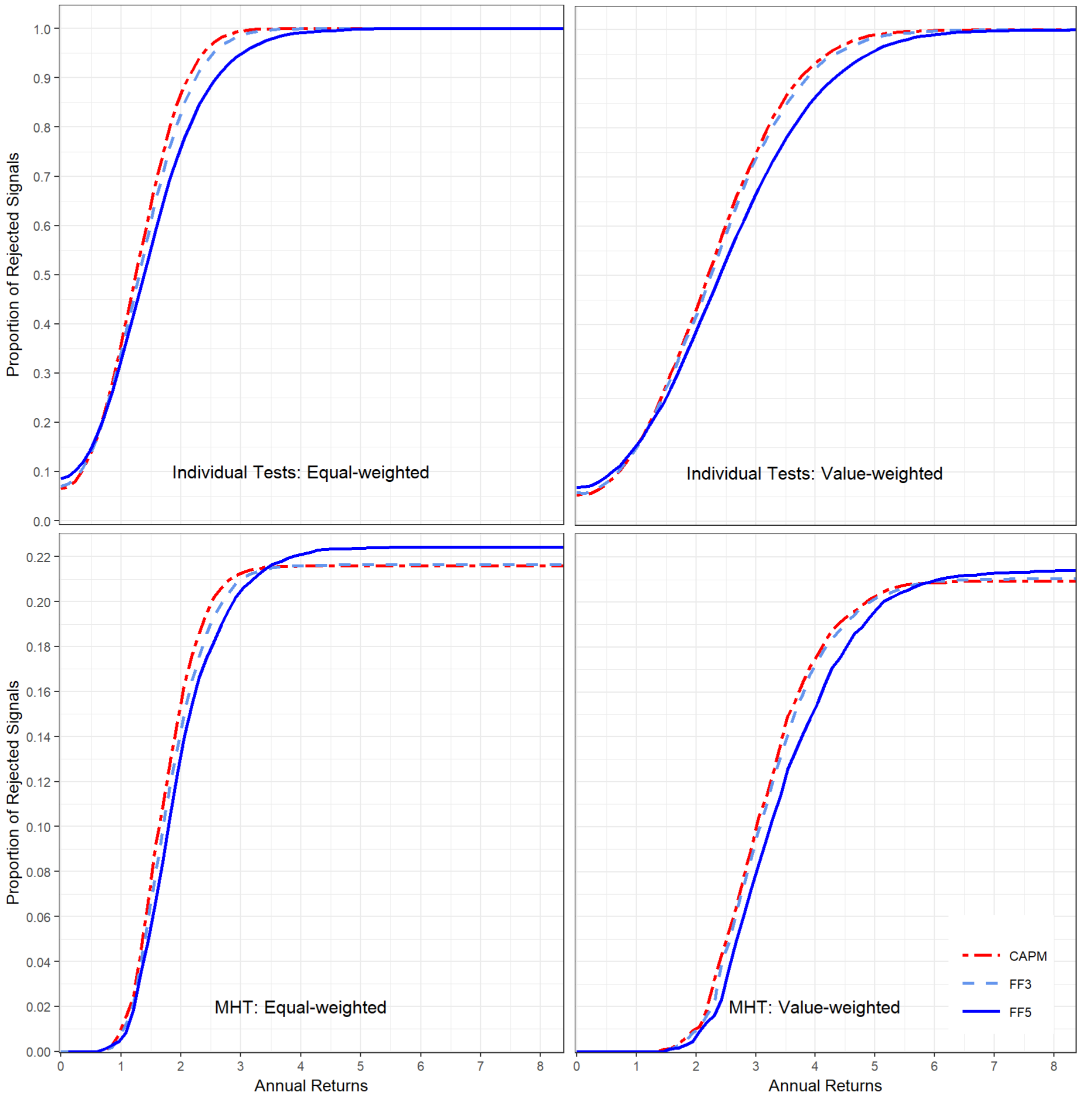


Figure 2.4: **Proportion of Significant Simulated Signals Depending on Strength of the Signals.** The figure shows the proportion of significant signals as a function of true annual returns on 10,000 simulated annual fundamental signals. The signals are deemed significant based on their bootstrapped p-values using 10,000 runs of the block bootstrap with block length 3 and standard errors adjusted for autocorrelation and heteroskedasticity as in Newey and West (1987) with 3 lags. The significance is determined either in individual tests at 5% significance level or multiple hypothesis tests (MHT) of Storey (2002) at 5% false discovery rate. The significance of individual signals is determined via regressions adjusting the portfolio returns for CAPM, FF3, or FF5 factors. The fundamental signals are randomly drawn from uniform distribution for all company-years (GVKEY in Compustat). The company-years are matched to market data from July to June each year so that the signal for each company changes at the beginning of each July and remains constant for the next 12 months. The equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the random signals and shorting stocks in the bottom decile of the signals. The annual return on the signals is simulated by adding constant monthly return to all the portfolios for individual tests and 20% of portfolios for MHT tests. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016.

(2002) at 5% false discovery rate. The bootstrap uses block length 3 and standard errors on the returns on the portfolio are estimated with Newey-West adjustment with 3 lags. The Figure 2.4 is supported with tabulated numbers in Panel A of Table 2.5. The table also contains estimated proportion of signals with zero true return $\pi = \pi_0/M$ in Storey (2002) test. The proportion of significant signals is provided for portfolio returns adjusted for CAPM, FF3, and FF5 risk factors.

The individual tests for significance in the upper two subplots in Figure 2.4 have visibly distorted size for FF5 risk factors. The proportion of rejected signals is close to 9% even through the size of the test is 5% and the true annual return is zero. The distorted size is also documented in Table 2.5 for both 5% and 1% intended significance levels. The FF5 risk factors also lead to noticeably lower power relative to CAPM. There is also a slight decrease in power for FF3 risk factors relative to CAPM. The power of the individual tests is lower for value-weighted portfolios relative to equal-weighted portfolios.

The results for multiple hypothesis tests (MHT) closely follow results for the individual signals. The bootstrapped p-values of individual signals are the basis of Storey (2002) test and their incorrect distribution translates into incorrect false discovery rate. Figure 2.4 documents that the MHT tests don't reject any signal when there are no outperforming strategies and reject all the truly outperforming strategies for sufficiently high annual returns. The power of the tests is higher for equal-weighted returns and 3% annual return is enough for the signal to be rejected with certainty at 5% false discovery rate. The difference in power between equal-weighted and value-weighted portfolios explain the previous results where the proportion of significant fundamental signals was much lower for value-weighted portfolios. The power also decreases with number of risk factors so that FF5 factors lead to the weakest power. The drop in proportion of significant signals with number of the risk factors documented in Figure 2.3 is therefore partly due to the drop in power.

There is 20% of simulated signals with truly positive annual returns which means that there should be about 21.05% of rejected signals with the desired 5% false discovery rate. The proportion of rejected signals is higher than that for FF5 in Figure 2.4 when the annual returns are larger than 4%. This is further supported by Panel A in Table 2.5. The bootstrapped false discovery rate for FF5 and equal-weighted returns is close to 11% which is far from the desired 5% rate. Estimated proportion of simulated strategies satisfying the null hypothesis of zero return is also much lower than its true value of 80%.

Panel A of Table 2.5 has shown that the proportion of rejected signals under the null of the individual tests does not correspond to their desired size. Panel B then studies impact of increasing block length of the bootstrap and lags in Newey-West adjustment of standard errors to 12. Panel B is supplemented with Figure 2.5 which plots the proportion of rejected signals depending on annual returns on the portfolios for both 3 and 12 lags in the adjustment. The figure also plots proportion of significant signals with 3 lags when critical threshold for the bootstrapped p-values is chosen to yield exactly 5% of significant

Table 2.5:
Simulations of Size of the Tests

The table shows bootstrapped size of statistical tests on 10,000 simulated annual fundamental signals. The signals are deemed significant based on their bootstrapped p-values using 10,000 runs of the block bootstrap in Panels A, B, and C. Block length of the bootstrap is 3 in Panel A and C while it is 12 in Panel B. p-values in Panels D and E are based on critical values of normal distribution. Standard errors adjusted for autocorrelation and heteroskedasticity as in Newey and West (1987) with either 3 lags in Panel A and D or 12 lags in Panel B, C, and E. The bootstrapped size of individual tests is defined as proportion of significant signals in individual tests where the bootstrapped p-values are lower than 1% or 5% threshold. The bootstrapped size of multiple hypothesis test of Storey (2002) at either 1% or 5% false discovery rate is defined as proportion of false positive discoveries among all the rejected signals. There are 20% true positive discoveries among all the signals for multiple hypothesis tests. The table also shows estimated proportion of signals with zero returns π. The significance of individual signals is determined via regressions adjusting the portfolio returns for CAPM, FF3, or FF5 factors. The fundamental signals are randomly drawn from uniform distribution for all company-years (GVKEY in Compustat). The company-years are matched to market data from July to June each year so that the signal for each company changes at the beginning of each July and remains constant for the next 12 months. The equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the random signals and shorting stocks in the bottom decile of the signals. The sample is restricted to industrial stocks with price over \$1 and capitalization larger than bottom decile in NYSE at the end of previous June. The sample spans July 1963 to December 2016.

	Monthly Rebalanced Portfolios						Annually Rebalanced Portfolios					
	Equal-weighted			Value-weighted			Equal-weighted			Value-weighted		
	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5
Panel A: Bootstrapped p-values with 3 Block Length and 3 Lags in HAC Adjustment												
Single Hypothesis Tests												
1%	1.25	1.35	1.45	0.82	0.96	1.44	1.46	1.55	1.89	1.01	1.11	1.47
5%	5.32	5.68	6.19	4.84	4.88	6.51	6.54	6.99	8.52	5.38	5.88	6.94
Multiple Hypothesis Tests												
1%	1.28	1.43	1.38	0.89	1.14	1.67	1.57	2.10	2.68	0.84	1.14	1.14
5%	5.88	6.15	7.11	4.85	5.03	7.66	7.36	7.54	10.8	4.44	4.90	6.59
π	79.0	77.6	76.1	80.6	79.3	74.5	75.3	73.0	66.4	78.0	77.8	74.7
Panel B: Bootstrapped p-values with 12 Block Length and 12 Lags in HAC Adjustment												
Single Hypothesis Tests												
1%	1.21	1.27	1.19	0.95	1.13	1.31	1.15	1.03	1.07	0.99	1.12	1.25
5%	5.42	5.43	5.71	5.09	5.33	5.86	5.19	5.02	5.72	4.74	5.15	5.73
Multiple Hypothesis Tests												
1%	1.33	1.48	1.57	0.79	0.94	0.94	0.84	0.89	1.04	0.89	0.94	0.99
5%	5.48	5.53	6.89	4.53	4.49	5.12	4.53	4.67	5.79	5.30	4.81	5.88
π	81.0	78.8	75.7	80.0	79.0	76.6	79.5	79.2	70.5	80.5	78.6	74.5

signals under the null hypothesis. Increasing the number of lags limits and block length the distortions in size of the individual tests. It also limits distortions in false discovery rate in the multiple hypothesis setting. Figure 2.5 documents that power of the individual tests decreases for 12 lags but it is nonetheless higher than for proper critical threshold for bootstrapped p-values with 3 lags. The larger number of lags therefore decreases number

Table 2.5 Continued

	Monthly Rebalanced Portfolios						Annually Rebalanced Portfolios					
	Equal-weighted			Value-weighted			Equal-weighted			Value-weighted		
	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5
Panel C: Bootstrapped p-values with 3 Block Length and 12 Lags in HAC Adjustment												
Single Hypothesis Tests												
1%	1.20	1.30	1.40	0.96	1.30	1.40	1.31	1.33	1.27	1.23	1.20	1.39
5%	5.92	6.09	6.36	5.41	5.58	6.13	5.70	5.80	6.40	5.47	5.42	6.07
Multiple Hypothesis Tests												
1%	1.28	1.38	1.62	0.89	1.04	0.89	1.23	1.23	1.57	1.04	0.79	1.67
5%	5.70	6.67	6.85	4.90	4.67	6.37	6.50	6.15	7.66	5.39	5.57	6.89
π	78.4	77.8	76.4	76.0	78.4	74.4	78.0	75.6	68.0	76.7	77.1	72.8
Panel D: 3 Lags in HAC Adjustment without Bootstrap												
Single Hypothesis Tests												
1%	1.21	1.42	1.78	0.97	0.89	1.63	1.39	1.66	2.80	1.11	1.20	1.79
5%	5.75	6.71	7.07	5.05	5.03	7.22	6.22	7.49	10.1	5.47	5.74	7.67
Multiple Hypothesis Tests												
1%	1.48	1.43	2.34	0.74	1.38	2.53	1.19	2.01	3.94	0.89	1.53	3.05
5%	6.98	7.83	8.63	6.02	6.28	10.4	7.36	9.34	16.0	5.93	6.76	9.62
π	76.6	73.7	72.5	77.0	76.8	72.4	76.7	73.5	65.7	77.6	74.6	69.5
Panel E: 12 Lags in HAC Adjustment without Bootstrap												
Single Hypothesis Tests												
1%	1.36	1.43	1.87	1.21	1.48	2.07	1.16	1.29	2.40	1.34	1.48	1.79
5%	5.99	6.23	6.79	5.53	6.05	7.82	5.74	6.16	9.22	5.81	5.95	7.79
Multiple Hypothesis Tests												
1%	1.96	1.91	2.10	1.23	1.57	2.53	1.43	1.72	3.71	0.94	1.04	2.58
5%	6.93	8.05	8.80	5.79	6.76	9.91	7.32	8.38	13.3	4.94	5.84	9.26
π	77.8	76.0	73.8	76.6	76.6	73.4	77.4	74.6	65.8	77.6	76.2	71.0

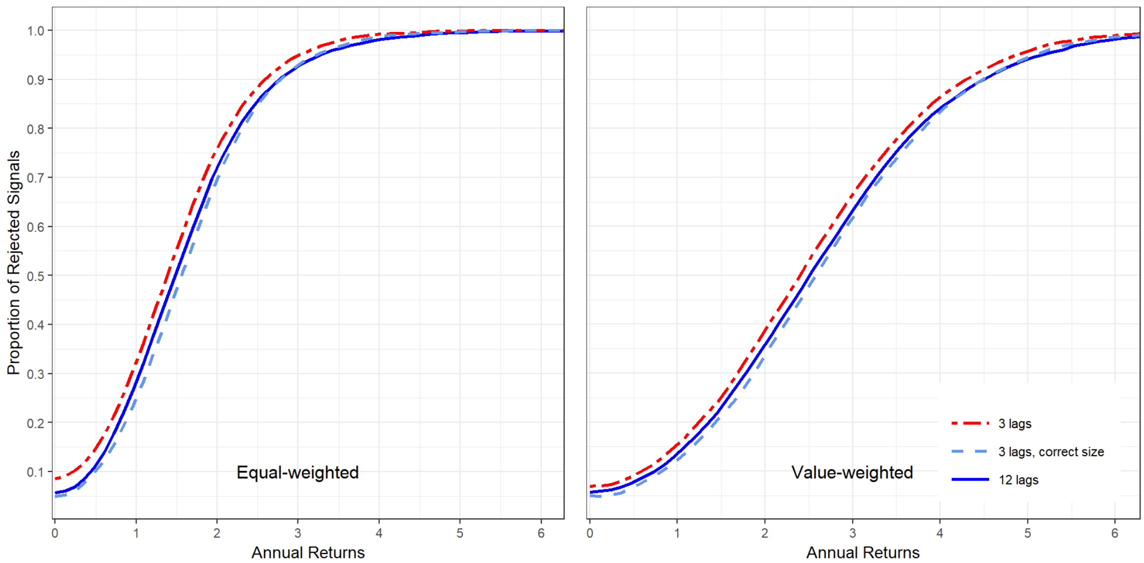


Figure 2.5: **Proportion of Significant Simulated Signals Depending on Strength of the Signals.** The figure shows the proportion of significant signals as a function of true annual returns on 10,000 simulated annual fundamental signals. The signals are deemed significant based on their bootstrapped p-values using 10,000 runs of the block bootstrap with block length 3 (or 12) and standard errors adjusted for autocorrelation and heteroskedasticity as in Newey and West (1987) with 3 (or 12) lags. The significance of individual signals is determined in individual tests at 5% significance level. The significance is determined via regressions adjusting the portfolio returns for FF5 factors. The line denoted "3 lags, correct size" uses critical threshold for the bootstrapped p-values that leads to exactly 5% significant signals when the excess return is equal to zero. The fundamental signals are randomly drawn from uniform distribution for all company-years (GVKEY in Compustat). The company-years are matched to market data from July to June each year so that the signal for each company changes at the beginning of each July and remains constant for the next 12 months. The equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the random signals and shorting stocks in the bottom decile of the signals. The annual returns on the signals are simulated by adding constant monthly return to all the portfolios. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016.

of signals mainly by correcting the size of the individual tests and not by decreasing power of the tests.

It is not clear from the previous evidence whether (a) increasing block length in the bootstrap or (b) increasing number of lags in auto-correlation and heteroskedasticity adjustment or (c) doing both (a) and (b) jointly is behind the desired correction of size of the tests. Panel C in Table 2.5 is a middle step between Panel A and B as it increases only the block length to 12 while keeping the number of lags in adjustment of the standard errors for auto-correlation at 3. It is apparent that the size of the tests has improved significantly relative to Panel A with block length 3. The block length in the bootstrap is therefore responsible for a large part of the correction in the size of the tests. Increasing the number of lags in the standard error adjustment in Panel B, however, has significant impact for the false discovery rate in the multiple hypothesis tests.

Panels D and E in Table 2.5 investigate importance of the bootstrap in estimation of p-values on the individual portfolios. p-values in Panels D and E do not rely on bootstrap but are derived from critical values of standard normal distribution, as is mostly done in the literature. The normal distribution is a limiting distribution of t-distribution as the sample increases to infinity and is a good approximation of the t-distribution of the t-statistics with larger sample size. Panel D relies on adjustment of standard errors for heteroskedasticity and auto-correlation for up to 3 lags while Panel E for up to 12 lags. It is evident that the bootstrapped size of the individual tests and false discovery rate of the multiple hypothesis tests are far from their desired values even for the 12 lags. The adjustment of standard errors for auto-correlation and heteroskedasticity is therefore alone not sufficient to provide correct size of the tests. The conclusion derived from this section therefore is that not only the number of lags in the adjustment of the standard error is important for the correct size of the tests but the p-values also need to be bootstrapped with a suitable method.

2.3.4 Auto-correlation or Conditional Heteroskedasticity?

Table 2.6:
Tests of Presence of Auto-correlation and Heteroskedasticity on the Simulated Signals

The table shows proportion of simulated signals created as in Table 2.5 for which null hypothesis of no auto-correlation or no auto-regressive heteroskedasticity of up to 12 lags is rejected. The proportion of signals is provided in percentage points for 1% and 5% significance levels. Presence of no auto-correlation is tested in Ljung-Box test while presence of no autoregressive conditional heteroscedasticity (ARCH) is tested in ARCH LM test and McLeod and Li (1983) test.

	Monthly Rebalanced Portfolios						Annually Rebalanced Portfolios					
	Equal-weighted			Value-weighted			Equal-weighted			Value-weighted		
	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5	CAPM	FF3	FF5
Tests of Presence of Auto-correlation and Heteroskedasticity												
Ljung-Box Test for Auto-correlation of up to 12 Lags												
1%	4.35	3.93	3.66	20.5	18.9	17.3	20.6	15.5	12.7	23.4	21.2	19.5
5%	13.4	12.7	12.4	35.8	33.9	31.9	33.1	27.5	24.4	38.6	36.1	34.0
ARCH LM Test for Auto-regressive Heteroskedasticity of up to 12 Lags												
1%	57.4	52.2	47.6	98.1	97.7	96.9	80.7	75.9	70.8	96.7	96.0	95.0
5%	71.0	67.1	62.6	99.2	99.0	98.7	86.4	83.6	79.7	98.4	98.0	97.4
McLeod-Li Test for Auto-regressive Heteroskedasticity of up to 12 Lags												
1%	65.7	61.4	56.1	99.2	98.9	98.5	83.3	79.3	74.7	98.2	97.8	97.2
5%	76.2	72.3	68.3	99.6	99.4	99.2	87.8	85.1	81.8	99.0	98.7	98.4

The importance of block size in the bootstrap hints that auto-correlation in returns not the sole driver of inappropriate critical values in the t-statistic. Table 2.6 provides further support for this claim. It provides proportion of signals, simulated in the same way as in Table 2.5, where null hypothesis of no auto-correlation of up to 12 lags in residual portfolio returns after adjusting them for the risk factors is rejected at either 1% or 5%.

The test rejects auto-correlation far more frequently than what would be expected by pure chance but not all the signals exhibit significant auto-correlation. The auto-correlation is therefore only weak among the generated signals. Table 2.6 also presents proportion of cases where null hypothesis of no autoregressive heteroskedasticity in residual portfolio returns after adjusting them for the risk factors is rejected at either 1% or 5%. The autoregressive heteroskedasticity is tested with autoregressive conditional heteroscedasticity (ARCH) Lagrange multiplier (LM) test and McLeod and Li (1983) test. Almost all the signals exhibit ARCH effects which sheds some light on why increasing the block length was so important for the block bootstrap. Short block length does not allow for the bootstrap to generate sample with the significant ARCH effects as in the original series.

The autoregressive heteroskedasticity can be explained by shifting leverage over the year for the annually rebalanced portfolios. The reason for this is that the annual rebalancing induces dependence in portfolio returns over the year. The dependence is induced by the fact that returns in a given month are influenced by cumulative returns since the last portfolio rebalancing. Suppose that the market return is 50% over July-December period, the volatility of payoffs from the given long-short strategy should then also be 50% higher over the remaining 6 months before the portfolio is rebalanced. Similar effects emerge when value of either short or long legs of the portfolio increases in value. The time-dependence of portfolio returns is then manifesting itself though the documented ARCH effects when leverage of the portfolios significantly drifts away from its baseline value in months following the annual rebalancing period. Larger block length in the bootstrap is therefore required to simulate the dependence in the data.

The analysis in this paper has so far focused on annually rebalanced zero-cost long-short strategies. Table 2.5 also shows bootstrapped size and false discovery rate for monthly rebalanced strategies. Monthly rebalanced strategies generally don't suffer from the large distortions in size and false discovery rate. These results further support that the annual rebalancing is the driving force behind the problems related to inappropriate critical values in the t-statistic.

2.3.5 Impact on Proportion of Significant Signals in Multiple Hypothesis Tests

We study the impact of unconditionally non-pivotal nature of t-statistics on the multiple hypothesis tests in Table 2.7. The table is mostly constructed in the same way as Table 2.3. The only difference in Panel A with respect to Table 2.3 is that p-values are based on a block bootstrap with block length of 12 months and Newey-West adjustment of standard errors of 12 months instead of 3 previously. Panel B and C rely on the unchanged number of lags but the portfolios are rebalanced monthly instead of annually. Both of the changes should limit the incorrect size of the multiple hypothesis tests.

The impact of increasing the number of lags in Newey-West adjustment in Panel A depends on risk factors used in estimation of alphas on the portfolio returns. It was

Table 2.7:

Multiple Hypothesis Tests - The Impact of HAC Adjustment

The table shows proportion of significant fundamental signals (N) in percentage points and corresponding critical p-values (p-val) under Storey (2002) multiple hypothesis framework controlling for 5% FDR. The table corresponds to Panel A in Table 2.3. The only exception is that the standard errors are adjusted as in Newey and West (1987) with 12 lags in Panel A (otherwise 3 lags) and the portfolios are rebalanced monthly in Panel B and C (otherwise annually). Risk factors in Panel C are constructed in the same way as the portfolios of the fundamental signals and follow Fama and French (2015) otherwise.

	Equal-weighted Portfolios									Value-weighted Portfolios								
	CAPM			FF3			FF5			CAPM			FF3			FF5		
	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val	N	N_A	p-val
Panel A: 12 Lags in HAC Adjustment																		
USA 1963-2016																		
Reduced	40	58	5.0	22	31	2.3	5.7	14	0.5	9.0	22	0.7	6.5	14	0.6	0.7	2.2	0.1
Base	32	54	3.4	15	26	1.4	3.2	7.5	0.3	6.6	19	0.5	4.3	13	0.4	0.0	0.0	0.0
Extended	33	56	3.7	17	26	1.4	5.7	14	0.6	1.1	7.5	0.1	0.9	5.4	0.1	0.0	0.0	0.0
Japan 1990-2016																		
Reduced	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	15	26	1.4	18	27	1.7	0.9	3.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	13	25	1.2	13	25	1.2	1.2	3.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	8.2	23	0.7	8.0	20	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	14	25	1.3	13	17	1.3	1.0	1.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	13	24	1.1	10	15	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	2.4	9.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Panel B: Monthly Rebalanced Portfolios																		
USA 1963-2016																		
Reduced	48	68	6.3	48	68	6.1	35	54	3.9	17	34	1.5	22	38	2.2	18	28	1.6
Base	43	67	4.9	45	67	5.6	32	53	3.3	12	31	1.0	17	34	1.6	18	32	1.5
Extended	42	66	5.1	38	65	4.3	37	54	4.4	8.9	29	0.7	11	24	0.9	9.4	22	0.7
Japan 1990-2016																		
Reduced	1.4	5.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	36	55	3.9	42	60	4.9	8.0	24	0.4	0.8	2.2	0.1	1.2	4.3	0.1	0.0	0.0	0.0
Base	33	54	3.4	38	60	4.3	9.9	24	0.7	1.4	5.4	0.1	3.1	6.5	0.2	0.9	1.1	0.1
Extended	23	47	2.1	23	54	2.1	2.8	16	0.2	0.0	0.0	0.0	1.1	4.3	0.1	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	25	33	2.2	22	30	2.3	2.5	1.1	0.2	1.6	6.5	0.1	2.0	4.3	0.2	0.0	0.0	0.0
Base	22	30	1.9	20	29	2.1	2.3	1.1	0.2	0.0	0.0	0.0	1.0	2.2	0.1	0.0	0.0	0.0
Extended	13	28	1.1	4.1	13	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Panel C: Monthly Rebalanced Portfolios with Equivalently Constructed Risk Factors																		
USA 1963-2016																		
Reduced	49	69	6.6	42	65	5.2	18	39	1.7	16	34	1.4	19	40	1.8	5.4	18	0.5
Base	43	66	4.9	41	62	4.8	20	41	2.0	12	32	1.0	16	39	1.4	3.3	14	0.3
Extended	42	67	5.1	40	62	4.8	19	40	1.8	8.6	29	0.7	9.4	29	0.8	1.1	8.6	0.1
Japan 1990-2016																		
Reduced	0.8	4.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Base	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Europe 1990-2016																		
Reduced	36	55	3.9	35	55	4.2	14	35	1.2	0.8	2.2	0.1	2.7	5.4	0.2	0.0	0.0	0.0
Base	32	54	3.3	31	52	3.6	19	39	1.7	1.3	3.2	0.1	3.0	5.4	0.2	0.0	0.0	0.0
Extended	24	48	2.2	16	44	1.4	5.2	32	0.4	0.0	0.0	0.0	1.1	3.2	0.1	0.0	0.0	0.0
Asia Pacific 1990-2016																		
Reduced	23	30	2.0	23	31	2.2	0.0	0.0	0.0	1.3	4.3	0.1	2.4	5.4	0.2	0.0	0.0	0.0
Base	24	32	2.0	22	30	2.0	0.0	0.0	0.0	0.6	3.2	0.1	2.5	7.5	0.2	0.0	0.0	0.0
Extended	13	26	1.1	13	24	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

previously shown with simulations that problems with size of the single hypothesis tests are the most severe for FF5 but almost nonexistent for CAPM. The results in Panel A support this evidence. There are only slight changes in proportion of significant signals with respect to Table 2.3 for CAPM. The proportion of significant signals and critical values for p-values, however, drop to about one third for FF5. The drop is partly caused by correcting the size of the tests but it is also partly caused by a drop in power of the tests. The larger number of lags therefore carries some costs.

The monthly rebalancing in Panel B also shrinks problems with incorrect size of the tests. The proportion of significant signals is generally larger than in Table 2.3. The main reason for the increase in significant signals is that the risk factors are no longer properly controlling for their underlying fundamental risk as they are constructed in a different way than the portfolios. Panel C changes the risk factors' construction so that it is the same as construction for the fundamental signals. The new construction of the risk factors reduces the proportion of significant signals to about one half for FF5 and equal-weighted portfolios and even more for value-weighted portfolios. The portfolio setting is therefore intrinsically connected to many issues and the regression framework introduced in the next section is suitable to overcome them.

2.4 Multiple Hypothesis Tests: Independent Signals

All the multiple hypothesis tests covered so far correctly select signals in the individual tests of significance but they say nothing about which signals are marginally useful in explaining the future returns in relation to the other signals. It could be the case that the multiple hypothesis tests will give us hundreds of signals that are closely related to each other and most of the signals are then a linear combination of the others. The multiple hypothesis tests based on portfolios are unfit to select the signals that each provide new information. Therefore, it is necessary to look for other methods.

Lewellen et al. (2015) and Green et al. (2017) used Fama and MacBeth (1973b) regressions of individual stock returns on their past characteristics to select the independently significant signals. The regressions test whether the characteristics are related to cross-section of returns. The characteristics are normalized to have values between zero and one by considering quantiles of the given characteristic among all the stocks in the given region and month. There are some problems with this approach including multicollinearity and a need to adjust resulting p-values for multiple hypothesis setting. Green et al. (2017) try to overcome the multiple hypothesis problems by adjusting the p-values for FDR as in Benjamini and Yekutieli (2001) and the multicollinearity problems by discarding signals with the high variance inflation factors (VIF). We follow this approach and discard collinear signals with VIF larger than either 3 or 7. Having the universe of potential signals also allows us to use Storey (2002) multiple hypothesis test with a better power to reject truly significant signals relative to Benjamini and Yekutieli (2001).

In essence, we regress returns on the individual stocks over 12 months starting in July in calendar year t on M rescaled fundamental signals from business year ending in calendar year $t-1$

$$r_{i,t} = \beta_0 + \sum_{j=1}^{M} \beta_j x_{i,j,t-1} + \epsilon_{i,t}. \tag{2.4}$$

Where $x_{i,j}$ describes a transformed fundamental signal j for a given stock i. All the raw fundamental signals are normalized to between -0.5 and 0.5 by transforming them into empirical quantiles within each region-year and subtracting 0.5. The missing values are filled with zeros. We exclude 4 binary signals from the set of published anomalies. The regressions are fitted on annual data rather than on monthly data as was done in Green et al. (2017), since all signals under the study are updated only annually. The regressions pool together all the available stocks for all the time periods and the whole cross-sections in a given region. It is not possible to run Fama-MacBeth regressions as in Green et al. (2017) because the number of characteristics is sometimes larger than the number of stocks in the cross-section. The goal of the regressions is to select fundamental signals x_j that are statistically significant. Selection of the method to estimate the standard errors therefore plays some role. Heteroskedasticity robust standard errors are always used.[28]

Our selection problem is more complicated than in the existing studies as we have many signals. Consequently, the number of signals is larger than the number of stocks in cross-section in some regions. Discarding signals based on VIF can lead to a loss of signals that are closely correlated to some other signals but have incremental predictive power over the other signals. Therefore, we also adopt LASSO of Tibshirani (1996) with L1 penalty to tackle the multidimensionality problems without discarding any signals before running the regressions.[29]

Inference on the signals selected by LASSO is very problematic since coefficients from LASSO are heavily biased. Lee et al. (2016) explain the problem: "*For example, one common approach when the number of variables is not too large is to fit a linear model with all variables included, observe which ones are significant at level α, and then refit the linear model with only those variables included. The problem with this is that the p-values can no longer be trusted, since the variables that are selected will tend to be those that are significant. Intuitively, we are "over-fitting" to a particular realization of the data.*"[30] The whole argument boils down to the same reasoning as in the case of missing unpublished anomalies in the multiple hypothesis framework in that the confidence intervals are not

[28]We have tried several adjustments of the standard errors including HAC robust and clustered on time but the choice makes little difference for the conclusions. There appears to be no problem with autocorrelation here given that both signals and returns are at annual frequency.

[29]LASSO equally penalizes all coefficient in least square minimization problem and thus shrinks most of them to zero and thus selects the most important variables. LASSO does not have oracle property but we prefer it over adaptive LASSO as it requires fewer specified parameters. The results from adaptive lasso are almost identical so this does not have any impact on our findings. Freyberger et al. (2017) proposed to use non-parametric approach with additive models but we have found little improvement from adopting it. This is possibly due to larger number of signals under study here. Linear models are thus good enough approximation and seem to capture most of the predictability.

[30]Another good explanation is provided in Berk et al. (2013).

conservative enough if we do not account for all tried signals. There are now many available methods to adjust the confidence intervals for proper post-selection inference. P-values from LASSO are adjusted with method suggested in Tibshirani et al. (2016) and then Benjamini and Yekutieli (2001) FDR correction is applied on the adjusted p-values.[31]

2.4.1 Simulation Evidence

We start with the simulations to compare the various methods. The financial setting for regression methods is very challenging and it is therefore valuable to study the methods in a controlled environment before we move to the empirical setting. There are two main sources of difficulties. First, the predictability of returns on individual stocks is extremely small and out-of-sample (OOS) R^2 is mostly below 1%. Next, there are some cases of large multicollinearity between the signals due to their generation process. Both, multicollinearity and high noise-to-signal ratio, can lead to a complete breakdown of the methods.

We simulate a setting that is close to the empirical data in the US. We simulate 100,000 observations of annual returns on individual stocks that are log normal with zero mean and are driven by K signals

$$r_i = \exp\left(-\sigma^2/2 + 0.03\sum_{j=1}^{K} x_{ij} + \epsilon_i\right) - 1, \tag{2.5}$$

where $K \in \{10, 50\}$, $\epsilon \sim N(0, \sigma)$, and $\sigma = 0.4$. There are 1500 simulated signals in total; K of which are true drivers of returns, 600 - K are a mix of true drivers and noise, and 900 are pure noise. All the signals are uniformly distributed on (-.5, .5) interval. The 600 - K mixed signals are generated by randomly drawing three true drivers of returns and then randomly mixing them together with noise:

$$x_j = \rho(\bar{a}x_a + \bar{b}x_b + \bar{c}x_c)/(\bar{a} + \bar{b} + \bar{c}) + (1-\rho)\gamma, \tag{2.6}$$

where γ is uniform over (-0.5, 0.5), $\bar{a}$, $\bar{b}$, $\bar{c}$ are uniform on (0,1), and ρ is uniform on either (0, 0.5) or (0, 0.75). The mixed signals are then transformed to empirical quantiles minus 0.5 so that they have the same distribution as the other signals. ρ smaller than 0.75 guarantees that there are no extreme problems with collinearity. Two settings are considered; with few true drivers of the returns ($K = 10$) and many true drivers of the returns ($K = 50$). Any time dependence issues are ignored and the sample is iid over time and individual stocks to simplify the problem.

We first generate additional 20,000 observations to test how successful the OLS and the LASSO predictions are in our simulations in terms of OOS R^2. The maximum feasible OOS R^2 is about 1.82% for setting with 50 true independent signals and 0.37% for 10 signals. The maximum can only be obtained when all the truly independently significant signals are used with their true coefficients. LASSO leads to about 90% of the maximum feasible OOS R^2 while OLS leads to negative OOS R^2 in the case with 10 truly significant

[31] Storey (2002) method is not feasible here as the number of selected signals is often small.

signals and less than 1% with 50 truly significant signals. The ability of the methods to cope with the high noise-to-signal ratio deteriorates rapidly when all the true coefficients are shrank from 0.03 to 0.02, which is more consistent with empirical data in the US. R^2 from LASSO drops to about 50-70% of the maximum feasible OOS R^2 while OLS leads to negative OOS R^2 everywhere.

Table 2.8:
Simulations for Independently Significant Signals
The table shows number of significant coefficients and corresponding simulated false discovery rate (FDR) in brackets for regressions of simulated returns on 1,500 signals. We present average values from 100 runs. LASSO (no FDR) counts all signals selected from the LASSO that are significant at 5% level in a further regression (OLS) or that are significant with proper post-selection inference (P-S Inf). We then control for 5% FDR in the regressions of the return on all the signals (OLS) or again in post-selection setting (LASSO: OLS and LASSO: P-S Inf). We control for FDR with either Benjamini and Yekutieli (2001) (BY) or Storey (2002) (STO) method. Port STO stands for number of signals that are significant in decile long-short portfolio setting with STO adjustment at 5% FDR. The penalty in the LASSO is selected with three-fold cross-validation.

				5% FDR					
	Port	LASSO: (no FDR)		OLS		LASSO: OLS		LASSO: P-S Inf	
K	STO	OLS	P-S Inf	STO	BY	STO	BY	STO	BY
					$\rho = \mathbf{0.5}$				
10	329	26 (0.749)	8 (0.247)	1 (0.031)	0 (0.000)	11 (0.158)	4 (0.265)	4 (0.009)	6 (0.068)
50	346	110 (0.544)	47 (0.130)	49 (0.047)	41 (0.006)	51 (0.051)	52 (0.074)	40 (0.052)	34 (0.024)
					$\rho = \mathbf{0.75}$				
10	432	12 (1.000)	13 (0.993)	0 (0.050)	0 (0.000)	6 (0.790)	0 (0.170)	0 (0.010)	0 (0.070)
50	448	85 (0.722)	90 (0.728)	11 (0.050)	4 (0.003)	88 (0.724)	7 (0.148)	0 (0.010)	8 (0.043)

Table 2.8 presents number of significant coefficients (and corresponding simulated FDR in brackets) from regressions of simulated returns on all the 1,500 simulated signals. LASSO (no FDR) shows number of significant coefficients at 5% level from regressions of returns on signals selected from the LASSO. OLS stands there for a second step OLS estimated with the selected signals and P-S Inf for significance under proper post-selection inference. 5% FDR category then adjusts for FDR either in regressions of all the 1,500 signals (OLS) or post-selection (LASSO: OLS or LASSO: P-S Inf). We control for FDR with either Benjamini and Yekutieli (2001) (BY) or Storey (2002) (STO) methods. Penalty in the LASSO is selected with three-fold cross-validation. The LASSO usually selects about 250 strategies.

We first study the impact of preselecting signals without properly adjusting p-values on the selected signals. This is analogous to focusing only on published anomalies and disregarding tried but unpublished signals. LASSO (no FDR) compares number of significant coefficients at 5% level without accounting for FDR. It is obvious that there are many more significant signals without proper post-selection inference. The same is also true after adjusting for FDR at 5% level. The bootstrapped true FDR is mostly higher than the 5% desired level. This means that the standard multiple hypothesis methods fail when they are applied to a preselected sample, as is the case for the published anomalies in Green et al. (2017).

We next focus on the differences between OLS with FDR and LASSO with post-selection inference and FDR. Both of these methods lead to correct FDR rates. OLS tends to select more signals. True FDR for STO method are close to the desired rates and it thus has the largest power to reject the true significant signals. BY then tends to be too strict and under-reject. The difference between the two FDR methods is minimal for the LASSO. Larger collinearity between the signals decreases power of all the tests. Selection of truly significant signals with $K = 10$ becomes infeasible.

We also show the number of signals that are significant in decile long-short portfolios with STO adjustment and 5% FDR (Port STO). The number is in line with our previous empirical analysis using multiple hypothesis methods in the US.

To conclude, we have documented that both small OOS R^2 and multicollinearity can lead to severe problems with the methods and the setting can easily become too challenging for them to work properly. Then, it is not possible to find any significant signal after properly correcting for FDR at 5%, although there are many of them in reality. Failure to find significant signals is not an evidence of no existing important signals, it could simply be a result of the poor power of the tests.

2.4.2 Empirical Evidence

Green et al. (2017) found that only 12 characteristics are reliably independent determinants in non-microcap stocks in the US from 1980 to 2014. They relied only on a set of published anomalies and it is highly possible that their conclusions were influenced by the absence of tried but unpublished signals. It is important to adjust p-values in regressions for the number of all tried signals for the very same reason as in the other multiple hypothesis tests. We will here revisit the issue with our universe of signals and methods.

Panel A in Table 2.9 presents number of selected independently significant signals in regressions of individual stock returns on the fundamental characteristics. We focus on three LASSO settings based on selection of penalty λ in the regressions. We set $\lambda = 1\%$ in the simplest setting. The value is chosen so that it yields parsimonious model with a few signals, but yet the same out-of-sample predictive power as those with more parameters. The value of λ is also close to those estimated with BIC criterion in the second setting. λ in the last setting is estimated to minimize the mean square error in 10-fold cross-validation.[32] This should give us the upper estimate since it tends to select model with too many parameters. We further control for FDR with either Benjamini and Yekutieli (2001) (BY) or Storey (2002) (STO) methods.

The number of selected signals widely differs depending on the values of λ. The number of statistically significant signals is, however, very similar. There are no significant signals in Japan, which is in line with the previous evidence of no significant signals in multiple hypothesis tests on portfolios. There is one clear trend in that anomalies tend to be more

[32] AIC criterion tends to select values that are between BIC and cross-validation.

Table 2.9:
Independently Significant Signals

The table shows number of independently significant signals from regressions of individual stocks returns on transformed fundamental signals. We either regress all the signals using weighted-least-squares regressions (WLS) or rely on LASSO to select smaller number of signals that do not suffer from collinearity. We control for FDR at 5% level with either Benjamini and Yekutieli (2001) (BY) or Storey (2002) (STO) method. Panel B then presents the significant signals chosen with LASSO in the US. Signals with bold p-values are also significant with FDR of 5%. The LASSO regressions are conducted on 1,590 fundamental signals; 1,497 data-mined and 93 from published studies. The data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. Panels C and E consider 861 (Reduced), 1,586 (Base), and 48,476 (Extended) signals, 89 of which are anomalies and the rest data-mined. Panels D and F further restrict the set of published anomalies to 24 that are closely tied to the data-mined signals. The sample is restricted to industrial stocks with price over $1 and capitalization larger than the bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 in the US and July 1995 to December 2016 in other regions.

Panel A: Number of Significant Signals

	Anom and d-m Signals				Data-mined Signals				Anomalies			
	USA	E	J	AP	USA	E	J	AP	USA	E	J	AP
WLS All Signals												
STO	6	3	0	0	0	1	0	0	25	19	3	0
Anomalies	4	2	0	0								
BY	2	1	2	0	0	2	0	0	19	9	3	0
Anomalies	2	1	1	0								
WLS Signals with VIF < 7												
STO	24	7	2	0	9	13	0	0	25	19	3	0
Anomalies	16	5	2	0								
BY	8	3	1	0	5	3	1	0	19	9	3	0
Anomalies	8	2	2	0								
LASSO with $\lambda = 1\%$												
Selected by LASSO	22	31	15	39	20	25	14	34	10	14	6	14
Post-selection inference	8	6	6	3	7	7	6	5	10	12	2	4
Anomalies	5	2	1	1								
Post-selection inference BY	5	2	3	0	4	2	1	0	10	4	0	1
Anomalies	3	1	0	0								
LASSO with λ Minimizing BIC												
λ	1.3%	2.8%	2.2%	3.7%	1.7%	3.3%	2.2%	3.7%	0.9%	2.1%	1.9%	3.1%
Selected by LASSO	12	3	0	0	6	0	0	0	10	3	0	0
Post-selection inference	8	3	0	0	6	0	0	0	10	3	0	0
Anomalies	7	2	0	0								
Post-selection inference BY	8	3	0	0	5	0	0	0	10	3	0	0
Anomalies	7	2	0	0								
LASSO with λ Minimizing MSE in 10-fold Cross-validation												
λ	0.4%	0.5%	2.2%	1.4%	0.6%	0.6%	2.2%	1.4%	0.2%	0.5%	1.0%	1.5%
Selected by LASSO	62	62	0	22	32	46	0	20	39	21	6	9
Post-selection inference	14	10	0	3	12	5	0	3	20	10	2	2
Anomalies	7	4	0	1								
Post-selection inference BY	4	2	0	0	4	0	0	0	11	5	0	0
Anomalies	2	2	0	0								

Panel B: List of the Significant Signals in the US Chosen with LASSO and $\lambda = 1\%$

Signal	Original study	p-value
BE/ME	Fama and French (1992)	0.021
Growth in LTNOA	Fairfield et al. (2003)	0.021
R&D/MV	Chan et al. (2001)	**0.000**
CBOP	Ball et al. (2016)	**0.006**
Earnings Predictability	Francis et al. (2004)	**0.000**
Change in OIBDP/SEQ		0.014
(trailing CHE)/ME		**0.000**
XSGA/ME		**0.007**

Table 2.9 Continued

	Reduced Set				Base Set				Extended Set			
	USA	E	J	AP	USA	E	J	AP	USA	E	J	AP
Panel C: Number of Significant Signals for STO, Base Set of Anomalies, VIF < 7												
All Signals	24/319	10/319	2/319	0/319	24/339	7/339	2/339	0/339	9/1194	2/1194	0/1194	0/1194
Anomalies	17/76	7/76	2/76	0/76	16/76	5/76	2/76	0/76	7/76	2/76	0/76	0/76
Panel D: Number of Significant Signals for STO, Reduced Set of Anomalies, VIF < 7												
All Signals	7/260	4/260	0/260	0/260	4/273	9/273	0/273	0/273	5/1130	0/1130	0/1130	0/1130
Anomalies	2/19	2/19	0/19	0/19	2/17	2/17	0/17	0/17	3/15	0/15	0/15	0/15
Panel E: Number of Significant Signals for STO, Base Set of Anomalies, VIF < 3												
All Signals	19/137	12/137	0/137	0/137	23/139	15/139	0/139	0/139	17/316	9/316	0/316	0/316
Anomalies	13/48	7/48	0/48	0/48	14/50	9/50	0/50	0/50	9/45	5/45	0/45	0/45
Panel F: Number of Significant Signals for STO, Reduced Set of Anomalies, VIF < 3												
All Signals	11/106	16/106	0/106	0/106	17/105	10/105	0/105	0/105	10/289	9/289	0/289	0/289
Anomalies	5/10	5/10	0/10	0/10	7/10	4/10	0/10	0/10	4/9	4/9	0/9	0/9

frequent among the selected signals in the US. Data mining thus leads to a selection of signals that are similar to those published in journals. This is expected since all the covered anomalies were discovered in the US with CRSP and Compustat data. The researchers thus tend to focus on the US market much more.

WLS section in Panel A deals with weighted-least-squares regression of returns on all the signals with consequent FDR correction. The weight in the regressions is proportional to one over number of stock in the cross-section each year to give equal weight to each time period. This is very similar to the approach in Green et al. (2017) when only anomalies are considered. The multi-collinearity issues are either ignored and all the signals are used, or they are partially dealt with by iteratively discarding anomalies that have variance inflation factor (VIF) of more than seven when considered with respect to the other anomalies.[33] The initial focus on independent anomalies is introduced in order to keep as many of them as possible. Data-mined signals with VIF larger than seven are then also iteratively discarded. STO leads to a larger number of significant signals than any specification of LASSO. BY, however, provides fewer signals due to poor power of the test. Green et al. (2017) therefore likely underestimate number of independently significant anomalies due to their reliance on BY. Dealing with multi-collinearity turns out to be very important and there are many more significant signals when the closely related signals are discarded. One striking feature is that there are many more significant anomalies when no data-mined signals are considered. This is due to the pre-selection problem described previously in the simulations. Any analysis focusing just on published anomalies therefore likely suffers from biases caused by omitting signals that were tried but not published.

The results reported so far were describing the base case with 1,497 data-mined signals and 89 published anomalies. Panel C extends the analysis to the reduced and extended sets of data-mined signals with 768 and 48,383 signals, respectively. Panel D reduces the number of published anomalies in Panel C from 89 to 24 so that all of them are closely related to the universe of data-mined signals. The results in both Panel C and D are based

[33] VIF is defined as $1/(1-R^2)$ in a regression of a given explanatory variable on all the other explanatory variables.

on Storey (2002) method applied to p-values from WLS regressions based on signals with VIF lower than 7, as in Panel A. The format of reporting is x/y where x is the number of significant signals and y is the number of remaining signals after the closely related signals are discarded. Panel C documents that the reduced set of data-mined signals has only a small impact on the number of significant independent signals. The number of significant signals, however, drops for the extended set of signals. The drop can be simply explained by a finite number of signals that can predict the individual stock returns and increasing proportion of noise signals with the new data-mined signals. The new noise signals contaminate the existing signals and make it harder to statistically reject them. The same is also true for the reduced set of published anomalies in Panel D. The number of significant signals is much smaller in Panel D relative to Panel C. The smaller number is again caused by lower number of signals with small p-values due to the reduced set of anomalies. The academic process is therefore able to create new signals that cannot be captured by a naive data-mining process in the regression setting. Panel E and F correspond to Panel C and D, respectively, with the exception that only signals with VIF lower than 3 are kept. There is only a small difference in Panel E relative to Panel C. There is notably a larger number of significant signals in Europe and for the extended set of data-mined signals. There are many more significant signals in Panel F relative to Panel D which supports the conjecture that there were too many noise signals in Panel D relative to true drivers of returns for the MHT tests to distinguish between them. To conclude, the number of data-mined signals plays a large role in the regression setting and too many data-mined signals can lead to fewer rejected anomalies.

We next turn to a detailed analysis of signals that were selected in the US in Panel B. There are five anomalies and three data-mined signals for the base case of LASSO with $\lambda = 1\%$ and without accounting for FDR. The original source of anomalies is described in the table. We prefer to interpret signals selected from LASSO as the selection guarantees that the signals have economically significant predictive ability along with their statistical significance. All the three data-mined signals are similar to some published anomaly.[34] There are seven anomalies among the 8 selected signals for λ minimizing BIC criterion. The anomalies include the five with $\lambda = 1$ plus cash flow over market value of equity of Lakonishok et al. (1994) and change in net non-current operating assets of Richardson et al. (2006). The eighth signal is again trailing cash over market value of equity. The shift between selected anomalies based on slight change in λ documents that the selection process is very unstable and could lead to very different outcomes for different research designs. The impact of FDR correction depends on the chosen λ and there is notably no reduction in significant signals for BIC.

The significant signals selected from just data-mined signals without the anomalies are all closely related to some published anomalies. They include operating income (OIDBP)

[34] The relevant paper is Soliman (2008) for change in operating income over book value of equity, Palazzo (2012) in the case of trailing cash (transformation 5) over market value, and Eisfeldt and Papanikolaou (2013) in the case of sales, general, and administrative expenses over market value of equity.

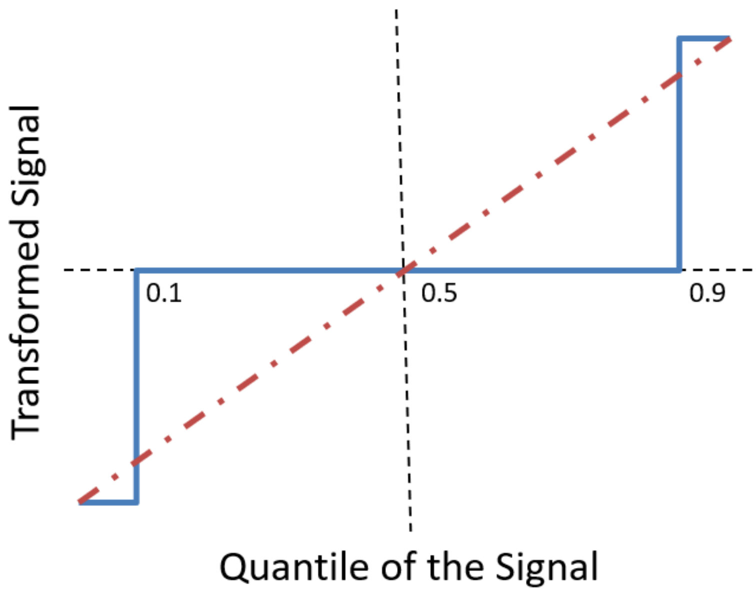


Figure 2.6: **Transformations of Signals in the Regressions.**

over total assets (similar to profitability), cash over market cap (similar to Palazzo (2012)), change in assets over market cap (similar to investments), SG&A over market cap (similar to Eisfeldt and Papanikolaou (2013)), and R&D expense over market cap (similar to Chan et al. (2001)). This documents that data mining without any regard for economic theory tends to find similar drivers of returns as the academic research. Furthermore the selected signals are close to those generally thought of as important in the academic literature. All the fundamental factors in Fama and French (2015) five factor model, or their close substitutes, appear somewhere among the selected significant signals in the US.

2.4.3 Portfolio-shaped Signals

The regressions have so far assumed that the fundamental signals predict mean stock returns in a linear form. That is, the individual stock returns are a linear function of cross-sectional quantiles of the signals. The linear form is, however, different from the previous portfolio-level analysis in that the fundamental signals have to predict the whole cross-section of the stock returns and not just the extreme deciles. Figure 2.6 compares the linear form with portfolio-shaped transformation of the signals. The portfolio-shaped transformation sets the upper $1-x$ cross-sectional quantile of the signal equal to 1, lower x quantile equal to -1, and is equal to zero otherwise. The x corresponds to 0.1 for decile sorts. We will now investigate whether the portfolio-shaped transformation of the signals has any impact on the previous conclusions.

Table 2.10 presents number of significant signals with the portfolio-shaped transformation of the signals. The table corresponds to Panels C to F in Table 2.9. The weighed least squares regression is again estimated with $1/N_t$ weights so that all the years have the same role in optimization regardless of the number of stocks in each of them. Three specifications of the portfolio-shaped regressions corresponding to decile (10/90 breakpoints), quintile (20/80 breakpoints), or third decile (30/70 breakpoints) long-short portfolios are considered. That is, x is either 0.1, 0.2, or 0.3. Only the base case with 1,490 data-mined signals is considered.

Table 2.10:
Independently Significant Portfolio-shaped Signals

The table shows number of independently significant signals from regressions of individual stocks returns on transformed fundamental signals. The transformed signals are shaped to correspond to kernels for portfolios, that is, the bottom x quantile of original signals each year is transformed to -1 and upper $1-x$ quantile of the original signals is transformed to 1. Where x is either first decile (10/90 breakpoints), second decile (20/80 breakpoints), or third decile (30/70 breakpoints) and the transformed signals are equal to 0 otherwise. We regress all the signals using weighted-least-squares regressions (WLS). We control for FDR at 5% level with Storey (2002) (STO) method. Signals with variance inflation factor (VIF) larger than 7 are discarded before running the regressions in Panel A and B and with VIF larger than 3 in Panel C and D. The regressions are conducted on 1,586 fundamental signals; 1,497 data-mined and 89 from published studies. Panel B and D further restricts the number of anomalies to 24 that are closely related to the data-mined signals. The data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. The sample is restricted to industrial stocks with price over $1 and capitalization larger than the bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 in the US and July 1995 to December 2016 in other regions.

	10/90 Breakpoints				20/80 Breakpoints				30/70 Breakpoints			
	USA	E	J	AP	USA	E	J	AP	USA	E	J	AP
Panel A: Base Set of Anomalies, VIF <7												
All Signals	2/1115	0/1115	0/1115	0/1115	3/1028	2/1028	0/1028	0/1028	5/905	0/905	0/905	0/905
Anomalies	2/87	0/87	0/87	0/87	3/88	2/88	0/88	0/88	3/88	0/88	0/88	0/88
Panel B: Reduced Set of Anomalies, VIF <7												
All Signals	0/1052	0/1052	0/1052	0/1052	0/957	0/957	0/957	0/957	3/835	0/835	0/835	0/835
Anomalies	0/22	0/22	0/22	0/22	0/22	0/22	0/22	0/22	0/835	0/835	0/835	0/835
Panel C: Base Set of Anomalies, VIF <3												
All Signals	3/464	2/464	1/464	0/464	20/285	9/285	0/285	0/285	30/210	15/210	1/210	0/210
Anomalies	3/76	2/76	1/76	0/76	11/71	6/71	0/71	0/71	19/60	7/60	0/60	0/60
Panel D: Reduced Set of Anomalies, VIF <3												
All Signals	1/410	1/410	0/410	0/410	12/235	8/235	0/235	0/235	15/169	10/169	0/169	0/169
Anomalies	1/21	1/21	0/21	0/21	5/16	3/16	0/16	0/16	6/15	5/15	0/15	0/15

There are almost no significant signals for VIF smaller than 7 in Panel A and B. The main reason for the low number of the significant signals is that the number of signals discarded because the VIF was larger than 7 is also low. The setting with about one thousand signals corresponds to the case with extended set of data-mined signals in Table 2.9, where the number of significant signals was low as well. The results in Panel C and D with VIF smaller than 3 for 20/80 and 30/70 specification of breakpoints in the portfolio-shaped transformation then roughly correspond to the linear specification in Table 2.9. To conclude, the results with the linear transformation can be replicated with the portfolio-shaped transformation which provides further robustness to the findings in Table 2.9.

2.5 Out-of-sample Tests

The analysis so far has focused on the selection of significant signals for the full available sample. We now turn to profitability out-of-sample. Previous evidence has shown that the selected anomalies in academic journals are very similar to the selected data-mined signals. This is why it could be the case that data mining has a similar predictive power as the academic research. There is, however, also a good reason why this does not have to be the case. All the published anomalies in good journals have to undergo a vetting procedure during their publication. The vetting guarantees that they are backed with sound reasoning which should in turn increase profitability out-of-sample.

There are two types of out-of-sample comparison that we offer here. First, there is a comparison in the US in that we select the historically most successful signals and observe how they fare out-of-sample. This comparison is within the original market but outside the original sample's time period. We also study how they perform outside their original market; in Europe, Japan, and Asia Pacific. The international test should provide a good setting to test the hypothesis of better external validity of the published anomalies as none of the original studies included international tests. We first study out-of-sample profitability of the signals in investment strategies and then examine predictive ability of the signals with respect to returns on individual stocks in formal tests.

2.5.1 Out-of-sample Profitability

Panel A of Table 2.11 shows out-of-sample returns on a simple strategy that equally invests into individual significant fundamental signals. Long-short decile portfolios are created based on all of the fundamental signals. Every June, we select portfolios based on all the published anomalies with returns significant at 5% level in one sided hypothesis test and hold the selected portfolios for the next year.[35] Only significant anomalies are selected as some of the anomalies cannot be replicated after the micro-caps have been excluded from the sample. The strategy essentially evaluates mean out-of-sample return on average published anomaly that can be replicated. A similar strategy is then also created for the 1,497 data-mined anomalies with one change that the same number of the most significant signals is selected as for published anomalies in the given year. The same number of data-mined signals is chosen so that the out-of-sample performance captures the ability of academic studies to better select important signals in contrast to just looking at significance level. Panel B selects all long-short decile portfolios based on data-mined signals that are significant in multiple hypothesis tests of Storey (2002) at $\gamma = 5\%$. Panels A and B therefore study out-of-sample performance at portfolio level of the simplest aggregate strategy that equally invests in the individual historically significant quantitative strategies. The selection process of significant signals is separately conducted for

[35]We consider only anomalies published before the formation of the mixed portfolio so that the results are not driven by forward looking bias.

Table 2.11:
Performance Persistence

The table shows out-of-sample performance of data-mined or published fundamental signals. Strategy in Panel A and B selects n_t of the historically most significant long-short portfolios created based on individual fundamental signals in the US at the end of each June starting in 1995 and equally invests in them for one year. n_t corresponds to a number of significant (t-statistic larger than 1.65) anomalies published by the time t in Panel A and to number of data-mined signals significant based on Storey (2002) and $\gamma = .05$ in Panel B. We select the strategies separately from a subsample of 93 anomalies published by the time of portfolio formation and then from 1,497 data-mined signals. Panel C is based on long-short decile portfolio from strategy that combines all the available signals through predictive LASSO regressions of individual stocks' returns on transformed fundamental signals in the US. The out-of-sample performance is observed in the US, Europe, Japan, and Asia Pacific. The 1,497 data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. The sample is restricted to industrial stocks with price over $1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 in the US and July 1995 to December 2016 in other regions. The value-weighted or equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals. The alphas are estimated with Fama and French (2015) five factor model (FF5). The returns are in percentage points per month. Standard errors in t-statistics are HAC adjusted, as in Newey and West (1987) with 12 lags.

	Equal-weighted Portfolios				Value-weighted Portfolios			
	USA	Europe	Japan	Asia Pacific	USA	Europe	Japan	Asia Pacific
Panel A: Individual Signals Selected Based on Their Significance								
Data Mined Signals								
Mean Return	0.323	0.236	-0.098	0.422	0.285	0.101	-0.337	0.260
	(3.030)	(1.610)	(-0.816)	(3.520)	(2.550)	(0.891)	(-1.080)	(0.698)
FF5 alpha	0.196	0.023	-0.113	0.332	0.193	-0.022	-0.351	-0.367
	(2.180)	(0.238)	(-1.270)	(2.590)	(1.870)	(-0.210)	(-1.490)	(-0.748)
Published Anomalies								
Mean Return	0.278	0.306	0.128	0.446	0.281	0.253	0.184	0.432
	(3.970)	(5.790)	(2.640)	(5.360)	(3.350)	(3.350)	(1.930)	(4.250)
FF5 alpha	0.206	0.180	0.093	0.337	0.202	0.093	0.089	0.224
	(3.190)	(4.490)	(2.430)	(3.590)	(2.450)	(1.550)	(1.350)	(2.520)
Diff wrt the data mining	-0.046	0.070	0.226	0.024	-0.004	0.152	0.521	0.172
	(-0.583)	(0.589)	(2.270)	(0.184)	(-0.036)	(1.910)	(2.000)	(0.538)
Panel B: Individual Data-mined Signals Selected Based on Storey (2002) and $\gamma = .05$								
Mean Return	0.027	0.047	-0.011	0.043	0.114	0.072	0.016	0.298
	(1.030)	(2.020)	(-0.561)	(1.880)	(1.630)	(1.070)	(0.168)	(2.460)
FF5 alpha	0.035	0.012	0.005	0.015	0.120	0.067	0.068	0.100
	(1.750)	(0.583)	(0.326)	(0.639)	(1.740)	(1.020)	(0.933)	(0.850)
Panel C: Signals aggregated with LASSO Regressions								
Data Mined Signals								
Mean Return	0.644	1.250	0.775	1.270	0.533	0.434	0.549	1.010
	(1.830)	(3.260)	(2.460)	(4.060)	(1.800)	(1.340)	(1.100)	(2.850)
FF5 alpha	0.274	0.689	0.554	0.873	0.346	-0.043	0.096	0.722
	(1.190)	(3.130)	(2.480)	(2.070)	(1.280)	(-0.132)	(0.261)	(1.470)
Published Anomalies								
Mean Return	0.819	1.110	0.565	1.180	0.643	0.510	0.398	0.975
	(1.530)	(2.590)	(1.610)	(3.830)	(1.640)	(1.600)	(1.250)	(3.070)
FF5 alpha	0.062	0.421	0.218	0.492	0.025	0.024	-0.060	0.060
	(0.193)	(1.810)	(1.090)	(1.650)	(0.110)	(0.126)	(-0.311)	(0.161)
Diff wrt the data mining	0.175	-0.137	-0.209	-0.092	0.110	0.076	-0.151	-0.033
	(0.516)	(-0.681)	(-1.370)	(-0.341)	(0.377)	(0.271)	(-0.438)	(-0.081)

equal-weighted and value-weighted portfolios.[36]

There is some slight under-performance of the published anomalies in the US although not significant at 5% level. There is a good reason why performance in the US should be lower than expected for the published anomalies. McLean and Pontiff (2016) showed that there is a large drop in profitability of anomalies after they are published. They ascribe this drop to informed trading where investors take the opportunity to make profits. The published anomalies are, however, much more successful in the international setting. Equal-weighted portfolios are significant in all regions for published anomalies but they are insignificant in Japan and Europe for the data-mined signals. This is even more evident for value-weighted portfolios which are significant in three regions for published anomalies but only in the US for data-mined signals. These differences are, however, significant at 5% level only in Japan.

Panel B tests whether a more advanced technique to select all outperforming signals adds any value. Multiple hypothesis tests were developed in order to select the largest possible number of individually significant signals and they should, at least theoretically, provide better out of sample performance. The table documents that it is not the case and selecting a fixed small number of signals works better. This is in line with the findings in Bajgrowicz and Scaillet (2012) who found no out-of-sample predictability of technical trading rules using Storey (2002) test to select likely outperforming signals. As as result, the ability of the formal multiple hypothesis tests to select individual signals with persistent profitability is questionable. Chordia et al. (2017) came essentially to the same conclusion with their 2.1 million strategies.

The evidence presented so far supports the hypothesis that published anomalies outperform data mining out-of-sample. This is as expected since publishing anomalies means overcoming a detailed scrutiny of referees, who require valid theoretical underpinning for the new findings. This rigorous process should then lead to a better selection of signals. The profitable anomalies can be identified with important underlying risks, which is connected to risk premia or to behavioural biases that are innate to humans.

We now turn to what happens when supervised machine learning techniques are applied in Panel C. To do this, we use LASSO regressions of individual stock returns on fundamental signals in the US using data available up to the portfolio formation at the end of each June. The approach is described in Section 2.4 and we set $\lambda = 1\%$. Future returns on individual stocks are predicted with the latest available fundamentals. Equal-weighted and value-weighted portfolios are created based on sorts on the predictions. That is, the zero-cost long-short portfolios are created by buying stocks in the top decile of predicted returns and shorting stocks in the bottom decile of predicted returns. Only anomalies that have been published by the time of portfolio formation are used in order not to create look ahead bias.[37] Shrinking the number of signals should lead to more profitable

[36]The value-weighted portfolios here label a strategy that equally invests in the value-weighted portfolios based on the individual signals, but is not value-weighted itself per se.

[37]The results are not hugely influenced by inclusion of all anomalies for the whole period. The perfor-

strategies as it should provide optimal combination of signals that jointly lead to highest returns. This is indeed the case and returns have increased in all the regions for both data-mined and published anomalies. Data mining now provides positive returns in all the regions and it is not significantly different from the published anomalies. The same applies for the risk adjusted returns. Although the risk adjusted returns are noticeably lower now due to lower diversification among the profitable signals.

2.5.2 Out-of-sample Return Predictability on All the Stocks

Returns on the decile portfolios are not the objective being optimized when minimizing squared loss in the LASSO regressions. The previous results do not properly access the predictive ability of the data mined signals and the anomalies. We will now investigate the out-of-sample (OOS) predictive power for returns on all stocks and not only those in extreme deciles. We follow Gu et al. (2018) and define absolute predictive ability of the individual forecast with OOS R^2

$$1 - \frac{\sum_{it}(r_{it} - \hat{f}(x_{it}))^2}{\sum_{it} r_{it}^2} \tag{2.7}$$

where $\hat{f}$ is a predictive function fitted on data preceding year t. Both estimation and OOS R^2 is done on annual signals and returns. Unlike Gu et al. (2018), we use demeaned returns in each year and region instead of excess returns (over value-weighted market returns) since we do the same in the LASSO regressions. Raw returns without demeaning lead to almost identical performance. We compare individual forecasts using anomalies relative to using data mining in the Diebold and Mariano (1995) test. Specifically, we adopt the approach in Gu et al. (2018) and create a time-series of differences of cross-sectional sums of squared losses of the two forecasts. We then test significance of the differences by testing significance of their time-series average with simple t-test. We adjust the standard errors in t-statistics for heteroskedasticity and autocorrelation with Newey-West procedure with 3 lags.

The results in Table 2.12 are in line with our previous results for portfolio returns. The anomalies research and data mining lead to a very similar OOS predictive performance but data mining slightly wins here. The difference is significant only in Europe. The fact that academic research does not have superior predictive power should not be surprising in the light of our previous evidence. We have shown that signals that are selected as independently significant from LASSO are very similar for data mining and anomalies. Academic research thus identifies important drivers of returns which are then very similar to what pure data mining approach finds, at least in the US, where most of the anomalies were found.

mance in the US does not improve but rather shrinks to one third for value-weighted portfolios.

Table 2.12:
Out-of-sample R^2

This table shows out-of-sample R^2 of the predictions of individual stocks' returns using data-mined or published fundamental signals. The predictions follow Panel C of Table 2.11 and are based on a strategy that combines all the available signals through predictive LASSO regressions of individual stocks returns on transformed fundamental signals in the US. The out-of-sample performance is observed in the US, Europe, Japan, and Asia Pacific over July 1990 to December 2016 period. The 1,497 data-mined fundamental signals are created by various transformations of 49 accounting variables, as described in the Section 2.1.2. The sample is restricted to industrial stocks with price over \$1 and capitalization larger than bottom decile in NYSE at the end of previous June. It spans July 1963 to December 2016 in the US and July 1990 to December 2016 period in other regions. The OOS R^2 are in percentage points.

	USA	Europe	Japan	Asia Pacific
Anomalies	0.488	0.744	1.150	0.818
Data mining	0.551	1.290	1.680	1.050
Difference	-0.339	-2.830	-1.870	-0.587

2.6 Conclusion

We have documented that it is very difficult to select outperforming strategies outside micro-cap universe of stocks. After carefully accounting for biases in returns, it is possible to do reasonably well at explaining the returns on a wide range of fundamental anomalies. Critical values for significance of signals in multiple hypothesis tests are higher than the critical values for single hypothesis tests but they also heavily depend on specific setting. The equal-weighted returns tend to lead to higher number of significant anomalies and this in turn leads to less strict critical values for t-statistics for a given level of significance. The number of significant signals also critically depends on adjustment of standard errors of returns for heteroskedasticity and autocorrelation. The number of significant anomalies can shrink to one third if twelve lags are included in the HAC robust adjustment of standard errors instead of just one lag. We have shown that the individual significant anomalies identified in the US are profitable in all the regions out-of-sample but this is not the case for the data-mined signals. Selecting individual data-mined signals solely based on their past long-short portfolio returns is not a profitable strategy in Japan when micro-caps are excluded. Using machine learning tools, however, shrinks this advantage of academic research. There is no significant difference in predictive ability of the data-mined signals and the published academic anomalies when LASSO regressions are used to synthesize the individual signals into one mispricing measure.

The sole focus of this study is fundamental anomalies. The analysis could be easily extended into other types of anomalies such as those based on past returns or quarterly fundamental data. We leave this to future research.

Appendix E

49 Fundamental Variables Used in Construction of the Data-mined Fundamental Signals

The list of variables is provided in Table E.1. The table also gives a link on how the corresponding variables in Datastream are constructed. Appendix F then provides the full names of variables in Datastream matched to their shortcuts.

Table E.1:
Fundamental Variables Used for Construction of Fundamental Strategies
The table shows all fundamental variables that were required for construction of our fundamental anomalies.

BALANCE SHEET		
ASSETS		
Current Assets		
Cash and Short-Term Investments	CHE	WC02001
Receivables - Total	RECT	WC02051
Inventories - Total	INVT	WC02101
Current Assets - Other - Total	ACO	WC02149 + WC02140
Prepaid Expenses	XPP	WC02140
Current Assets - Total	ACT	WC02201
Non-Current Assets		
Long-Term Investments	IVAO	WC02258 + WC02250
Property Plant and Equipment - Total (Net)	PPENT	WC02501
Property Plant and Equipment - Total (Gross)	PPEGT	WC02301
Property Plant and Equipment Buildings at Cost	FATB	WC18376
Property Plant and Equipment Leases at Cost	FATL	WC18381
Investment and Advances - Equity	IVAEQ	WC02256
Intangible Assets - Total	INTAN	WC02649
Assets - Total	AT	WC02999
LIABILITIES AND SHAREHOLDERS' EQUITY		
Current Liabilities		
Debt in Current Liabilities	DLC	WC03051
Account Payable/Creditors - Trade	AP	WC03040
Current Liabilities - Other - Total	LCO	WC03066 + WC03054 + WC03063 + WC03061
Income Taxes Payable	TXP	WC03063
Current Liabilities - Total	LCT	WC03101
Long-Term Liabilities		
Long-Term Debt - Total	DLTT	WC03251
Liabilities - Other	LO	WC03273 + WC03262
Liabilities - Total	LT	WC03351
Minority Interest - Balance Sheet	MIB	WC03426
Shareholders' Equity		
Preferred/Preference Stock (Capital) - Total	PSTK	WC03451
Retained Earnings	RE	WC03495
Shareholders' Equity - Total	SEQ	WC03501 + WC03451
Common/Ordinary Equity - Total	CEQ	WC03501

INCOME STATEMENT		
Sales/Turnover (Net)	SALE	WC01001
Cost of Goods Sold	COGS	WC01051
Selling, General and Administrative Expenses	XSGA	WC01101
Research and Development Expense	XRD	WC01201
Earnings Before Interest, Taxes & Depreciation	OIBDP	WC01151 + WC01250
Depreciation and Amortization - Total	DP	WC01151
Earnings Before Interest and Taxes	OIADP	WC01250
Interest and Related Expense	XINT	WC01251
Pretax Income	PI	WC01401
Income Taxes - Total	TXT	WC01451
Income Before Extraordinary Items	IB	WC01551
CASH FLOW STATEMENT		
Indirect Operating Activities		
Operating Activities - Net Cash Flow	OANCF	WC04860
Investing Activities		
Capital Expenditures	CAPX	WC04601
Investing Activities - Net Cash Flow	IVNCF	- WC04870
Financing Activities		
Purchase of Common and Preferred Stock	PRSTKC	WC04751
Sale of Common and Preferred Stock	SSTK	WC04251
Cash Dividends	DV	WC04551
Dividends on Common Stock	DVC	WC05376
Long-Term Debt - Issuance	DLTIS	WC04401
Long-Term Debt - Reduction	DLTR	WC04701
Net Changes in Current Debt	DLCCH	WC04821
Financing Activities - Net Cash Flow	FINCF	WC04890

Appendix F

Names of Variables in Datastream Matched to Their Shortcuts

Table F.1:
Fundamental variables in Datastream

	Assets		35	Total Assets	02999
1	Cash & Short Term Investments	02001		**Liabilities & Shareholders' Equity**	
2	Cash	02003	36	Accounts Payable	03040
3	Short Term Investments	02008	37	Short Term Debt & Current Portion of Long Term Debt	03051
4	Receivables (Net)	02051	38	Accrued Payroll	03054
5	Inventories -Total	02101	39	Income Taxes Payable	03063
6	Raw Materials	02097	40	Dividends Payable	03061
7	Work in Process	02098	41	Other Current Liabilities	03066
8	Finished Goods	02099	42	Current Liabilities - Total	03101
9	Progress Payments & Other	02100	43	Long Term Debt	03251
10	Prepaid Expenses	02140	44	Long Term Debt Excluding Capitalized Leases	03245
11	Other Current Assets	02149	45	Non Convertible Debt	18281
12	Current Assets - Total	02201	46	Convertible Debt	18282
13	Long Term Receivables	02258	47	Capitalized Lease Obligations	03249
14	Investment in Associated Companies	02256	48	Provision for Risks and Charges	03260
15	Other Investments	02250	49	Deferred Income	03262
16	Property Plant and Equipment Net	02501	50	Deferred Taxes	03263
17	Property Plant and Equipment - Gross	02301	51	Deferred Taxes - Credit	18183
18	Land	18375	52	Deferred Taxes - Debit	18184
19	Buildings	18376	53	Other Liabilities	03273
20	Machinery & Equipment	18377	54	Total Liabilities	03351
21	Transportation Equipment	18380	55	Non-Equity Reserves	03401
22	Property Plant & Equipment under Capitalized Leases	18381	56	Minority Interest	03426
23	Property Plant & Equipment - Other	18379	57	Preferred Stock	03451
24	Accumulated Depreciation	02401	58	Common Equity	03501
25	Accumulated Depreciation - Land	18383	59	Common Stock	03480
26	Accumulated Depreciation - Buildings	18384	60	Capital Surplus	03481
27	Accumulated Depreciation - Machinery & Equipment	18385	61	Revaluation Reserves	03492
28	Accumulated Depreciation - Transportation Equipment	18388	62	Other Appropriated Reserves	03493
29	Accumulated Depreciation Other Property Plant & Equipment	18387	63	Retained Earnings	03495
30	Accumulated Depreciation - PPE under Capitalized Leases	18389	64	ESOP Guarantees	03496
31	Other Assets	02652	65	Unrealized Foreign Exchange Gain/Loss	03497
32	Deferred Charges	02647	66	Unrealized Gain/Loss on Marketable Securities	03498
33	Tangible Other Assets	02648	67	Treasury Stock	03499
34	Total Intangible Other Assets - Net	02649	68	Total Liabilities & Shareholders' Equity	03999

Table F.1 continued

	Additional items		102	Preferred Dividend Requirements	01701
69	Trade Receivables - Net	18297	103	Net Income after Preferred Dividends (Basic EPS)	01706
70	Provision for Bad Debt	18298	104	Extraordinary Items & Gain/Loss Sale of Assets	01601
71	Other Accrued Expenses	03069		**Additional items**	
72	Current Portion of Long Term Debt	18232	105	Research & Development Expense	01201
	Income statement		106	Restructuring Expense	18227
73	Net Sales or Revenues	01001		**Cash Flow statement**	
74	Cost of Goods Sold	01051	107	Net Income / Starting Line	04001
75	Depreciation, Depletion & Amortization	01151	108	Depreciation, Depletion & Amortization	04051
76	Depreciation	01148	109	Depreciation and Depletion	04049
77	Amortization of Intangibles	01149	110	Amortization of Intangible Assets	04050
78	Amortization of Deferred Charges	01150	111	Deferred Income Taxes & Investment Tax Credit	04101
79	Gross Income	01100	112	Deferred Income Taxes	04199
80	Selling, General & Administrative Expenses	01101	113	Total Other Cash Flow	04151
81	Other Operating Expenses	01230	114	Funds from Operations	04201
82	Operating Expenses - Total	01249	115	Extraordinary Items	04225
83	Operating Income	01250	116	Funds from/for Other Operating Activities	04831
84	Extraordinary Credit - Pre-tax	01253	117	Decrease/Increase in Receivables	04825
85	Extraordinary Charge - Pre-tax	01254	118	Decrease/Increase in Inventories	04826
86	Non-Operating Interest Income	01266	119	Increase/Decrease in Accounts Payable	04827
87	Pre-tax Equity in Earnings	01267	120	Increase/Decrease in Income Taxes Payable	04828
88	Other Income/Expense - Net	01262	121	Increase/Decrease in Other Accruals	04829
89	Interest Expense on Debt	01251	122	Decrease/Increase in Other Assets/Liabilities	04830
90	Interest Capitalized	01255	123	Net Cash Flow - Operating Activities	04860
91	Pre-tax Income	01401	124	Capital Expenditures (Additions to Fixed Assets)	04601
92	Income Taxes	01451	125	Additions to Other Assets	04651
93	Current Domestic Income Tax	18186	126	Net Assets from Acquisitions	04355
94	Current Foreign Income Tax	18187	127	Increase in Investments	04760
95	Deferred Domestic Income Tax	18188	128	Decrease in Investments	04440
96	Deferred Foreign Income Tax	18189	129	Disposal of Fixed Assets	04351
97	Minority Interest	01501	130	Other Uses/(Sources) - Investing	04797
98	Equity in Earnings	01503	131	Other Uses - Investing	04795
99	After Tax Other Income/Expense	01504	132	Other Sources - Investing	04796
100	Discontinued Operations	01505	133	Net Cash Flow - Investing	04870
101	Net Income before Extraordinary Items/Preferred Dividends	01551	134	Net Proceeds from Sale/Issue of Common & Preferred	04251

Table F.1 continued

135	Proceeds from Stock Options	04301
136	Other Proceeds from Sale/Issuance of Stock	04302
137	Com/Pfd Purchased, Retired, Converted, Redeemed	04751
138	Long Term Borrowings	04401
139	Increase/Decrease in Short Term Borrowings	04821
140	Reduction in Long Term Debt	04701
141	Cash Dividends Paid - Total	04551
142	Common Dividends (Cash)	05376
143	Preferred Dividends (Cash)	05401
144	Other Sources/(Uses) - Financing	04448
145	Other Sources - Financing	04446
146	Other Uses - Financing	04447
147	Net Cash Flow - Financing	04890
148	Effect of Exchange Rate on Cash	04840
149	Increase/Decrease in Cash & Short Term Investments	04851
	Additional items	
150	Total Sources	04501
151	Total Uses	04811

Appendix G

93 Published Anomalies

Table G.1:
List of Published Fundamental Anomalies

Accruals	
Accruals	Sloan (1996)
Change in Common Equity	Richardson et al. (2006)
Change in Current Operating Assets	Richardson et al. (2006)
Change in Current Operating Liabilities	Richardson et al. (2006)
Change in Financial Liabilities	Richardson et al. (2006)
Change in Long-Term Investments	Richardson et al. (2006)
Change in Net Financial Assets	Richardson et al. (2006)
Change in Net Non-Cash Working Capital	Richardson et al. (2006)
Change in Net Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Liabilities	Richardson et al. (2006)
Change in Short-Term Investments	Richardson et al. (2006)
Discretionary Accruals	Dechow et al. (1995)
Growth in Inventory	Thomas and Zhang (2002)
Inventory Change	Thomas and Zhang (2002)
Inventory Growth	Belo and Lin (2011)
M/B and Accruals	Bartov and Kim (2004)
Net Working Capital Changes	Soliman (2008)
Percent Operating Accrual	Hafzalla et al. (2011)
Percent Total Accrual	Hafzalla et al. (2011)
Total Accruals	Richardson et al. (2006)
Intangibles	
$\triangle$ Gross Marging - $\triangle$ Sales	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Acounts Receivable	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Inventory	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ SG and A	Abarbanell and Bushee (1998)
Asset Liquidity	Ortiz-Molina and Phillips (2014)
Asset Liquidity II	Ortiz-Molina and Phillips (2014)
Cash-to-assets	Palazzo (2012)
Earnings Conservatism	Francis et al. (2004)
Earnings Persistence	Francis et al. (2004)
Earnings Predictability	Francis et al. (2004)
Earnings Smoothness	Francis et al. (2004)
Earnings Timeliness	Francis et al. (2004)
Herfindahl Index	Hou and Robinson (2006)
Hiring rate	Belo et al. (2014)
Industry Concentration Assets	Hou and Robinson (2006)
Industry Concentration Book Equity	Hou and Robinson (2006)
Industry-adjusted Organizational Capital-to-Assets	Eisfeldt and Papanikolaou (2013)
Industry-adjusted Real Estate Ratio	Tuzel (2010)
Org. Capital	Eisfeldt and Papanikolaou (2013)
RD / Market Equity	Chan et al. (2001)
RD Capital-to-assets	Li (2011)
RD Expenses-to-sales	Chan et al. (2001)
Tangibility	Hahn and Lee (2009)
Unexpected RD Increases	Eberhart et al. (2004)
Whited-Wu Index	Whited and Wu (2006)

Table G.1 continued

Investment	
△ CAPEX - △ Industry CAPEX	Abarbanell and Bushee (1998)
Asset Growth	Cooper et al. (2008)
Change Net Operating Assets	Hirshleifer et al. (2004)
Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
Composite Debt Issuance	Lyandres et al. (2007)
Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
Debt Issuance	Spiess and Affleck-Graves (1995)
Growth in LTNOA	Fairfield et al. (2003)
Investment	Titman et al. (2004)
Net Debt Finance	Bradshaw et al. (2006)
Net Equity Finance	Bradshaw et al. (2006)
Net Operating Assets	Hirshleifer et al. (2004)
Noncurrent Operating Assets Changes	Soliman (2008)
Share Repurchases	Ikenberry et al. (1995)
Total XFIN	Bradshaw et al. (2006)
Profitability	
Asset Turnover	Soliman (2008)
Capital Turnover	Haugen and Baker (1996)
Cash-based Operating Profitability	Ball et al. (2016)
Change in Asset Turnover	Soliman (2008)
Change in Profit Margin	Soliman (2008)
Earnings / Price	Basu (1977)
Earnings Consistency	Alwathainani (2009)
F-Score	Piotroski (2000)
Gross Profitability	Novy-Marx (2013)
Labor Force Efficiency	Abarbanell and Bushee (1998)
Leverage	Bhandari (1988)
O-Score (More Financial Distress)	Dichev (1998)
Operating Profits to Assets	Ball et al. (2016)
Operating Profits to Equity	Fama and French (2015)
Profit Margin	Soliman (2008)
Return on Net Operating Assets	Soliman (2008)
Return-on-Equity	Haugen and Baker (1996)
Z-Score (Less Financial Distress)	Dichev (1998)
Value	
Assets-to-Market	Fama and French (1992)
Book Equity / Market Equity	Fama and French (1992)
Cash Flow / Market Equity	Lakonishok et al. (1994)
Duration of Equity	Dechow et al. (2004)
Enterprise Component of Book/Price	Penman et al. (2007)
Enterprise Multiple	Loughran and Wellman (2011)
Intangible Return	Daniel and Titman (2006)
Leverage Component of Book/Price	Penman et al. (2007)
Net Payout Yield	Boudoukh et al. (2007)
Operating Leverage	Novy-Marx (2010)
Payout Yield	Boudoukh et al. (2007)
Sales Growth	Lakonishok et al. (1994)
Sales/Price	Barbee Jr et al. (1996)
Sustainable Growth	Lockwood and Prombutr (2010)

Chapter 3

Does It Pay to Follow Anomalies Research? Machine Learning Approach with International Evidence

Low interest rates environment after the Financial Crisis of 2008 has caused a surge in search for alternative ways of how to earn steady returns that are uncorrelated with the stocks market. One response of the financial industry was an explosion in a number of "smart beta" funds that provide exposure to various risk factors, which have been historically connected to risk premia. This larger interest should, however, in turn lead to their lower profitability. McLean and Pontiff (2016) document the decrease of 58% in post-publication returns relative to the in-sample returns of anomalies. Jacobs and Müller (2017b) however show that the United States is the only country with a reliable post-publication decline in returns of anomalies, emphasizing the importance of international evidence in asset pricing. Apart from the lack of the post-publication decline in the international setting, Jacobs and Müller (2017a) find that combining anomalies into one mispricing signal using least squares leads to superior out-of-sample risk-adjusted returns relative to focusing on individual anomalies. The benefit of combining individual anomalies through predictive regressions is further emphasized by Gu et al. (2018) who conclude that sophisticated machine learning methods offer higher out-of-sample predictability in the US compared to the traditional methods in Jacobs and Müller (2017a). This study extends the use of machine learning methods to international sample and finds internationally unprecedented out-of-sample profitability using anomalies as predictors in machine-learning-based predictive regressions.

In order to benchmark machine learning based strategy (mispricing strategy hereafter) we look at out-of-sample profitability of a portfolio-level strategy that invests in the individual published anomalies (portfolio-mixing strategy hereafter). Having all the constructed anomalies at our disposal, we examine degree of predictability of future profitability of the individual anomalies based on their past profitability in various regions. To our knowledge, this is the first study to focus the question whether international evidence for individual anomalies can actually help with predictions of their future returns. We also study the value of international evidence for the prediction of out-of-sample stock returns in the mispricing strategy. Furthermore, even though machine learning methods are notoriously hard to interpret we look at the marginal variable importance in our predictive regressions and find substantial heterogeneity across the regions, forecasting methods, as well as liquidity-based subsamples. Next, we examine limits to arbitrage associated with our strategies. We are the first to extensively estimate transaction costs associated with strategies leveraging predictive power of anomalies internationally and document that strategies remain profitable even after accounting for the transaction costs as well as short-selling constraints. Since we only include anomalies as predictors after their publication we also examine the marginal value of the new anomalies for the out-of-sample predictions after accounting for the already published anomalies and show that it remains positive over time, confirming added value of recent anomalies literature.

153 published anomalies are studied in the US, Japan, Europe, and Asia Pacific. The anomalies in this study describe characteristics related to individual stocks that can predict their future returns. No distinction is being made between characteristics that are related to risk premia and characteristics that are related to mispricing due to frictions or other market imperfections. The studied anomalies are, for example, accruals of Sloan (1996), earnings over price of Basu (1977), composite equity issuance of Daniel and Titman (2006), and R&D over Market Equity of Chan et al. (2001). The focus in this study is restricted to a liquid universe of stocks. The liquid stocks are defined as the largest stocks with capitalization in the top 90% of the overall market's capitalization and dollar trading volume over the previous year in the top 90% of the overall market's volume in the individual regions. Only about 500 most liquid stocks pass the criteria in 2010s in a given month in the US. Excluding small-capitalization stocks leads to results more relevant to investors and limits effect of microstructure noise.[1]

The portfolio-mixing strategy describing average return on the individual anomalies is first considered. The portfolio-mixing strategy equally invests in portfolios created based on individual anomalies that are significant in the US at 5% level.[2] Hou et al. (2017) show that many of the published anomalies disappear on liquid universe of stocks. Our stock universe is far more liquid relative to Hou et al. (2017). The focus on significant anomalies in the strategy therefore guarantees that the conclusions are not driven by inclusion of these irrelevant strategies, as would be the case for the simplest strategy taking into account all the published anomalies in Hou et al. (2017). The weighting in the strategy is the simplest possible and the strategy's average returns can be interpreted as average return on individual anomalies that were historically significant. The average returns are expected to be positive if there is any persistence in returns on the anomalies.

[1] See Asparouhova et al. (2010) for description of the effect of microstructure noise.
[2] It is later shown that the results do not depend on the 5% significance level.

The significant anomalies are selected once a year, at the end of June. Only anomalies that are published by the time of selection are considered. Green et al. (2017) documented a significant drop in performance of all anomalies in the US after 2003. A similar drop is observed on the portfolio-mixing strategy and it's average annualized return drops to less than 2% after accounting for transaction costs.

The strategy that synthesize information from all the anomalies into one mispricing signal is studied next. The strategy first predicts next-month returns on individual stocks from their past characteristics (cross-sectional quantiles of the anomalies). Investment portfolios are then constructed by buying stocks in top decile of the predicted returns and short-selling stocks in the bottom decile of the predicted returns. Historical relation between the past characteristics and future returns is estimated on the past data. The next month returns on individual stocks are predicted from the latest characteristics. The historical relationships are typically linearly approximated using Fama and MacBeth (1973a) least squares regressions in the academic literature, as in Lewellen et al. (2015). Gu et al. (2018) showed that machine learning methods can significantly outperform the linear approximation in the US. The use of machine learning methods is extended here from the US to international markets. The least squares regressions are compared to gradient boosting regression trees, random forest, and neural networks. The machine learning methods lead to significant gains in performance of the mispricing strategy in all the regions.

Value of international evidence for the prediction of out-of-sample returns on the anomalies is evaluated. Hou et al. (2017) and Harvey et al. (2016) showed that many anomalies cannot be replicated and many others are significant only due to the in-sample data snooping. New anomalies are discovered using the same historical datasets in the US which can lead to false positive discoveries. International data provides new information with respect to the US, and it could therefore limit the number of false discoveries.[3] International data also increases sample size which in turn leads to more powerful statistical tests. One problem could be that some anomalies are specific to the US as they depend on the local institutional setting. For example accruals depend on country-specific accounting rules. The institutional uniqueness then limits the value of data outside the US for predictions in the US. There is a little gain from forecasting the expected future returns in the US based on historical data outside the US relative to focusing solely on the historical US data in the mispricing strategy. The forecasts in the other regions, however, gain accuracy from historical data in the respective regions when it is added to historical data in the US. Mispricing of stocks estimated on historical data in the US captures most

[3]Note that many anomalies have been individually studied in the international markets. For examples of studies investigating cross-sectional predictability of individual signals outside the US see Chui et al. (2010), Barber et al. (2013), McLean et al. (2009), Rouwenhorst (1998), Lam and Wei (2011), Titman et al. (2013), and Watanabe et al. (2013). The goal here is not the study of performance of the anomalies outside of the US but rather the use of international historical performance of the anomalies to better select anomalies that are likely to outperform in the future.

of predictability of stock returns outside the US.[4]

Marginal value of new anomalies for out-of-sample predictions after accounting for the already published anomalies is evaluated. Most of the widely accepted risk factors have been published before 1995. Examples include size and book-to-market ratio in Fama and French (1992) and momentum of Jegadeesh and Titman (1993). The new discoveries should therefore have lower marginal explanatory power over time as the strongest predictors of stock returns have been already revealed. It is also possible that the vetting procedure that authors have to undergo during the publishing process limits these decreasing returns to the new discoveries. The value of recent anomalies is examined by comparing out-of-sample returns of the mispricing strategy that synthesizes anomalies published either before 1995, 2000, or 2005. There is a gradual increase in mean returns and Sharpe ratio on the mispricing strategy over 2005 to 2016 period, when the more recently published anomalies are added. Investors can therefore benefit from following the recent academic anomalies research.

Limits to arbitrage could explain the strategies' profitability and it might not be possible to invest into the mispriced stocks. Several robustness checks are therefore conducted. The returns on the long-short portfolios are decomposed into long-only and short-only components. It is often impossible to short-sell due to insufficient supply of borrowable stocks. Both the long-only and short-only legs of the mispricing strategy, however, offer an investment opportunity with respect to returns on the market. Short-selling constraints cannot therefore fully explain the profitability. Transaction costs on the investment strategies are studied next. It is concluded that both the portfolio-mixing strategy and the mispricing strategy remain profitable after the transaction costs.[5]

The focus of this study is the closest to Jacobs and Müller (2017c) and Jacobs and Müller (2017a) who analyzed returns on anomalies outside the US. This study is, however, different in many aspects. Firstly, it focuses on liquid universe of stocks which should make the results more relevant to any investor. Secondly, the role of international evidence in the strategies is investigated. Jacobs and Müller (2017c) and Jacobs and Müller (2017a) focused solely on strategies that were using data in the respective regions without evaluating the possible benefits of using the global data. Thirdly, the prediction methods differ. The introduction of advanced machine learning techniques significantly improves the out-of-sample fit of the predictions in this study.

The study is the closest in methodology and application of machine learning techniques

[4]The role of international evidence for the mispricing signal is broadly related to variety of factor structures outside the US. The international evidence is likely to add little value if there is no proximity of factor structures across the regions. For examples of papers investigating factor structure of international returns see Fama and French (2012), Fama and French (2017), Rouwenhorst (1999), Griffin (2002), Griffin et al. (2010), Hou et al. (2011b), and Bartram and Grinblatt (2018b).

[5]Novy-Marx and Velikov (2015) studied transaction costs on a range of anomalies in the US and concluded that the transaction costs are important mainly for high-turnover anomalies whose returns net of transaction costs often turn negative. Frazzini et al. (2012) demonstrated that real-life transaction costs for large portfolio managers are much lower than assumed by academics. In particular, returns on momentum and value style premia survive transaction costs and have large investment capacity. The transaction costs can be further lowered by appropriate optimized portfolio rebalancing.

to Gu et al. (2018) who, however, focused solely on the US. Gu et al. (2018) in other respects, differ from this study with their focus on full universe of stocks which has profound effects on their conclusions. The most important anomalies in their estimation are liquidity, size, and return over the past month (short-term return reversal). Asparouhova et al. (2010) argue that these variables are connected to future returns mainly through microstructure biases and have nothing to do with true predictability of stock returns that is of interest to investors.[6] The machine learning methods were built to find all patterns in the dependent variable and this leads to sub-optimal outcome when predicting stock returns on illiquid stocks. Focus on large cap universe helps to address these concerns. Secondly, a large difference with respect to Gu et al. (2018) is that this study allows only already published anomalies to enter predictions in each year. That is, the information set of existing anomalies was available to investors by the time they would make a decision of where to invest their money. Ignoring this assumption can lead to illusory profits that cannot be obtained in practice.

The contributions of this study are multiple. Firstly, the role of international evidence for predictions of future returns on individual stocks is evaluated. Most of academic anomalies research focuses solely on the US and benefits of international evidence have not been systematically studied before. It is shown that training sample outside the US does not largely improve forecasts of expected returns on the individual stocks in the US. Secondly, the marginal value of recent anomalies, while controlling for the well established anomalies, is evaluated. It is shown that the recently published anomalies are providing new information about the cross-section of stock returns.

The key takeaways of this chapter are that machine learning methods are a superior method for predicting future individual stock returns not only in the US but also in all the other regions. The past evidence from the US encompasses most of information contained in the stocks-level historical data and international data adds only little to profitability of the mispricing strategy. Furthermore, the more recently published anomalies are important for out-of-sample profitability and following anomalies research is therefore beneficial for the investors.

3.1 Data and Methodology

3.1.1 Data

The source of accounting and market data for the US is Merged CRSP/Compustat database from Wharton Research Data Service (WRDS). The sample spans 1926 to 2016 period and contains all New York Stock Exchange (NYSE), Amex, and NASDAQ common stocks (CRSP share code 10 or 11). The returns are adjusted for delisting following guidance in Hou et al. (2017).[7]

[6]See Roll (1984) for a simple model decomposing stock returns into microstructure noise and changes in true prices.

[7]If the delisting is on the last day of the month, returns over the month are used. The relevant delisting return is then added as a return over the next month. Delisting return (DLRET) from monthly

The international data is sourced from Reuters Datastream. It is filtered following Ince and Porter (2006), Lee (2011), and Griffin et al. (2010). The procedure comprises of manually checking names of the shares in the database for over 100 expressions describing their share class. Only the primary quotes of ordinary shares of the companies are retained, with few exceptions where fundamental data in Datastream is linked to other share classes.[8] Real Estate Investment Trusts (REITs) are excluded from the sample. All the international returns and financial statements in this study are converted to US dollars. The daily returns are deleted for days when the stock market was closed in a given country. The quality of data is further improved with procedures described in Chapter 1. Chapter 1 studies implications of the choice of fundamental database on the measurement of performance of individual fundamental anomalies. It shows that statistical significance of the individual anomalies varies across Datastream and Compustat. The research inference can therefore change when a different fundamental database is used. The differences across the databases are mainly due to imperfect historical fundamental coverage. Studies of aggregated performance of anomalies, however, do not suffer from these problems. Analysis in this study is therefore not impacted.

The sample includes 23 developed countries. The countries are sorted into 4 regions: the USA; Europe (E) - Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom; Japan (J); and Asia Pacific (AP) - Australia, New Zealand, Hong Kong, and Singapore.

Another important source of data for the anomalies is Institutional Brokers' Estimate System (I/B/E/S) which is obtained from WRDS. I/B/E/S is merged on Datastream directly as it is one of databases provided by Thompson Reuters and Datastream includes the respective tickers in its static file. The merger with CRSP is done indirectly through CUSIPs. The databases are merged on 8 digit CUSIP and then on 6 digit CUSIP if unsuccessful. The success of the merger is checked manually by comparing quoted tickers on the exchanges and names of the companies. All the variables in I/B/E/S are transformed to US dollars with original Reuters exchange rates which are provided by WRDS.

This study focuses only on the liquid universe of stocks. The liquid universe category covers only stocks that are both (a) within the top 90% of the overall capitalization of all stocks in each region at the end of previous June and (b) within the top 90% of the overall dollar trading volume over the previous 12 months of all stocks in each region. The stocks are further required to have price larger than $1 ($.1 for Asia Pacific) at the end of the previous June. The restriction on capitalization in the US roughly corresponds to 50% percentile of the largest stocks on NYSE. Stocks outside the US are further restricted to have capitalization larger than the bottom decile at NYSE. This further constraint

file is used if it is not missing. $(1 + ret_{cum}) * (1 + DLRET_d) - 1$ is used if it is missing, where ret_{cum} is cumulative return in the given month of delisting and $DLRET_d$ is delisting return from the daily file. Lastly, the gaps are filled with $(1 + ret_{cum}) * (1 + DLRET_{avg}) - 1$, where $DLRET_{avg}$ is average delisting return for stocks with the same first digit of delisting code (DLSTCD).

[8] The description in Griffin et al. (2010) on classification of common shares is followed.

guarantees that the stocks are not only liquid with respect to other stocks in the region but also with respect to the stocks in the US.

Table 3.1 shows average, minimum, and maximum number of stocks in the cross-section of the individual regions. Full sample category includes all the available stocks without any restrictions. There are on average about 500 stocks in the US that satisfy the criteria. The average number of stocks in even smaller in the other regions. Average capitalization of stocks in the liquid universe after July 1995 is $24 billion in the US, $21 billion in Europe, $9 billion in Japan, and $11 billion in Asia Pacific. Average size of the stocks in the sample is therefore balanced over the regions.

Table 3.1:
Number of Stocks in the Cross-section

	Full sample			Liquid Universe		
	mean	min	max	mean	min	max
Asia Pacific	2430	1012	3706	132	71	238
Europe	5194	4440	6121	350	208	826
Japan	3141	2074	3678	331	208	744
USA	4768	1993	7525	495	263	829

3.1.2 Anomalies

The sample includes 153 anomalies published in academic studies. The full list of the anomalies is provided in Appendix H. Anomalies that have been described in McLean and Pontiff (2016), Hou et al. (2017), or Harvey et al. (2016) are primarily selected. The study focuses only on anomalies that are valid in the cross-section of stocks so that long-short portfolios can be formed out of them. Any anomalies that are specific to the US, and which cannot therefore be constructed outside the US, are excluded.[9] Fundamental signals are updated annually at the end of every June using financial statements from financial years ending in the previous calendar year.[10]

Some anomalies, such as Herfindahl Index of Hou and Robinson (2006), require classification of industries for individual firms. The choice in the original papers is mostly with respect to Standard Industrial Classification (SIC). Third level Datastream classification, sorting industries into 19 groups, is applied here instead. The larger industry groups should make the results more robust and consistent across the data vendors. The industry classification in Datastream is available only from the static file, which means that only the latest values are available. Data vendors may slightly differ in the classification of individual firms over time because the differences between individual SIC categories

[9]This includes anomalies: based on quarterly fundamental data since there is only short coverage internationally; connected to hand collected data in the US such as IPOs, SPOs, and mergers; requiring segment information and NBER data; and that are institutionally specific, such as, share turnover or effective tax rate. Some fundamental anomalies could not be implemented in Datastream as the required items are missing there.

[10]Section J.1 documents that the annual refreshing of fundamental signals provides very similar results to monthly refreshing.

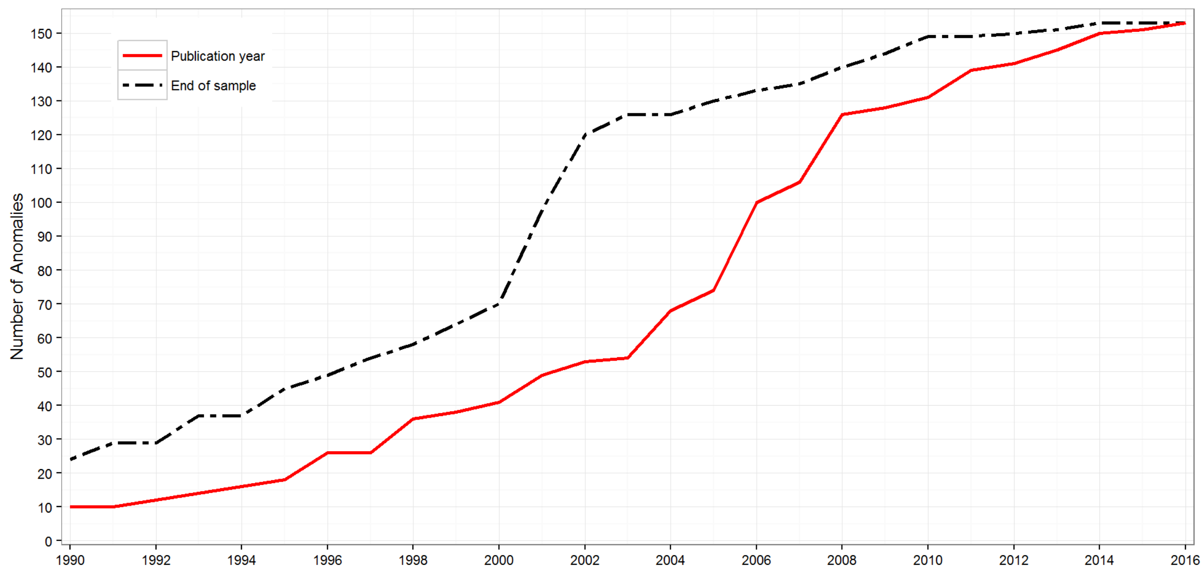


Figure 3.1: **Number of the Published Anomalies Over Time.**

are often subtle. A translation table between SIC classification and the Datastream classification is provided in the Appendix C.

There are 93 fundamental, 11 I/B/E/S, and 49 market friction anomalies in the sample. The anomalies come almost exclusively from the top finance and accounting journals. Figure 3.1 graphs number of the published anomalies over time. The second line is capturing number of anomalies whose in-sample period in their respective studies has ended. The number of anomalies has been gradually increasing over time without any apparent jumps.

3.1.3 Portfolio-mixing Strategy

This section describes the portfolio-mixing strategy that equally weights returns from the portfolios on individual anomalies. It serves as a benchmark for the more complicated mispricing strategy described in the next section. The strategy is especially useful when studying the role of international evidence in the selection of quantitative strategies that outperform out-of-sample as it can be understood as a combination rule of multiple quantitative strategies based on the past evidence. Portfolio construction for the individual anomalies is first described and the logic for how the individual portfolios are combined is discussed next.

Portfolio Construction for the Individual Anomalies

The portfolios are constructed on the liquid universe of stocks. The focus on liquid universe should make the findings more realistic to someone trying to trade the anomalies. The stocks with small capitalization (micro-cap) account for only a small fraction of the overall capitalization of the market, often cannot be traded at significant volumes due to their high illiquidity. Chapter 1 documents that the fundamental coverage of micro-cap stocks outside the US is very problematic in Datastream and the imperfect coverage can introduce huge biases into the analysis. Both equal-weighted and value-weighted returns are always provided. The preference should be given to interpretation of the value-weighted returns since they do not suffer from the market microstructure biases documented in Asparouhova et al. (2010). These biases can be substantial and can heavily influence the analysis.

The portfolios on individual anomalies start in July 1963 in the US and July 1990 in Europe, Japan, and Asia Pacific.[11] The period before 1963 in the US is omitted due to the quality of returns and number of available stocks in CRSP is very low during that time. The fundamental coverage of stocks in Compustat is also very low which makes the construction of majority of the anomalies impossible. Further restrictions of the sample of stocks, based on industries, age of the firms, and the length of history of the firms' fundamental data, follow the original studies when constructing portfolios on the individual anomalies. The original studies are also followed regarding rebalancing period of the portfolios so that most of anomalies in I/B/E/S and market friction categories are rebalanced monthly, whereas, fundamental anomalies are mostly rebalanced annually at the beginning of every July. The zero-cost long-short portfolios on the individual anomalies are constructed by buying stocks in the top decile of the signals and shorting stocks in the bottom decile of the signals.[12]

Table 3.2:
Average Time-series Correlations of Returns on Portfolios Created for the Individual Anomalies Across the Regions.

	USA	E	J	AP
USA	1.000	0.239	0.105	0.120
E	0.239	1.000	0.126	0.122
J	0.105	0.126	1.000	0.094
AP	0.120	0.122	0.094	1.000

Table 3.2 presents average of time-series correlations of returns on the long-short portfolios created from identical anomalies across the different regions. The anomalies are not closely correlated across the regions. The international evidence should therefore be very useful as it can serve as an independent source of information for stock return predictability.

Combining Individual Portfolios into One Strategy

The portfolios on individual anomalies are combined into one meta-portfolio through a simple strategy. The portfolio-mixing strategy selects all the anomalies whose portfolio returns are significant at 5% level and equally weights them into a single portfolio. The

[11] International studies using fundamental data, such as Fama and French (2017), usually start in 1990. The reason for this is that there is an insufficient fundamental coverage before that.

[12] The zero-cost portfolios are preferred since some annually rebalanced anomalies experience lower than -100% return during some years which creates problems with the definition of return in terms of relative change in value of the invested money with respect to the previous month. It would be necessary to introduce leverage constraints which would unnecessarily complicate the analysis.

selection is repeated at the end of every June from 1995 to 2016. Many of the published anomalies cannot be replicated on the liquid universe of stocks and the selection based on historical significance guarantees that only robust strategies are used. Significance of the anomalies is determined based on returns available up to the given June. Only anomalies published by the given June are considered. The significance is determined based on p-values that are adjusted for heteroskedasticity and auto-correlation for up to 12 lags.

The equal-weighting of portfolios on individual anomalies adds robustness to the strategy. It could be beneficial to use information of historical covariance structure between the strategies. DeMiguel et al. (2007), however, show that 1/N weighting provides a very robust performance out-of-sample and no other simple weighting strategy is able to beat it.

3.1.4 Mispricing Strategy

The focus has so far been on portfolio level analysis of the individual anomalies. The rest of this section covers the strategy that shrinks all the anomalies into a single mispricing signal ("mispricing strategy"). Lewellen et al. (2015) defined the prediction problem as follows: the goal is to devise a forecasting method that predicts which stocks are likely to have the highest returns in the next month and which the lowest based on stock characteristics (the cross-sectional anomalies). To do this, monthly returns on individual stocks are regressed on their past characteristics. The future return are then predicted from the latest available characteristics. The regressions are estimated by pooling all the available stock returns up to the date of portfolio formation. The past characteristics have to be available before the start of measurement period of the returns. The characteristics are normalized to their cross-sectional quantiles within each region to reduce problems with outliers.

To summarize, the following equation is estimated

$$r_{it} = f(x_{i,t-1,1}, x_{i,t-1,2}, ..., x_{i,t-1,M}) + \epsilon_{it} \tag{3.1}$$

where r_{it} is return on stock i in month t and $x_{i,t-1,1}$ is cross-sectional quantile of a given anomaly (characteristic) for the stock i available before the start of month t. The returns are demeaned by subtracting average cross-sectional returns in every region-month. A simpler case with linear $f()$ is first covered. It is then extended to a more general structure using machine learning. The machine learning exercise follows Gu et al. (2018) who applied a suite of standard machine learning algorithms and showed that they outperform the linear models in the US. Readers are referred to Gu et al. (2018) or any advanced machine learning textbook for a detailed theoretical description of the machine learning methods and only basic definitions are covered here.[13]

The machine learning methods have both some benefits and some negatives. They provide better out-of-sample forecasts through limitation of in-sample over-fitting. They also allow for a very general interaction between the explanatory variables. This general

[13]See, for example, Friedman et al. (2001) for the textbook treatment.

form, however, makes the fitted models hard to estimate and the estimates hard to interpret due to the black-box approach. The intractability of the estimates is not a large concern in this study since even the linear method becomes intractable given the number of exogenous variables. The main metric of this study is out-of-sample performance and not the interpretation of the estimated parameters, which is in line with the optimization objective of the machine learning methods.

The machine learning methods usually depend on some pre-specified meta-parameters. This study follows the common approach in machine learning literature to choose the meta-parameters in data-dependent way through three-fold cross-validation (CV). The CV splits the historical sample into pairs of mutually exclusive validation samples and training samples. The model is estimated on the training sample with various meta-parameters and its performance is captured on the validation samples. The meta-parameters, maximizing the performance over all the validation samples, are then selected for the estimation. The CV splits divide the historical sample into three consecutive parts with similar length.[14]

Weighted Least Squares

The benchmark model uses weighted least square estimation for linear approximation of the relationship in equation (3.1). That is, a weighted least square regressions of the stock returns on the rescaled characteristics is estimated

$$r_{it} = \beta_0 + \beta_1 x_{i,t-1,1} + \beta_2 x_{i,t-1,2} + ... + \beta_M x_{i,t-1,M} + \epsilon_{it} \tag{3.2}$$

where the weight on individual observation is the inverse of number of stocks in the each time period and region. The weights are introduced to give equal importance to the each time period. The weighting makes the moment conditions equivalent to Fama and MacBeth (1973a) regressions in Lewellen et al. (2015). The linear specification has already been applied in international context in Jacobs and Müller (2017c) and Jacobs and Müller (2017a). It is therefore selected as a benchmark for the more complicated machine learning methods.[15]

Penalized Weighted Least Squares

The linear regression model with many explanatory variables can overfit the realization of past data since it has many degrees of freedom. One way how to reduce the overfitting is to introduce L1 and L2 penalties on the coefficients during the estimation. The penalties are chosen by the three-fold cross-validation. The cross validation mostly selects only L1 penalty. The case with just L1 penalty is denoted least absolute shrinkage and selection operator (LASSO) and was introduced in Tibshirani (1996).

[14] The sample splits for the initial historical sample period 1963 - 1995 are, for example, 1973 and 1984. The pairs of training and validation samples are then [1963 - 1984, 1985 - 1995], [1963 - 1973 plus 1985 - 1995, 1974 - 1984], and [1974 - 1995, 1963 - 1973].

[15] Capitalization-weighted regressions as in Green et al. (2017) have been also tried. The capitalization-weighting puts lower weight on small cap stocks and is more suited for value-weighted portfolios. The weighting did not outperform the selected method and the results are therefore not reported here.

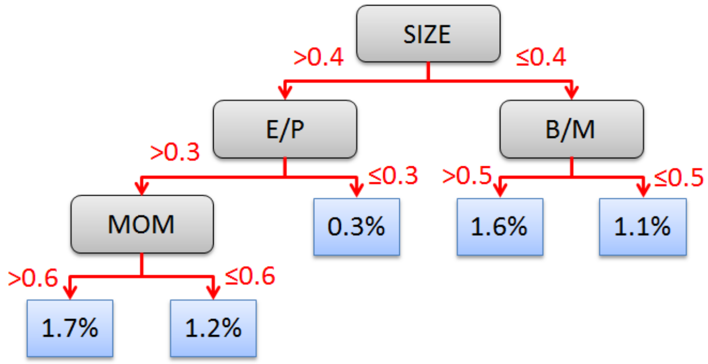


Figure 3.2: **Decision Tree.**

Random Forest

The regression tree family of methods is easy to estimate and requires a few specified meta-parameters. One such tree is depicted in Figure 3.2. The decision tree consists of nodes (the round-edged boxes) and outcomes (sharp-edged boxes). The outcomes are in percent return per month.[16] The tree starts with a decision whether a given stock is within the smallest 40% of stocks in the cross-section. The decision can then continue to the split based on the book to market ratio. The depicted tree is of depth 3, which is the maximum number of nodes in the longest branch. The tree allows for arbitrary cross-effects between the variables up to the (depth - 1) degree. This study deals mainly with relatively shallow trees. The shallow trees are nonetheless able to capture various important interactions between the explanatory variables. Random Forest and Gradient Boosting Regression Trees are based on a combination of the individual trees. These methods cannot be easily visualized but they lead to a better out-of-sample forecasting performance relative to simpler regression trees.

Random forest is one of the most widely used ensemble tree method. It combines forecasts from the individual decision trees that are based on subsamples of the training data. Explanatory variables are also subsampled in the individual trees to increase variety among the individual forecasts. Random forest is frequently among the top 10% of best performing machine learning methods in various competitions and it is therefore a very robust method that is powerful in most of the settings. It requires only few specified meta-parameters. The specification of the meta-parameters is furthermore not very important for its performance. It can therefore be used almost out-of-box. This is a large benefit with respect to neural networks where performance heavily depends on specification of the model. The largest downside is that its estimates is time consuming.

The results in this study are based on a combination of 500 trees. The trees use randomly selected 50% of the overall training observations and square root of the overall available explanatory variables. Minimum node size is chosen to be 0.1% of all the training observation to leave the method completely meta-parameter free. The 0.1% is large enough to limit over-fitting but small enough to allow the method to approximate the

[16] The numbers are arbitrary and do not reflect real data.

true expected returns on stocks.[17]

Gradient Boosting Regression Trees

Gradient boosting regression trees (GBRT) of Friedman (2001) rely on a different way of combining the regression trees than random forest. All the trees in random forest are chosen independently, whereas, they are selected in a dependent fashion in GBRT. The idea is to estimate a tree and use only a fraction of its fit for forecasts. The next iterations then proceed on residuals of the dependent variable after removing the fraction of the fitted values in the previous iteration. Shrinkage of the individual predictions guarantees that the learning can correct itself if the fitted values are selected suboptimally in some iterations. The fraction of individual predictions that is retained for the forecast is called a learning rate. Number of the learning iterations, given the learning rate, then determines how closely the particular realization of the sample from the whole population (the training sample) is over-fitted. A selection of fewer iterations reduces the risk of over-fitting (estimation error) but decreases the overall fit of the estimation (i.e. introduces an approximation error). It is therefore important to select the number of iterations with optimal estimation and approximation error trade-off. One way to do this is to rely on a cross-validation. The method requires a specification of learning rate, number of iterations (trees), and maximum depth of the trees.

The analysis in this study is conducted with a fast version of the gradient boosting - extreme gradient boosting (XGBOOST) of Chen and He (2017). The reason for this is that it is ten times faster to estimate and thus requires far less computational power. Gu et al. (2018) benchmarked the different machine learning methods and only neural networks provided significantly better forecasts than GBRT. GBRT is therefore a good candidate for the empirical application and it captures most of the gains from the machine learning methods over the standard finance methods. That is why GBRT is used to examine the benefits of international training sample in section 3.3 and the benefits of recent anomalies in section 3.4.

The specification of the GBRT is set as follows: the maximum depth of the trees is determined by a cross-validation. Depth of up to 9 nodes is considered. Gu et al. (2018) showed that cross-validation selects similar values in their analysis. The learning rate is set to 10%.[18] Number of iterations is again determined via the three-fold cross-validation.

Neural Networks

Arguably the most powerful machine learning method of today is (deep) neural networks. Gu et al. (2018) show that they outperform any other method if they are optimally specified. The neural networks are a very flexible tool that encompasses many specifications.[19]

[17] Ignoring this this parameter completely, and leaving unlimited note size, leads to almost identical results. It is thus not an important assumption.

[18] Experimenting with the learning rate did not lead to any increase in the predictive power. There is an extreme amount of noise in the financial data and slower learning is thus not necessary.

[19] A linear regression is the simplest specification.

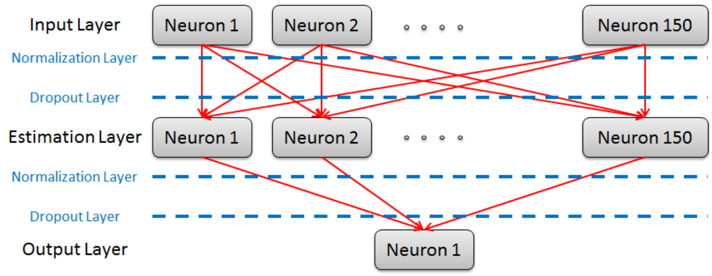


Figure 3.3: **Neural Network.**

The flexibility is also their largest disadvantage as it requires a long experimentation and possible over-fitting of the sample.[20]

Sequential neural networks consist of layers of neurons with information flowing between the layers in only one direction, from input layer to output layer. The information is fed in batches consisting of n sample points. Processing of the full training sample is called an epoch. The speed of change in estimated parameters with new processed batches is determined through the learning rate. It is often an advantage to slow the learning rate over time to allow for finer details to be captured. The neural networks are estimated with back-propagation and stochastic gradient descent.

Figure 3.3 plots specification of the neural network in this study. It is based on three layers. The initial layer has 150 neurons. The second hidden layer also has 150 neurons. The last output layer only has one neuron. The first two layers use a rectified linear unit (ReLU) activation function while the last layer uses a linear activation. Input into each layer is batch normalized. The network is regularized with dropout layers where output of fifteen randomly selected neurons is dropped in the first and the second layer in each epoch. Early stopping callbacks then provide further regularization and stop the learning process once the mean squared loss stops improving in the validation sample in four consecutive epochs. Another callback reduces the learning rate when the mean squared loss stops improving from one epoch to another.

The final forecast is produced from a combination of three estimated neural networks with different initial random seeds. Each run also uses different validation-training sample splits to further increase variety over the forecasts. The combination forecast leads to a great improvement in the performance of the mispricing strategy based on the neural networks.

Portfolio Construction

The mispricing portfolios start in July 1995, unless stated otherwise. They are again long-short self-financing and are rebalanced every month. The long leg of the strategy buys stocks in the upper decile of the predicted next month's returns. The short leg of

[20] The over-fitting should be a large cause of worry and all results based on neural networks should be taken with a grain of salt. The tree-based methods work well out of box even with default setting but neural networks require a long fine tuning. The fine tuning will translate into problematic performance out-of-sample of this study.

the strategy short-sells stocks in the bottom decile of the predicted next month's returns.

The portfolios are constructed based on sorts of the predicted returns in the individual regions. Global strategy invests into stocks from all the four regions. The global strategy is again based on stocks in the extreme deciles of the predicted returns in the individual regions.

The portfolio returns now also correspond to an investable strategy that holds $1 in cash, invests $1 in the stocks that are likely to have the largest return in the next month, and shorts $1 worth of stocks that are likely to have the smallest return in the next month. The portfolios are rebalanced to have an equal position in cash, long, and short leg of investment in the stocks at the beginning of each month.

3.1.5 Liquidity Measures

Liquidity costs on the strategies are studies with several liquidity proxies. The proxies are: VoV(% Spread) of Fong et al. (2017), Gibbs proxy of Hasbrouck (2009), and closing quoted spread proxy of Chung and Zhang (2014). They are defined in detail in Appendix I.

The proxies were selected to capture a fixed component of transaction costs and ignore variable component that measures price impact of larger orders. The variable component is very volatile and depends on the precise trade execution algorithm of each asset manager. The large capitalization universe of stocks reduces concerns about the variable component and it should be possible to avoid any execution costs altogether through the use of limit orders.

All of the proxies have some missing observations. The missing observations are backfilled from the other proxies. Quoted spread is used first for the backfilling, followed by VoV(% Spread), and the remaining missing observations are backfilled with Gibbs proxy. Less than 0.02% of the observations is missing in all the three proxies and these observations are filled by 5% costs.

3.2 Profitability

3.2.1 Portfolio-mixing Strategy

The portfolio level analysis of the individual anomalies is a good starting point as it provides a simple indication of out-of-sample profitability of the anomalies. The more complicated method, that synthesizes information embedded in the individual anomalies to one mispricing signal, is just a refinement of this simple strategy.

Table 3.3 presents returns on the portfolio-mixing strategy that invests equally in all the portfolios on anomalies that have significantly positive returns at 5% significance level as described in Section 3.1.3. That is, it corresponds to a setting where someone is following anomalies research, replicates the published findings, and equally invests into all published anomalies that he was able to replicate on the liquid universe of stocks. The performance of the portfolio-mixing strategy is followed in all the regions. The out-of-sample forecasts begin in July 1995. Global strategy equally invests in the portfolio-mixing strategy in the four developed regions.

Table 3.3:
Out-of-sample Performance of the Portfolio-mixing Strategy
The table shows returns of the strategy that equally invests in all the anomalies that are significant in the US at 5% significance level as described in Section 3.1.3. The significant anomalies are selected once a year, at the end of June. Only anomalies that are published by the time of selection are considered. The reported returns are for July 1995 to December 2016 period and are in percentage points.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Mean	0.174	0.297	0.001	0.663	0.284	0.301	0.180	0.253	0.882	0.404
Sharpe Ratio	0.227	0.484	0.002	0.695	0.566	0.387	0.270	0.198	0.816	0.598
Skewness	0.083	-0.085	-1.885	-1.087	-0.436	0.356	0.197	-0.046	1.871	1.358
Kurtosis	9.963	9.230	14.68	13.16	6.920	6.481	9.243	25.09	16.59	22.41
Max Drawdown	-29.40	-17.96	-27.63	-27.43	-12.95	-18.12	-26.33	-61.07	-17.35	-20.79

The portfolio-mixing strategy is not statistically significant in the US for both equal-weighted and value-weighted returns over 1995-2016 period and Sharpe ratio is also low there. The profitability is sometimes higher in the other regions. The strategy is the most profitable in Asia Pacific. The returns are higher outside the US despite the fact that the anomalies have been chosen in the US without any regard for evidence from the other countries. The anomalies documented in academic literature in the US are therefore successful in capturing risk premia outside the US. Diversification among the regions also provides some benefits. The global strategy has Sharpe ratio close to 0.6.

Maximum drawdown (DD) is defined as

$$\min_{s>t} 100 * (P_s/P_t - 1) \tag{3.3}$$

where P_t is market value of all assets held in the strategy at time t. That is, DD is the largest relative drop in value of the invested money over the 1995 to 2016 period. DD is the smallest in Asia Pacific regions for value-weighted returns, which is in line with the highest returns and Sharpe ratio there. It is, nonetheless, also small in other regions, except for Japan.

Green et al. (2017) showed that the profitability of all anomalies has decreased significantly after 2003. The same decline in profitability is documented in Figure 3.4. The figure presents evolution of cumulative returns on the portfolio mixing strategy since June 2002. The profitability of the individual anomalies in the US has dropped to the point that they yielded only about 20% in this whole period. The strategy was more profitable in other regions.

The portfolio-mixing strategy relies on a specific threshold for the decision whether to include a given anomaly in the mix. Figure 3.5 documents that the results are robust to the choice of this threshold. The figure shows annualized mean returns and Sharpe ratios for the portfolio-mixing strategy that equally invests into all anomalies whose historical returns have t-statistic larger than threshold specified at x-axis. Mean returns are increas-

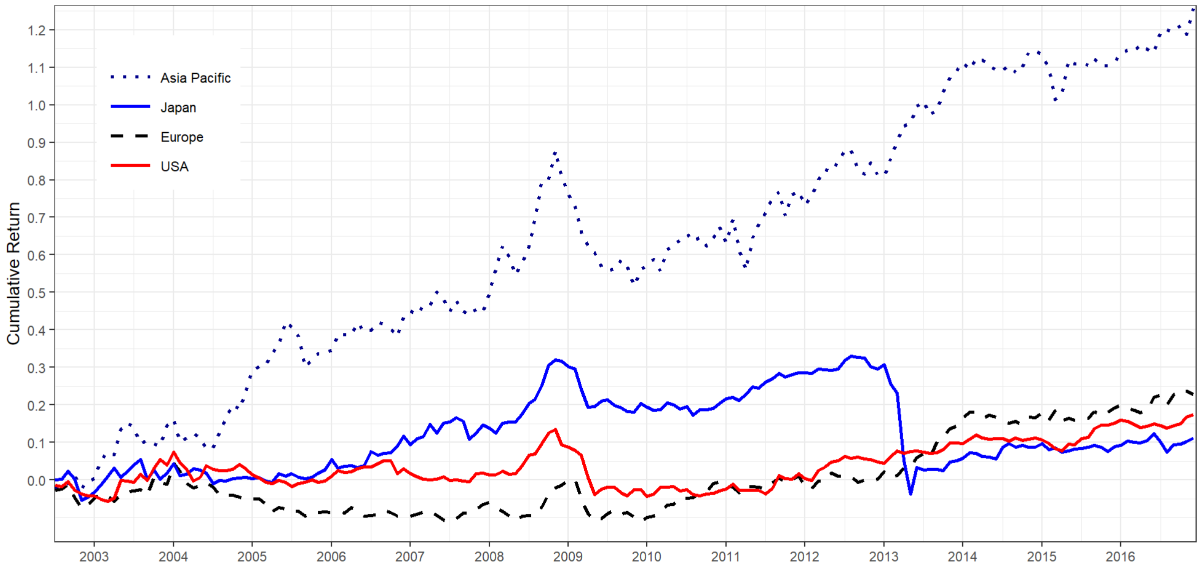


Figure 3.4: **Cumulative returns on the Portfolio-mixing Strategy.**

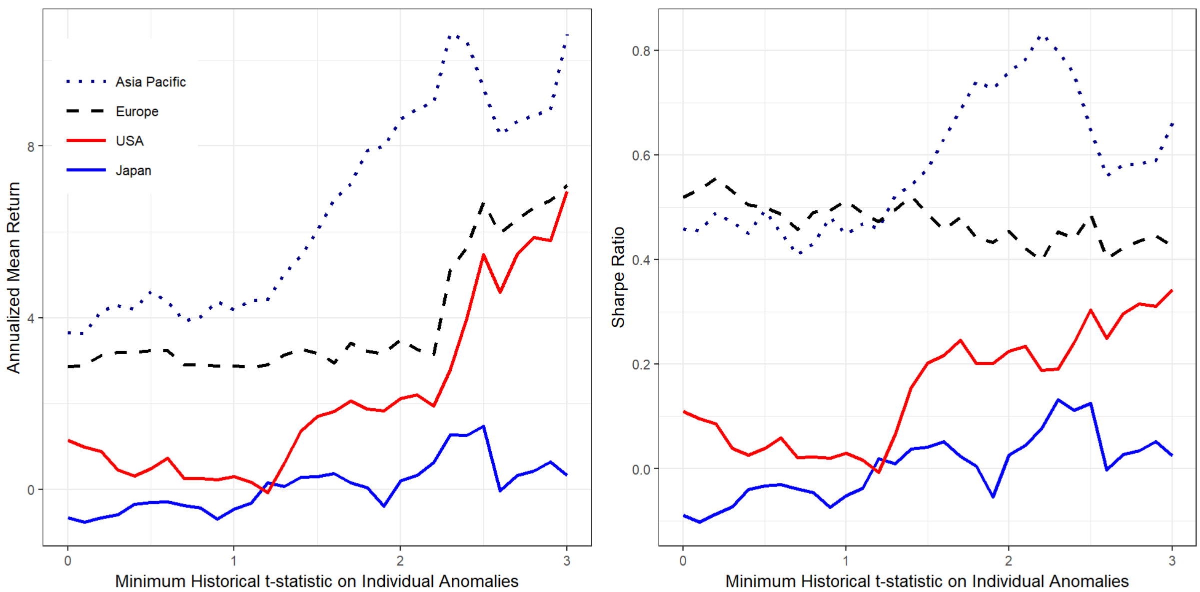


Figure 3.5: **Annualized Mean Returns and Sharpe Ratios on the Portfolio-mixing Strategy Depending on Significance Threshold for Individual Anomalies.**

ing with the threshold in all the region. The mean return on the anomalies are therefore larger the more significant they were historically. Sharpe ratio of the portfolio-mixing strategy does not depend that strongly on the significance threshold.[21]

To conclude, returns on the anomalies in all the regions are positive, which suggests that it is profitable to invest in the anomalies before transaction costs.

3.2.2 Mispricing Strategy

Performance of the mispricing strategy is examined next. Jacobs and Müller (2017a) showed that the mispricing strategy estimated with least squares leads to higher returns in both absolute term and on risk adjusted basis relative to mixing of portfolios on individual anomalies. Gu et al. (2018) then documented that the more sophisticated machine learning methods provide higher out-of-sample predictability relative to least squares. The machine learning methods are extended to the international sample to determine whether their benefits persist outside the US.

Table 3.4 presents mean returns on portfolios created, based on the mispricing strategy. The regressions of stock returns on their characteristics are fit on data available up to June every year and the future stock returns are then predicted with the latest available characteristics for each of the next 12 months. The regressions are estimated with least squares, penalized least squares, random forests, gradient boosting regression trees, and neural networks. The estimates in table 3.4 are based on the US data from July 1963. The long-short decile portfolios that invest into stocks in the top decile of the predicted future returns and short-sell stocks in the bottom decile of the predicted returns are then created. The reported returns on portfolios are in percent per month and are from July 1995 to December 2016.

Both the tree based methods and neural networks outperform simple least squares. In particular, gradient boosting regression trees and neural networks outperform least squares in all the regions for both mean returns and risk adjusted Sharpe ratios. The machine learning methods are therefore more powerful for stock return predictions outside the US as well as inside the US. The superior performance outside the US provides robustness to findings in Gu et al. (2018) who focused solely on the US. The average returns on the mispricing strategies are about 4 times higher than for the portfolio level strategy in the previous section.

Gradient boosting regression trees and neural networks also have the smallest maximum drawdowns and investing in them is therefore the least risky. Diversification over the four regions (in the global columns) further reduces the maximum drawdowns and increases the Sharpe ratios.

Figure 3.6 plots cumulative returns on the gradient boosting regression tree mispricing strategy in Table 3.4. The returns are presented in decimal logarithms and 1 on the left scale therefore corresponds to 1000% return on the initial investment. There is a small

[21] Note that there are only few anomalies with t-statistic larger than 2.5 and the results become unstable after that.

Table 3.4:
Performance of the Mispricing Strategy Estimated in the US
The table shows out-of-sample performance of the mispricing strategy as defined in Section 3.1.4. It is based on long-short decile portfolios from the strategy that combines all the available anomalies through predictive regressions of individual stock returns on transformed characteristics. The estimation methods are least squares, penalized least squares, random forests, gradient boosting regression trees, or neural networks. That is, pooled regressions of monthly stock returns on cross-sectional quantiles of their characteristics observable before each month start are estimated and future returns from the latest available characteristics are predicted. The value-weighted or equal-weighted long-short portfolios are constructed by buying stocks in the top decile of the predicted next month returns and shorting stocks in the bottom decile of the predicted next month returns. The regressions are rerun at the end of each June with only those anomalies that have been published by that time. The out-of-sample performance is observed in the US, Europe, Japan, and Asia Pacific. The training sample spans July 1963 to December 2016 in the US and July 1990 to December 2016 in other regions. The regressions are estimated only on the past US data and the future returns are predicted in all the regions. The reported returns are for July 1995 to December 2016 period and are in percentage points.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Weighted Least Squares										
Mean	0.801	0.680	0.922	0.782	0.810	0.575	0.647	0.648	0.633	0.639
Sharpe Ratio	0.479	0.541	0.701	0.497	0.763	0.348	0.472	0.410	0.318	0.550
Skewness	-0.340	0.239	-0.425	-0.356	0.025	-0.121	-0.017	-0.688	-0.238	-0.041
Kurtosis	8.521	5.673	4.167	3.713	8.488	7.214	6.483	5.544	4.824	7.880
Max Drawdown	-64.70	-37.10	-43.36	-47.61	-43.51	-69.75	-34.16	-44.52	-49.86	-50.85
Penalized Weighted Least Squares										
Mean	0.756	0.728	0.886	0.854	0.800	0.644	0.794	0.596	0.754	0.683
Sharpe Ratio	0.443	0.557	0.665	0.526	0.724	0.381	0.554	0.372	0.365	0.568
Skewness	-0.487	0.067	-0.620	-0.365	-0.263	-0.316	0.123	-0.692	-0.409	0.023
Kurtosis	8.703	6.662	4.540	3.834	9.301	7.297	7.579	5.715	5.080	9.111
Max Drawdown	-65.36	-35.46	-42.32	-48.50	-45.10	-68.02	-37.51	-49.44	-57.38	-49.59
Gradient Boosting Regression Trees										
Mean	1.165	0.870	1.173	1.650	1.163	1.391	0.591	1.011	1.415	1.033
Sharpe Ratio	0.720	0.644	0.766	1.005	1.146	0.831	0.412	0.525	0.800	0.870
Skewness	0.319	-1.160	0.682	-0.437	-0.449	0.561	-1.314	0.800	-0.112	-0.433
Kurtosis	6.653	10.17	8.274	5.575	6.812	9.287	12.03	8.611	4.718	7.797
Max Drawdown	-38.31	-48.25	-34.37	-36.65	-27.45	-43.93	-42.31	-41.79	-39.58	-35.62
Random Forest										
Mean	1.050	1.037	1.107	0.943	1.074	0.977	0.339	1.028	1.183	0.798
Sharpe Ratio	0.703	0.782	0.781	0.520	1.080	0.691	0.222	0.591	0.612	0.726
Skewness	-0.328	-1.132	-0.283	-0.789	-0.939	-0.594	-1.149	0.675	-0.062	-0.974
Kurtosis	5.989	9.399	5.558	7.201	7.191	4.951	12.27	8.613	5.855	7.549
Max Drawdown	-30.69	-48.18	-40.16	-46.87	-27.76	-30.59	-54.54	-42.12	-39.57	-31.17
Neural Networks										
Mean	1.416	1.097	1.295	1.752	1.346	1.420	0.826	1.100	1.177	1.093
Sharpe Ratio	0.905	0.880	1.130	1.086	1.582	0.905	0.649	0.693	0.697	1.042
Skewness	-0.083	-0.082	-0.149	0.244	-0.310	-0.167	-0.470	0.629	0.638	-0.255
Kurtosis	7.316	4.827	4.446	5.091	5.304	6.432	7.050	10.37	5.075	6.806
Max Drawdown	-44.60	-33.93	-24.70	-38.10	-18.90	-48.11	-31.93	-37.09	-54.45	-33.25

drop in profitability around 2003 in the US, which is in line with the evidence from portfolio-mixing strategy in Figure 3.4. The mispricing strategy is the least profitable in

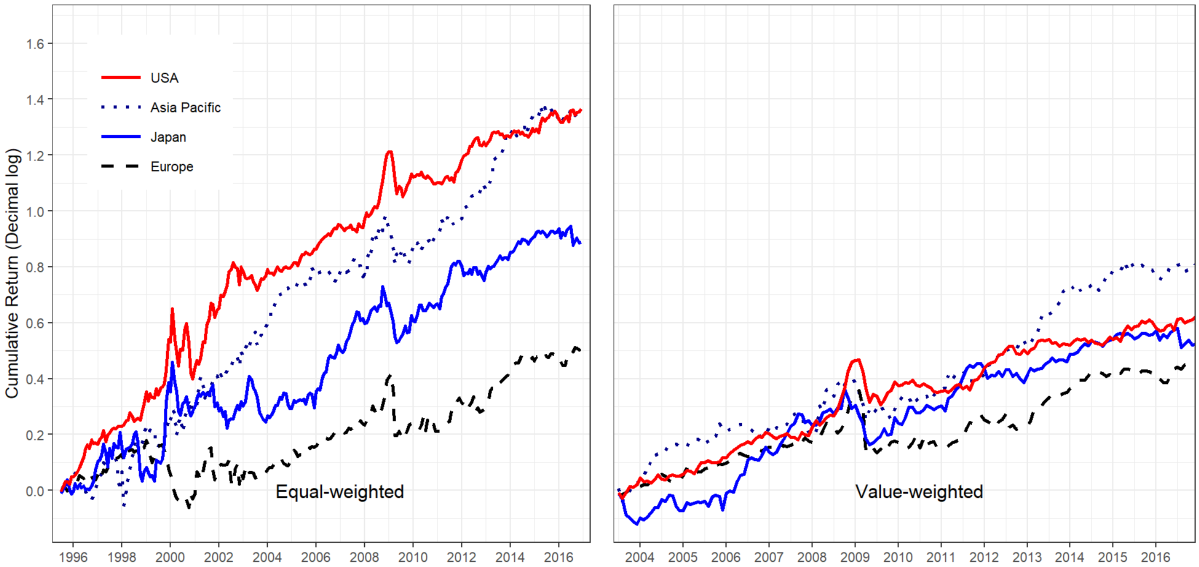


Figure 3.6: **Cumulative Returns on the Gradient Boosting Regression Trees Mispricing Strategy.** The figure shows cumulative returns for the mispricing strategy as described in Table 3.4 that is estimated on the individual stocks from the US.

the European region.

Long-only and Short-only Components of the Strategy

Short-selling can be connected to large costs and sometimes even outright impossible. That is why it might not be possible to replicate the returns on the mispricing strategy in practice.[22] The long-short strategy in Table 3.4 will now be decomposed into long-only and short-only components to determine the role of short-selling for the strategy's profitability. Table 3.5 decomposes the long-short returns separately for the individual machine learning methods. The long-only component can be compared to equal-weighted and value-weighted returns on the whole market as defined by the liquid universe of stocks in Panel A.

The panel A in the table documents that the mispricing strategy is more profitable than the whole market in all the regions. The long-only component is responsible for most of returns on the mispricing strategy. The short-only component then mainly serves as a hedge that increases Sharpe ratio and lowers maximum drawdown. The returns on long-only component of gradient boosting regression tree mispricing strategy are about 5% a year larger than returns on the market. The other machine learning methods also outperform the market.

The more advanced machine learning methods outperform simple least squares both on the short side and long side. To conclude, the positive returns on the mispricing strategy are robust to short-selling constrains. Even short-selling-constrained investors

[22]Short-selling constrains should not be a large issue on our liquid universe of stocks. Andrikopoulos et al. (2013) showed that although some stocks cannot be short-sold in practice, focusing only on those that can be short-sold does not statistically diminish returns on 8 quantitative strategies in the UK. They also showed that short-selling costs are small at about 1% annually in the UK.

Table 3.5:
Decomposition of the Returns on the Mispricing Strategy to Long-only and Short-only Components

The table shows returns of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. The returns on the long-short portfolios are decomposed to long-only and short-only components. Equal-weighted and value-weighted returns on the whole stock markets in the individual regions estimated on the liquid sample of stocks are also provided.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Panel A: Long-only Component of the Mispricing Strategy										
Whole Market										
Mean	0.829	0.739	0.300	0.737	0.655	0.786	0.668	0.222	0.806	0.617
Sharpe Ratio	0.507	0.445	0.184	0.376	0.439	0.609	0.442	0.145	0.471	0.477
Skewness	-0.656	-0.540	0.075	-0.565	-0.684	-0.651	-0.539	0.052	-0.492	-0.738
Kurtosis	4.391	5.215	3.260	5.405	5.447	3.883	4.249	3.103	4.707	4.384
Max Drawdown	-60.81	-63.61	-58.18	-64.72	-56.68	-51.41	-59.46	-65.98	-59.07	-54.54
Weighted Least Squares										
Mean	1.099	1.066	0.700	0.786	0.946	0.889	0.990	0.689	0.923	0.812
Sharpe Ratio	0.626	0.585	0.380	0.323	0.605	0.570	0.534	0.382	0.420	0.549
Skewness	-0.703	-0.611	0.072	-1.103	-0.811	-0.654	-0.535	-0.012	-0.814	-0.513
Kurtosis	5.328	5.045	3.908	8.174	5.229	5.225	5.024	3.970	8.804	4.554
Max Drawdown	-56.12	-62.66	-60.40	-76.62	-59.38	-48.56	-59.94	-63.29	-65.16	-52.87
Information Ratio	0.339	0.505	0.525	0.052	0.342	0.120	0.390	0.465	0.093	0.289
Penalized Weighted Least Squares										
Mean	1.036	1.063	0.643	0.848	0.913	0.867	1.064	0.614	1.057	0.794
Sharpe Ratio	0.595	0.594	0.352	0.343	0.592	0.563	0.582	0.347	0.480	0.552
Skewness	-0.837	-0.608	0.060	-1.094	-0.843	-0.749	-0.532	-0.063	-0.700	-0.606
Kurtosis	5.750	5.178	3.707	8.003	5.271	5.159	5.153	3.555	9.061	4.542
Max Drawdown	-56.87	-63.39	-59.58	-75.79	-59.31	-49.53	-58.57	-59.86	-67.70	-53.04
Information Ratio	0.251	0.480	0.466	0.108	0.300	0.093	0.471	0.405	0.201	0.265
Gradient Boosting Regression Trees										
Mean	1.235	1.154	0.676	1.414	1.078	1.367	0.986	0.653	1.396	1.084
Sharpe Ratio	0.569	0.654	0.357	0.600	0.629	0.684	0.586	0.360	0.625	0.650
Skewness	-0.338	-0.717	0.191	-0.530	-0.602	-0.020	-0.444	0.399	-0.485	-0.347
Kurtosis	6.051	5.357	3.818	5.681	4.314	6.596	4.746	5.293	5.455	4.035
Max Drawdown	-71.09	-63.32	-61.80	-63.02	-57.61	-65.67	-61.14	-73.43	-61.49	-62.35
Information Ratio	0.456	0.718	0.472	0.725	0.468	0.500	0.487	0.448	0.498	0.624
Random Forest										
Mean	1.127	1.275	0.577	0.971	0.994	0.951	0.968	0.620	1.003	0.868
Sharpe Ratio	0.527	0.709	0.315	0.396	0.585	0.523	0.553	0.327	0.446	0.547
Skewness	-0.985	-0.688	0.143	-0.501	-0.788	-0.975	-0.603	0.380	-0.316	-0.576
Kurtosis	6.545	5.300	3.468	5.211	4.452	6.248	5.125	4.849	4.467	3.825
Max Drawdown	-76.51	-62.27	-64.61	-75.60	-62.19	-69.23	-61.92	-74.02	-70.31	-63.97
Information Ratio	0.356	0.941	0.388	0.243	0.382	0.185	0.394	0.438	0.153	0.402
Neural Networks										
Mean	1.295	1.262	0.756	1.381	1.140	1.260	1.160	0.696	1.351	1.081
Sharpe Ratio	0.576	0.650	0.404	0.555	0.649	0.625	0.638	0.368	0.613	0.632
Skewness	-0.354	-0.081	0.151	-0.464	-0.431	-0.752	-0.313	0.301	-0.167	-0.505
Kurtosis	5.683	5.340	3.336	5.736	3.969	6.118	4.639	4.678	5.748	4.296
Max Drawdown	-74.67	-61.71	-58.40	-71.28	-57.78	-74.06	-60.45	-69.03	-60.95	-68.26
Information Ratio	0.480	0.734	0.694	0.611	0.500	0.418	0.593	0.493	0.450	0.577

can therefore benefit from the strategy.

Table 3.5 Continued

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Panel B: Short-only Component of the Mispricing Strategy										
Weighted Least Squares										
Mean	0.297	0.386	-0.223	0.004	0.136	0.313	0.342	0.041	0.290	0.173
Sharpe Ratio	0.131	0.206	-0.112	0.002	0.077	0.166	0.198	0.021	0.118	0.109
Skewness	0.119	-0.557	0.376	-0.204	-0.234	-0.142	-0.744	0.565	0.045	-0.492
Kurtosis	5.883	4.473	3.564	4.335	4.644	5.892	4.741	3.966	5.797	5.035
Max Drawdown	-83.94	-85.78	-71.55	-85.03	-72.04	-78.71	-79.76	-81.47	-86.74	-67.48
Penalized Weighted Least Squares										
Mean	0.280	0.335	-0.243	-0.006	0.114	0.222	0.270	0.017	0.303	0.111
Sharpe Ratio	0.123	0.172	-0.121	-0.003	0.063	0.116	0.149	0.009	0.123	0.068
Skewness	0.123	-0.414	0.349	-0.070	-0.166	-0.024	-0.781	0.518	0.156	-0.510
Kurtosis	5.582	5.149	3.510	4.296	4.659	5.650	5.428	3.824	5.814	5.204
Max Drawdown	-83.56	-86.54	-70.06	-85.41	-72.34	-72.79	-78.49	-79.48	-86.34	-66.80
Gradient Boosting Regression Trees										
Mean	0.069	0.284	-0.497	-0.236	-0.085	-0.023	0.395	-0.358	-0.019	0.051
Sharpe Ratio	0.029	0.117	-0.219	-0.087	-0.042	-0.012	0.174	-0.162	-0.007	0.028
Skewness	-0.296	0.008	0.182	0.474	-0.224	-0.395	0.177	0.087	0.019	-0.353
Kurtosis	4.847	6.517	3.632	6.785	5.078	5.682	6.697	3.818	5.182	6.167
Max Drawdown	-79.96	-88.78	-66.42	-79.12	-68.60	-79.31	-87.79	-68.29	-78.39	-68.80
Random Forest										
Mean	0.077	0.237	-0.529	0.028	-0.080	-0.026	0.628	-0.408	-0.180	0.070
Sharpe Ratio	0.031	0.098	-0.237	0.010	-0.039	-0.012	0.274	-0.186	-0.067	0.037
Skewness	-0.263	0.185	0.288	0.723	-0.185	-0.294	0.404	0.173	0.755	-0.275
Kurtosis	4.697	7.201	3.800	8.629	4.968	5.249	9.564	3.687	9.069	6.330
Max Drawdown	-80.78	-90.00	-62.61	-90.70	-71.49	-79.21	-93.16	-65.70	-83.39	-73.58
Neural Networks										
Mean	-0.121	0.165	-0.539	-0.371	-0.206	-0.159	0.333	-0.404	0.175	-0.012
Sharpe Ratio	-0.054	0.075	-0.254	-0.142	-0.109	-0.084	0.164	-0.198	0.073	-0.007
Skewness	-0.353	-0.134	0.218	0.031	-0.210	-0.416	-0.184	0.226	-0.079	-0.370
Kurtosis	5.104	6.669	3.433	4.608	4.942	5.608	6.763	3.867	4.007	5.798
Max Drawdown	-80.06	-83.69	-67.02	-80.60	-67.65	-77.82	-80.68	-68.21	-86.50	-64.90

Table 3.6:

Performance of the Mispricing Strategy on Risk-adjusted Basis

The table shows returns of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US adjusted for capital asset pricing model (CAPM) model and five Fama-French factors (FF5). The standard errors in t-statistics are adjusted for heterockedasticity and autocorrelation with Newey-West adjustment for up to 12 lags.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Weighted Least Squares										
Mean Return	0.801	0.680	0.922	0.782	0.810	0.575	0.647	0.648	0.633	0.639
	2.043	2.125	2.597	2.533	2.942	1.495	2.051	1.832	1.817	2.203
CAPM Alpha	0.951	0.672	0.930	0.744	0.851	0.688	0.588	0.657	0.707	0.666
	2.604	2.392	2.997	1.927	3.482	1.974	1.967	1.936	1.695	2.710
FF5 Alpha	0.328	0.133	0.672	0.232	0.263	0.080	0.134	0.402	-0.158	0.173
	1.114	0.543	2.189	0.562	1.336	0.262	0.486	1.121	-0.343	0.732
Penalized Weighted Least Squares										
Mean Return	0.756	0.728	0.886	0.854	0.800	0.644	0.794	0.596	0.754	0.683
	1.960	2.191	2.538	2.451	2.800	1.599	2.632	1.645	1.810	2.294
CAPM Alpha	0.922	0.760	0.894	0.839	0.864	0.767	0.776	0.606	0.832	0.737
	2.505	2.629	2.844	2.032	3.391	2.132	2.513	1.750	1.825	2.869
FF5 Alpha	0.332	0.179	0.604	0.283	0.252	0.128	0.299	0.320	-0.079	0.192
	1.093	0.750	1.998	0.639	1.187	0.412	1.105	0.898	-0.160	0.836
Gradient Boosting Regression Trees										
Mean Return	1.165	0.870	1.173	1.650	1.163	1.391	0.591	1.011	1.415	1.033
	4.266	2.978	4.021	5.470	6.641	4.465	2.198	2.662	5.236	4.998
CAPM Alpha	1.221	1.047	1.184	1.737	1.242	1.406	0.733	1.028	1.511	1.095
	3.630	3.589	3.643	4.883	5.999	3.726	2.539	2.408	4.109	4.283
FF5 Alpha	1.143	0.338	1.073	1.580	1.057	1.600	0.212	0.936	1.160	1.164
	3.680	1.116	3.222	3.841	4.377	4.834	0.730	2.211	2.595	4.535
Random Forest										
Mean Return	1.050	1.037	1.107	0.943	1.074	0.977	0.339	1.028	1.183	0.798
	4.103	4.108	4.038	3.269	5.894	3.932	1.270	2.386	3.222	4.391
CAPM Alpha	1.157	1.200	1.118	1.035	1.172	1.050	0.475	1.039	1.318	0.887
	3.614	4.407	3.469	3.064	5.526	3.253	1.564	2.430	3.566	3.762
FF5 Alpha	0.782	0.433	0.919	0.871	0.727	0.740	-0.092	0.836	0.971	0.626
	2.748	1.502	2.787	2.258	3.124	2.626	-0.275	1.974	2.304	2.343
Neural Networks										
Mean Return	1.416	1.097	1.295	1.752	1.346	1.420	0.826	1.100	1.177	1.093
	4.336	3.734	5.759	4.917	7.829	4.442	3.352	3.240	3.342	5.627
CAPM Alpha	1.402	1.179	1.301	1.788	1.383	1.354	0.903	1.108	1.247	1.103
	3.928	4.057	5.257	4.575	7.853	3.471	3.413	3.285	3.540	4.586
FF5 Alpha	1.482	1.038	1.185	1.435	1.334	1.584	0.758	1.008	0.749	1.323
	5.018	3.885	4.581	3.527	7.754	5.009	2.985	2.941	2.257	5.895

Risk-adjusted Performance of the Strategy

We have so far focused only on raw returns on the mispricing strategy without accounting for any risk factors. Table 3.6 presents performance of the strategy after accounting for market returns and five Fama-French factors. Accounting for market return should have little impact on the performance of the strategy since it is long-short, and thus close to market neutral, by construction. Table 3.6 confirms that it is indeed the case and capital asset pricing model (CAPM) alpha is close to the mean returns for all the estimation methods. The results are, however, very different when adjusting for five Fama-French factors. There is again almost no difference between the mean returns and alphas for more complicated estimation methods but there is a visible deterioration in risk-adjusted performance for the linear estimation methods. The linear estimation methods therefore lead to mispricing signal that is close to the traditional risk factors.

To conclude, the profitability of the mispricing strategy is significant even at risk-adjusted basis. The more complicated estimation methods then lead to returns that are unrelated to the traditional risk factors.

3.3 The Role of International Evidence

The evidence so far documented that the anomalies identified based on the past data in the US are profitable out-of-sample in all the regions. Can international data outside the US be used to better select the winning strategies?

There are some arguments for the usefulness of the international data. The international data increases sample size and therefore limits the possibility for data-mining and in-sample overfitting. The larger sample size also generally provides larger power to statistical tests which should lead to more precise selection of truly significant strategies. The international evidence extends the sample size mainly in the most recent period. The most recent data is also the most useful as the financial markets are changing rapidly and the older data may not be relevant anymore.

There are, however, also some problems with suitability of the international evidence. The individual global regions have very different institutional settings. Bankruptcy laws, tax laws, investor protection, and accounting standards vary widely across the regions. The institutional differences can lower the usefulness of historical data outside the respective regions. The larger estimation sample improves forecasts through consistency. The consistency, however, works only if the underlying true drivers of stock returns are uniform over the regions, which is in no way guaranteed.

The previous machine learning evidence was based on predictive regressions estimated solely on data from the US. This section first investigates whether estimating the predictive regressions in the respective regions is more suitable than estimating them only on data from the US. It then explores whether combining estimation samples from the individual regions can improve the profitability to the mispricing strategy.

There is surprisingly only a small difference between returns on strategies that are estimated on data from the US in table 3.4 and those that are estimated on data in the respective regions in table 3.7. One explanation for the similarity is that the sample size in the US is already large enough to capture the true drivers of stock returns that are globally valid. One exception is Asia Pacific region where there are only a few liquid stocks historically, which makes the predictive regressions imprecise. The performance of the mispricing strategy in Japan is also notably worse than when estimated on the US data. The explanation is again simple. Japan has undergone a slow eruption of an asset price bubble at the beginning of the estimation sample in early 1990s. The estimated relationships that are valid for this specific period fare badly out-of-sample where the stock market dynamics go back to their normal state.

Table 3.7:
Performance of the Mispricing Strategy Estimated in the Individual Regions
The table shows out-of-sample performance of the mispricing strategy as described in Table 3.4. The predictive regressions for individual stock returns are estimated in each respective region.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Weighted Least Squares										
Mean	0.801	0.682	0.750	1.117	0.811	0.575	0.483	0.348	0.876	0.596
Sharpe Ratio	0.479	0.450	0.396	0.507	0.786	0.348	0.338	0.157	0.375	0.541
Skewness	-0.340	-1.280	0.353	-1.957	-0.200	-0.121	-0.995	-0.255	-1.500	-0.349
Kurtosis	8.521	10.03	7.663	20.65	6.543	7.214	7.863	7.841	18.58	6.483
Max Drawdown	-64.70	-50.52	-66.56	-61.71	-36.44	-69.75	-42.84	-72.33	-63.57	-44.58
Penalized Weighted Least Squares										
Mean	0.756	0.753	0.701	1.333	0.823	0.644	0.616	0.423	1.359	0.656
Sharpe Ratio	0.443	0.458	0.363	0.535	0.749	0.381	0.393	0.197	0.534	0.577
Skewness	-0.487	-1.026	0.360	-1.665	-0.278	-0.316	-1.033	0.329	-1.072	-0.376
Kurtosis	8.703	8.700	7.280	21.06	6.723	7.297	8.447	6.321	15.03	7.534
Max Drawdown	-65.36	-46.61	-66.88	-69.25	-39.84	-68.02	-43.74	-58.29	-63.68	-39.76
Gradient Boosting Regression Trees										
Mean	1.165	0.725	0.951	1.766	1.107	1.391	0.319	0.678	1.522	0.915
Sharpe Ratio	0.720	0.596	0.636	0.761	1.183	0.831	0.238	0.400	0.581	0.850
Skewness	0.319	-0.884	0.445	-0.346	-0.012	0.561	-1.250	0.071	0.026	-0.450
Kurtosis	6.653	7.508	5.686	19.50	5.987	9.287	7.699	4.559	13.42	7.646
Max Drawdown	-38.31	-45.14	-34.11	-56.04	-22.56	-43.93	-58.13	-55.73	-55.96	-31.15
Random Forest										
Mean	1.050	0.353	1.022	0.960	0.892	0.977	0.140	0.792	1.112	0.711
Sharpe Ratio	0.703	0.265	0.779	0.544	1.007	0.691	0.094	0.503	0.516	0.688
Skewness	-0.328	-1.281	-0.201	0.591	-0.408	-0.594	-1.111	0.201	0.768	-0.953
Kurtosis	5.989	9.421	4.537	6.862	6.323	4.951	6.857	4.150	9.382	6.801
Max Drawdown	-30.69	-51.84	-32.79	-52.27	-22.31	-30.59	-60.13	-47.42	-51.77	-29.88
Neural Networks										
Mean	1.416	0.748	0.958	1.192	1.133	1.420	0.561	0.616	0.986	0.988
Sharpe Ratio	0.905	0.544	0.572	0.592	1.308	0.905	0.383	0.305	0.423	1.025
Skewness	-0.083	-0.637	0.464	-0.435	0.026	-0.167	-0.696	0.054	-0.151	-0.465
Kurtosis	7.316	6.991	7.300	5.796	4.891	6.432	7.182	8.016	5.195	6.920
Max Drawdown	-44.60	-50.30	-48.09	-55.31	-18.16	-48.11	-37.60	-72.91	-68.84	-21.88

Table 3.8 shows mean returns and other performance statistics for gradient boosting regression trees mispricing strategy as in table 3.4. The only difference with respect to table 3.4 is that the future individual stock returns are predicted from regressions estimated on historical data that are not solely from the US. Predictive regressions with training sample from the US, the US & Japan, the US & Europe, or the US & Japan & Europe & Asia Pacific are compared. These three regions cover most of the developed markets and global stock market capitalization. Corresponding evidence for least square mispricing strategy is provided in the Appendix J.[23]

[23]It is omitted here for the sake of space as all the findings are very similar to gradient boosting regression trees.

Table 3.8:

Performance of the Mispricing Strategy Estimated on the International Data

The table shows returns of the mispricing strategy based on gradient boosting regression trees described in Table 3.4. The historical predictive regressions are estimated on individual stocks from combinations of the four covered regions: the US, Japan, Europe, Europe, Asia Pacific.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Estimated in the US										
Mean	1.165	0.870	1.173	1.650	1.163	1.391	0.591	1.011	1.415	1.033
Sharpe Ratio	0.720	0.644	0.766	1.005	1.146	0.831	0.412	0.525	0.800	0.870
Skewness	0.319	-1.160	0.682	-0.437	-0.449	0.561	-1.314	0.800	-0.112	-0.433
Kurtosis	6.653	10.17	8.274	5.575	6.812	9.287	12.03	8.611	4.718	7.797
Max Drawdown	-38.31	-48.25	-34.37	-36.65	-27.45	-43.93	-42.31	-41.79	-39.58	-35.62
Estimated in the US and Cross-validated on the International Data										
Mean	1.114	1.012	1.021	1.192	1.085	1.318	0.665	0.675	0.846	0.968
Sharpe Ratio	0.704	0.747	0.686	0.655	1.057	0.776	0.473	0.355	0.410	0.785
Skewness	0.215	-0.422	0.313	-1.584	-0.558	0.270	-0.673	0.474	-1.266	-0.146
Kurtosis	6.706	11.44	8.626	12.11	7.354	7.886	13.89	8.906	12.23	9.259
Max Drawdown	-36.49	-44.69	-39.01	-49.81	-30.07	-39.70	-43.57	-44.37	-61.01	-35.13
Information Ratio	-0.108	0.235	-0.285	-0.415	-0.265	-0.108	0.084	-0.399	-0.379	-0.115
Estimated in the US & Japan										
Mean	1.317	1.015	1.353	1.537	1.289	1.616	0.812	1.001	1.722	1.262
Sharpe Ratio	0.809	0.911	1.030	1.016	1.413	0.907	0.647	0.550	1.008	1.093
Skewness	0.602	0.715	0.150	0.176	0.380	0.876	0.088	0.837	0.362	0.828
Kurtosis	7.612	8.349	6.909	3.599	8.273	10.26	9.448	10.76	3.516	9.092
Max Drawdown	-34.18	-25.91	-25.82	-29.65	-19.15	-42.98	-30.11	-45.32	-29.95	-31.87
Information Ratio	0.186	0.153	0.172	-0.083	0.224	0.208	0.203	-0.007	0.191	0.295
Estimated in the US & Europe										
Mean	1.361	1.016	1.173	1.555	1.268	1.513	0.716	0.786	1.397	1.111
Sharpe Ratio	0.854	0.812	0.763	0.892	1.241	0.875	0.501	0.431	0.654	0.944
Skewness	0.049	0.159	0.251	-1.266	-0.313	0.537	-0.463	-0.063	-1.763	-0.055
Kurtosis	6.672	6.745	6.599	11.62	6.270	8.410	7.574	5.445	17.04	6.332
Max Drawdown	-38.16	-40.92	-34.41	-45.58	-27.77	-38.96	-47.54	-49.58	-56.36	-30.43
Information Ratio	0.281	0.197	-0.000	-0.085	0.248	0.150	0.129	-0.214	-0.010	0.130
Estimated in the US & Japan & Europe & Asia Pacific										
Mean	1.325	1.009	1.281	2.295	1.373	1.432	0.955	1.066	2.317	1.225
Sharpe Ratio	0.808	0.870	0.960	1.486	1.394	0.803	0.680	0.615	1.056	1.048
Skewness	0.257	0.262	0.114	0.624	-0.034	0.991	0.745	-0.013	0.750	0.451
Kurtosis	7.133	5.071	5.931	5.360	7.274	11.46	7.285	4.745	8.252	7.276
Max Drawdown	-37.15	-33.19	-24.69	-20.31	-26.01	-38.84	-26.46	-44.00	-53.48	-30.43
Information Ratio	0.191	0.131	0.114	0.530	0.404	0.036	0.295	0.045	0.483	0.251

The table provides mixed results on the value of international evidence. There is a small gain from adding the international stocks to local training sample in the US for equal-weighted portfolios. Historical data in the US is therefore completely sufficient for the future predictions in the US. Profitability of the mispricing strategy in Europe improves with predictions based on the estimation sample from the US and Europe relative to from the US only. The profitability in Japan also improves with training sample from both the US and Japan instead of from the US only. The largest gains in profitability are in Asia Pacific region where training samples from Japan and Europe are jointly beneficial.

The table also shows the gradient boosting regression tree mispricing strategy estimated in the US using parameters cross-validated in the other three regions. The cross-validation on data outside the US could add some predictive power as the validation sample is coming from more recent period than when the training sample is from the US only. The table, however, documents that there is no gain from cross-validating outside the US.

To conclude, the regional institutional setting is indeed an important determinant of stock return drivers. There is no gain for the US investor to seek international evidence for quantitative strategies. The larger statistical power, caused by the larger sample, seems to be completely offset by the differences in institutional setting.

3.4 Importance of New Anomalies for Profitability of the Strategies

Figure 3.1 documented that the number of published anomalies is increasing roughly linearly over time. Harvey et al. (2016) found even sharper increase for published as well as unpublished anomalies. Researchers are looking at the same data again and again to find the new anomalies which should lead to a large proportion of false positive discoveries. The proportion of false discoveries is expected to increase over time as the strongest anomalies are likely already published. Harvey et al. (2016) therefore concluded that most of the recently published studies can be explained by data-mining and the standard critical values for statistical significance no longer apply. The data-mining should also lead to a lower predictive power of the new anomalies. Individual studies introducing new anomalies almost never properly control for all anomalies published previously. Many of the new anomalies are therefore subsumed by existing anomalies in proper multiple hypothesis setting as documented by Green et al. (2017).

Most of the widely accepted anomalies have been published before 1995.[24] It is therefore worth studying whether the more recently published drivers of stock returns are also important. This section investigates the marginal value of recently published anomalies for profitability of the mispricing strategy after accounting for anomalies published earlier.

Table 3.9 presents mean returns and Sharpe ratios on the mispricing strategy as specified in table 3.4 but with further restrictions on the universe of anomalies. The mispricing strategy is estimated using anomalies that were published before 1995, 2000, or 2005. Its performance is then tracked over the 2005-2016 period.[25] The different sets of anomalies provide a good indication for marginal value of the new signals published after 1995, while accounting for anomalies published before 1995.

There are improvements in mean returns and Sharpe ratios for both the equal-weighted and value-weighted portfolios in the US with addition of tje new anomalies. The new anomalies therefore have significant incremental value for out-of-sample forecasts. This benefit is smaller in Japan and Europe. The results are similar for both least squares and

[24]For example heavily cited size and book-to-value factor in Fama and French (1992) were introduced before 1990.

[25]Adding another set of anomalies published before 2010 and focusing on 2010-2016 out-of-sample period leads to identical findings. The corresponding results are available in table J.2 in the Appendix J.

Table 3.9:
Are the More Recent Anomalies Improving Profitability of the Mispricing Strategy?
The table shows returns of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. Anomalies in the estimation are restricted to those that were published before 1995, 2000, or 2005. The returns are reported in percentage points per month over the 2005-2016 period.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Weighted Least Squares										
Published by 1995										
Mean	0.044	0.082	0.361	0.294	0.169	0.268	0.297	0.914	0.284	0.256
Sharpe Ratio	0.034	0.068	0.266	0.194	0.171	0.210	0.226	0.615	0.155	0.234
Published by 2000										
Mean	0.055	0.278	0.341	-0.106	0.164	0.119	0.306	0.595	-0.238	0.232
Sharpe Ratio	0.047	0.237	0.234	-0.072	0.177	0.093	0.241	0.363	-0.140	0.234
Information Ratio	0.016	0.275	-0.025	-0.374	-0.010	-0.180	0.011	-0.326	-0.394	-0.038
Published by 2005										
Mean	0.612	0.078	0.854	0.270	0.506	0.685	0.105	1.087	-0.004	0.556
Sharpe Ratio	0.564	0.074	0.646	0.189	0.624	0.644	0.087	0.787	-0.002	0.684
Information Ratio	0.856	-0.265	0.702	0.291	0.724	0.545	-0.173	0.482	0.134	0.444
Gradient Boosting Regression Trees										
Published by 1995										
Mean	0.364	0.276	0.868	0.816	0.531	0.411	0.192	0.917	0.197	0.289
Sharpe Ratio	0.369	0.216	0.772	0.521	0.692	0.377	0.133	0.766	0.114	0.322
Published by 2000										
Mean	0.435	0.647	0.979	1.058	0.722	0.250	0.476	0.951	0.943	0.430
Sharpe Ratio	0.408	0.483	0.873	0.727	0.901	0.234	0.313	0.720	0.602	0.461
Information Ratio	0.091	0.453	0.122	0.226	0.430	-0.181	0.267	0.026	0.477	0.227
Published by 2005										
Mean	0.824	0.602	1.212	1.043	0.904	0.948	0.381	1.138	0.414	0.777
Sharpe Ratio	0.842	0.537	1.121	0.864	1.309	1.012	0.332	0.980	0.314	1.135
Information Ratio	0.585	-0.054	0.276	-0.012	0.403	0.819	-0.090	0.169	-0.349	0.515

gradient boosting regression trees methods but the returns from least squares are much more volatile. One explanation for the larger incremental value of the new anomalies in the US with respect to Europe and Japan is that there are more low-cost exchange traded funds in the US that arbitrage away the well-known strategies. It is therefore necessary to find new strategies to get the same predictability of stock returns over time.

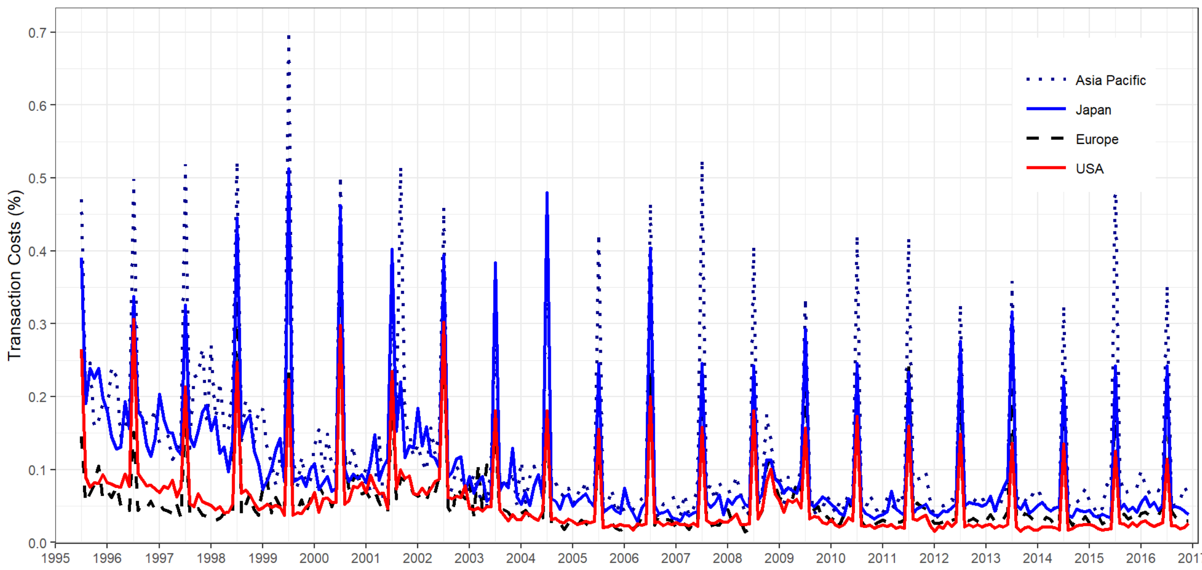
To conclude, the marginal value of the new anomalies remains positive over time. It is therefore valuable to follow recent academic research as it can increase returns to investors. The positive value of new anomalies is in line with the purpouse of academic publishing process where new findings are put under scrutiny and the authors have to prove that their findings provide incremental value with respect to the existing body of knowledge. The academic review process therefore fulfills its purpouse.

3.5 Transaction Costs

This section studies the out-of-sample performance of the strategies after the transaction costs. It is possible that the profits on the strategies are only virtual and transaction costs are larger than the returns. It is therefore important to examine the costs related to the strategies.

3.5.1 Transaction Costs on the Strategies

Panel A: Portfolio-mixing Strategy.



Panel B: Gradient Boosting Regression Trees Mispricing Strategy.

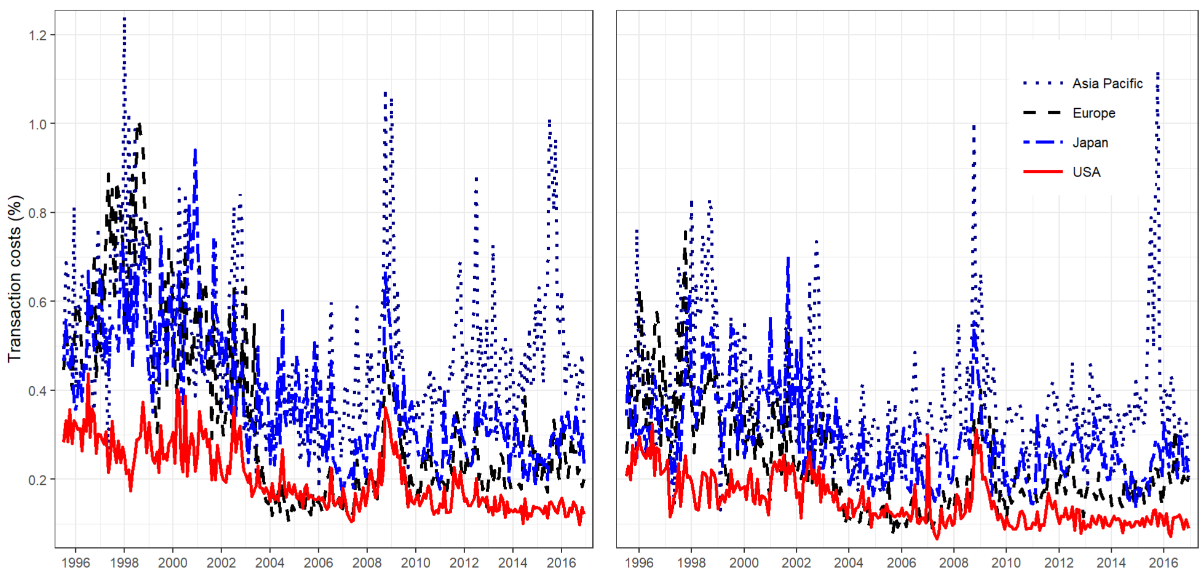


Figure 3.7: **Monthly Transaction Costs.** Panel A shows transaction costs for the portfolio mixing strategy that equally invests in all the significant anomalies as described in Table 3.3. Panel B shows transaction costs for the mispricing strategy described in Table 3.4 that is estimated on individual stock returns from the US. The transaction costs are estimated with VoV(% Spread) proxy of Fong et al. (2017).

Panel A in Figure 3.7 describes transaction costs on the portfolio-mixing strategy introduced in Section 3.2.1. The transaction costs are measured by VoV(% Spread) proxy introduced in Fong et al. (2017). It is evident that the trading costs are similar across the regions for the liquid sample of stocks. The highest transaction costs tend to be in

Asia Pacific region. The peaks in the figure appear every July because of the annual rebalancing of the fundamental strategies. The graph also documents that there are periods with significant spillover of illiquidity. Two such major episodes are Financial Crisis of 2008 and Dot-com bubble of early 2000s. The transaction costs have decreased significantly over time with the increase in market share of electronic trading in 2000s.

The transaction costs on the mispricing strategy are covered next. Panel B in Figure 3.7 maps transaction costs on the gradient boosting regression trees strategy estimated in the US. It is apparent that the transaction costs are larger than in case of portfolio-mixing strategy. The costs are larger because a large portion of the individual anomalies are fundamental anomalies that are rebalanced annually, whereas, the mispricing strategy is rebalanced monthly. The transaction costs have decreased significantly over time and there are again several historical episodes where they were heavily elevated, one being the Financial Crisis of 2008. The costs are smaller on value-weighted portfolios relative to equal-weighting which is expected because the value-weighting puts larger weight on more liquid stocks.

Table 3.10:
Transaction Costs on the Mispricing Strategy
The table shows transaction costs and turnover on the gradient boosting regression trees mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. The transaction costs are estimated either with VoV(% Spread) proxy of Fong et al. (2017), average daily closing quoted spread, or Gibbs proxy of Hasbrouck (2009). The transaction costs and turnover are in percentage points per month.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
VoV	0.203	0.343	0.388	0.493	0.357	0.151	0.232	0.288	0.386	0.264
Gibbs	0.819	0.651	0.816	0.784	0.767	0.712	0.567	0.733	0.714	0.681
Quoted Spread	0.111	0.511	0.482	0.791	0.473	0.101	0.384	0.418	0.645	0.387
Turnover	120.0	119.2	118.9	123.6	120.4	130.5	127.1	127.7	139.4	131.2

Table 3.10 presents average transaction costs on the gradient boosting regression trees mispricing strategy. The transaction costs are estimated with three liquidity proxies introduced in section 3.1.5. All the proxies provide very similar estimates of the transaction costs outside the US. Estimates from Gibbs proxy are significantly higher in the US than for the two other proxies. Gibbs proxy is, however, also the most noisy proxy since it is constructed at an annual frequency. It is furthermore not very suitable to measure transaction costs for the most liquid stocks due to its construction.

Table 3.10 also shows turnover of the mispricing strategy. The turnover is defined as

$$Turnover_t = \sum_i abs(w_{i,t} - w_{i,t-1} r_{i,t-1})/2 \tag{3.4}$$

where $w_{i,t}$ is weight of stock i in the investment portfolio at the start of period $t-1$ and $r_{i,t-1}$ is stock return over period $t-1$ to t. Sum of all absolute weights $w_{i,t}$ is equal to 2 since the portfolio is long-short. The turnover is close to 125% monthly in all the regions which means that over 60% of all the held stocks have to be sold and new bought for both the short and long leg of the strategy. The turnover can be easily reduced by staggered portfolio rebalancing but it is not a source of serious worries here due to the small average transaction costs on the liquid universe of stocks.

The sample of stocks has been selected to be liquid ex ante. Only about 500 most liquid US stocks fulfill this criterion. These stocks should be with virtually no fixed transaction costs. The depicted costs therefore correspond to unfavorable trade executions through aggressive marketable orders. Sophisticated trade execution systems using limit orders are able to execute the strategies without any transaction costs.

3.5.2 Performance of the Strategies after Transaction Costs

Portfolio-mixing Strategy

Panel A in Table 3.11 presents returns on the portfolio-mixing strategy introduced in Table 3.3 adjusted for the trading costs. The set of selected significant strategies is different from Table 3.3 as the strategies are selected on after cost basis here. The selection after adjusting for transaction costs leads to a more profitable meta-strategy as the anomalies with the largest profitability are also often those with the largest transaction costs.

Returns on the strategy remain positive outside Japan but they are generally smaller than without the transaction costs. The Sharpe ratios are also smaller. The global portfolio-mixing strategy, however, remains significantly profitable with Sharpe ratio close to 0.5 for value-weighted returns.

Mispricing Strategy

Panel B in Table 3.11 presents performance of the mispricing strategy after transaction costs. The mean returns on the strategy remain significantly positive at 5% level. The net mean annualized returns in the US are above 10% for the machine learning strategies. Sharpe ratios remain high, especially for the global strategy using neural networks where they are larger than one.

The mean returns after transaction costs for weighted least square method are again smaller than for the more advanced machine learning methods. The difference is even larger on risk adjusted basis. This difference in performance documents that the choice of appropriate forecasting method is very important for success of investing into the anomalies.

To conclude, the strategies remain profitable even after accounting for the transaction costs. The profitability of the strategies is therefore not illusory and can be capitalized by the investors.

3.6 Conclusion

This study has examined profitability of the quantitative strategies based on published anomalies around the globe. It has been shown that investing into individual anomalies is profitable after accounting for transaction costs even on liquid universe of stocks. The

Table 3.11:
Performance of the Strategies after Transaction Costs
Panel A shows returns minus transaction costs of the portfolio-mixing strategy described in Table 3.3. Panel B shows returns after transaction costs of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. The transaction costs are estimated with VoV(% Spread) proxy of Fong et al. (2017). The returns are reported in percentage points per month.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Panel A: Portfolio-mixing Strategy										
Mean	0.151	0.225	-0.119	0.612	0.217	0.109	0.066	0.185	0.589	0.237
Sharpe Ratio	0.186	0.340	-0.165	0.609	0.399	0.184	0.104	0.202	0.630	0.488
Skewness	0.272	0.024	-2.562	-1.255	-0.630	-0.044	0.202	0.260	1.264	0.643
Kurtosis	8.958	7.277	20.36	13.26	6.979	5.907	9.065	23.91	13.59	16.12
Max Drawdown	-29.47	-23.33	-42.78	-28.47	-15.07	-22.87	-34.18	-48.59	-20.75	-11.80
Panel B: Mispricing Strategy										
Weighted Least Squares										
Mean	0.583	0.336	0.514	0.269	0.439	0.583	0.336	0.514	0.269	0.439
Sharpe Ratio	0.348	0.267	0.390	0.170	0.412	0.249	0.294	0.217	0.124	0.313
Max Drawdown	-66.74	-44.12	-49.35	-58.25	-49.63	-71.40	-39.33	-59.36	-58.92	-55.91
Penalized Weighted Least Squares										
Mean	0.537	0.379	0.471	0.344	0.426	0.480	0.547	0.284	0.358	0.403
Sharpe Ratio	0.315	0.290	0.353	0.211	0.385	0.284	0.381	0.177	0.173	0.335
Max Drawdown	-67.37	-43.85	-48.82	-54.35	-51.57	-69.77	-38.97	-58.36	-65.62	-54.72
Gradient Boosting Regression Trees										
Mean	0.962	0.527	0.785	1.157	0.806	1.240	0.359	0.723	1.029	0.769
Sharpe Ratio	0.594	0.390	0.513	0.704	0.793	0.741	0.250	0.376	0.581	0.648
Max Drawdown	-39.92	-49.42	-36.00	-41.48	-29.34	-44.94	-45.90	-48.05	-41.42	-37.52
Random Forest										
Mean	0.844	0.681	0.714	0.414	0.703	0.825	0.089	0.751	0.770	0.525
Sharpe Ratio	0.565	0.513	0.504	0.228	0.706	0.584	0.058	0.431	0.398	0.477
Max Drawdown	-34.62	-49.52	-41.89	-63.52	-29.97	-31.22	-61.80	-44.32	-46.17	-32.73
Neural Networks										
Mean	1.222	0.785	0.934	1.296	1.016	1.282	0.610	0.834	0.829	0.851
Sharpe Ratio	0.782	0.630	0.815	0.804	1.195	0.818	0.479	0.526	0.492	0.812
Max Drawdown	-46.21	-35.29	-25.93	-41.99	-20.70	-49.19	-39.00	-38.28	-58.37	-35.83

performance of the strategy combining individual portfolios on anomalies can be improved by creating a single mispricing signal instead. Machine learning approach for construction of the mispricing signal was advocated and its benefits documented.

The machine learning methods lead to higher (risk adjusted) returns relative to standard methods applied in the academic finance literature. The quantitative strategy using machine learning is highly profitable even on liquid universe of stocks. Value of the more recent anomalies was then studied. The recently published anomalies improve average returns on the investment strategy even after accounting for the previously published anomalies. The recent anomaly studies are therefore successful in finding new sources of priced risk and investors' behavioural biases.

The role of international evidence on precision of predictions of future stock returns was studied. Out-of-sample performance in the US is not improved with international evidence in the training sample for the mispricing strategy. Most of the predictability of expected stock returns in all the global regions under study can be captured solely with the US training sample.

Appendix H

List of the Anomalies

Table H.1:
List of Anomalies

Fundamental	
Accruals	
Accruals	Sloan (1996)
Change in Common Equity	Richardson et al. (2006)
Change in Current Operating Assets	Richardson et al. (2006)
Change in Current Operating Liabilities	Richardson et al. (2006)
Change in Financial Liabilities	Richardson et al. (2006)
Change in Long-Term Investments	Richardson et al. (2006)
Change in Net Financial Assets	Richardson et al. (2006)
Change in Net Non-Cash Working Capital	Richardson et al. (2006)
Change in Net Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Assets	Richardson et al. (2006)
Change in Non-Current Operating Liabilities	Richardson et al. (2006)
Change in Short-Term Investments	Richardson et al. (2006)
Discretionary Accruals	Dechow et al. (1995)
Growth in Inventory	Thomas and Zhang (2002)
Inventory Change	Thomas and Zhang (2002)
Inventory Growth	Belo and Lin (2011)
M/B and Accruals	Bartov and Kim (2004)
Net Working Capital Changes	Soliman (2008)
Percent Operating Accrual	Hafzalla et al. (2011)
Percent Total Accrual	Hafzalla et al. (2011)
Total Accruals	Richardson et al. (2006)
Intangibles	
$\triangle$ Gross Margin - $\triangle$ Sales	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Accounts Receivable	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ Inventory	Abarbanell and Bushee (1998)
$\triangle$ Sales - $\triangle$ SG and A	Abarbanell and Bushee (1998)
Asset Liquidity	Ortiz-Molina and Phillips (2014)
Asset Liquidity II	Ortiz-Molina and Phillips (2014)
Cash-to-assets	Palazzo (2012)
Earnings Conservatism	Francis et al. (2004)
Earnings Persistence	Francis et al. (2004)
Earnings Predictability	Francis et al. (2004)
Earnings Smoothness	Francis et al. (2004)
Earnings Timeliness	Francis et al. (2004)
Herfindahl Index	Hou and Robinson (2006)
Hiring rate	Belo et al. (2014)
Industry Concentration Assets	Hou and Robinson (2006)
Industry Concentration Book Equity	Hou and Robinson (2006)
Industry-adjusted Organizational Capital-to-Assets	Eisfeldt and Papanikolaou (2013)

Industry-adjusted Real Estate Ratio	Tuzel (2010)
Org. Capital	Eisfeldt and Papanikolaou (2013)
RD / Market Equity	Chan et al. (2001)
RD Capital-to-assets	Li (2011)
RD Expenses-to-sales	Chan et al. (2001)
Tangibility	Hahn and Lee (2009)
Unexpected RD Increases	Eberhart et al. (2004)
Whited-Wu Index	Whited and Wu (2006)
Investment	
△ CAPEX - △ Industry CAPEX	Abarbanell and Bushee (1998)
Asset Growth	Cooper et al. (2008)
Change Net Operating Assets	Hirshleifer et al. (2004)
Changes in PPE and Inventory-to-Assets	Lyandres et al. (2007)
Composite Debt Issuance	Lyandres et al. (2007)
Composite Equity Issuance (5-Year)	Daniel and Titman (2006)
Debt Issuance	Spiess and Affleck-Graves (1995)
Growth in LTNOA	Fairfield et al. (2003)
Investment	Titman et al. (2004)
Net Debt Finance	Bradshaw et al. (2006)
Net Equity Finance	Bradshaw et al. (2006)
Net Operating Assets	Hirshleifer et al. (2004)
Noncurrent Operating Assets Changes	Soliman (2008)
Share Repurchases	Ikenberry et al. (1995)
Total XFIN	Bradshaw et al. (2006)
Profitability	
Asset Turnover	Soliman (2008)
Capital Turnover	Haugen and Baker (1996)
Cash-based Operating Profitability	Ball et al. (2016)
Change in Asset Turnover	Soliman (2008)
Change in Profit Margin	Soliman (2008)
Earnings / Price	Basu (1977)
Earnings Consistency	Alwathainani (2009)
F-Score	Piotroski (2000)
Gross Profitability	Novy-Marx (2013)
Labor Force Efficiency	Abarbanell and Bushee (1998)
Leverage	Bhandari (1988)
O-Score (More Financial Distress)	Dichev (1998)
Operating Profits to Assets	Ball et al. (2016)
Operating Profits to Equity	Fama and French (2015)
Profit Margin	Soliman (2008)
Return on Net Operating Assets	Soliman (2008)
Return-on-Equity	Haugen and Baker (1996)
Z-Score (Less Financial Distress)	Dichev (1998)
Value	
Assets-to-Market	Fama and French (1992)
Book Equity / Market Equity	Fama and French (1992)
Cash Flow / Market Equity	Lakonishok et al. (1994)
Duration of Equity	Dechow et al. (2004)
Enterprise Component of Book/Price	Penman et al. (2007)
Enterprise Multiple	Loughran and Wellman (2011)
Intangible Return	Daniel and Titman (2006)
Leverage Component of Book/Price	Penman et al. (2007)
Net Payout Yield	Boudoukh et al. (2007)
Operating Leverage	Novy-Marx (2010)
Payout Yield	Boudoukh et al. (2007)
Sales Growth	Lakonishok et al. (1994)
Sales/Price	Barbee Jr et al. (1996)
Sustainable Growth	Lockwood and Prombutr (2010)
Market Friction	
11-Month Residual Momentum	Blitz et al. (2011)

52-Week High	George and Hwang (2004)
Amihud's Measure (Illiquidity)	Amihud (2002)
Beta	Fama and MacBeth (1973a)
Betting against Beta	Frazzini and Pedersen (2014)
Bid-Ask Spread	Amihud and Mendelson (1986)
Cash Flow Variance	Haugen and Baker (1996)
Coefficient of Variation of Share Turnover	Chordia et al. (2001)
Coskewness	Harvey and Siddique (2000)
Downside Beta	Ang et al. (2006a)
Earnings Forecast-to-Price	Elgers et al. (2001)
Firm Age	Barry and Brown (1984)
Firm Age-Momentum	Zhang (2006)
Idiosyncratic Risk	Ang et al. (2006b)
Industry Momentum	Moskowitz and Grinblatt (1999)
Lagged Momentum	Novy-Marx (2012)
Liquidity Beta 1	Acharya and Pedersen (2005)
Liquidity Beta 2	Acharya and Pedersen (2005)
Liquidity Beta 3	Acharya and Pedersen (2005)
Liquidity Beta 4	Acharya and Pedersen (2005)
Liquidity Beta 5	Acharya and Pedersen (2005)
Liquidity Shocks	Bali et al. (2013)
Long-Term Reversal	Bondt and Thaler (1985)
Max	Bali et al. (2011)
Momentum	Jegadeesh and Titman (1993)
Momentum and LT Reversal	Kot and Chan (2006)
Momentum-Reversal	Jegadeesh and Titman (1993)
Momentum-Volume	Lee and Swaminathan (2000)
Price	Blume and Husic (1973)
Seasonality	Heston and Sadka (2008)
Seasonality 1 A	Heston and Sadka (2008)
Seasonality 1 N	Heston and Sadka (2008)
Seasonality 11-15 A	Heston and Sadka (2008)
Seasonality 11-15 N	Heston and Sadka (2008)
Seasonality 16-20 A	Heston and Sadka (2008)
Seasonality 16-20 N	Heston and Sadka (2008)
Seasonality 2-5 A	Heston and Sadka (2008)
Seasonality 2-5 N	Heston and Sadka (2008)
Seasonality 6-10 A	Heston and Sadka (2008)
Seasonality 6-10 N	Heston and Sadka (2008)
Share Issuance (1-Year)	Pontiff and Woodgate (2008)
Share Turnover	Datar et al. (1998)
Short-Term Reversal	Jegadeesh (1990)
Size	Banz (1981)
Tail Risk	Kelly and Jiang (2014)
Total Volatility	Ang et al. (2006b)
Volume / Market Value of Equity	Haugen and Baker (1996)
Volume Trend	Haugen and Baker (1996)
Volume Variance	Chordia et al. (2001)
I/B/E/S	
Analyst Value	Frankel and Lee (1998)
Analysts Coverage	Elgers et al. (2001)
Change in Forecast + Accrual	Barth and Hutton (2004)
Change in Recommendation	Jegadeesh et al. (2004)
Changes in Analyst Earnings Forecasts	Hawkins et al. (1984)
Disparity between LT and ST Earnings Growth Forecasts	Da and Warachka (2011)
Dispersion in Analyst LT Growth Forecasts	Anderson et al. (2005)
Down Forecast	Barber et al. (2001)
Forecast Dispersion	Diether et al. (2002)
Long-Term Growth Forecasts	La Porta (1996)
Up Forecast	Barber et al. (2001)

Appendix I

Definition of Liquidity Proxies

I.1 VoV(% Spread) Proxy

The fixed transaction costs are approximated with VoV(% Spread) proxy introduced in Fong et al. (2017). It is defined as

$$8 \frac{\sigma^{2/3}}{avg\ vol^{1/3}} \tag{I.1}$$

where σ is standard deviation of daily returns and *avg vol* is average daily trading volume in USD within a given month. The trading volume is in USD and deflated to 2000 prices. The proxy roughly measures fixed component of trading costs and excludes price impact. Including the price impact would further increase the transaction costs. Fong et al. (2017) show that the price impact component is very hard to measure. It is volatile over regions, and therefore, very dependent on execution strategy of individual asset managers. The focus is therefore solely on the fixed component of transaction costs (effective spread).

Kyle and Obizhaeva (2016) estimated a relationship between transaction costs and size of large institutional portfolio transfers depending on average daily trading volume and volatility of the stocks. The analysis was conducted on a proprietary dataset covering the 2002-2005 period. VoV(% Spread) roughly corresponds to the fixed component of their estimated transaction cost function.

Fong et al. (2017) benchmarked the proxy to other existing proxies and found that it can be outperformed only by closing quoted spread. The quoted spread is, however, not available for all the regions over the whole sample period.

I.2 Closing Quoted Spread

Closing quoted spread for a given month is defined as

$$QS = \frac{1}{T} \sum_{t=1}^{T} \frac{2(ask - bid)}{ask + bid} \tag{I.2}$$

where ask and bid are observed at the end of trading day on each stock exchange and T is number of days in the given month. Observations with missing or negative daily value of QS are excluded from the average. CRSP lists the best quote of bid and ask for NASDAQ stocks and the last representative quotes before the market close for NYSE and Amex stocks. Precise definition of QS can therefore vary over the exchanges.

Chung and Zhang (2014) first benchmarked the QS by comparing it to high frequency effective spread estimates from Trade and Quote (TAQ) database. They showed that QS has about 95% average cross sectional correlation with TAQ effective spread over the 1998 to 2009 period. Fong et al. (2017) document that it is also the best spread proxy in international setting. One problem with QS is that it is often missing in earlier periods and therefore has to be backfilled with other proxies.

I.3 Gibbs Proxy

Roll (1984) introduced one of the first spread proxies in the academic literature. He assumed that the true price of stock follows a random walk with bid-ask jumps. That is,

$$P_t^A = P_{t-1}^A + u_t, \quad P_t^O = P_t^A + sq_t \tag{I.3}$$

$$\triangle P_t^o = s \triangle q_t + u_t,\, u_t \sim N(0, \sigma_u^2) \tag{I.4}$$

where P_t^o is observed log price, P_t^A is price of the underlying Brownian motion, and s is a half spread. Indicator q_t is equal to one if the last trade in the day is buy, minus one if it is sell, and zero if no prices are available during the day. Serial correlation of the price changes $\triangle P_t^o$ should be negative and related to the spread through the following relationship

$$S_{roll} = 2\sqrt{-cov(\triangle P_t^o, \triangle P_{t+1}^o)}. \tag{I.5}$$

This can be contributed to the fact that

$$cov(\triangle P_t^o, \triangle P_{t+1}^o) = cov(s(q_t - q_{t-1}) + u_t, s(q_{t+1} - q_t) + u_{t+1}) = \mathbb{E}[-s^2 q_t^2] = -s^2. \tag{I.6}$$

The covariance can be positive in practice. In which case the estimate of spread is set equal to zero.

Hasbrouck (2009) proposed to extend the Roll model by estimating it with Gibbs sampler. The idea is to estimate the equation (I.4) augmented with another dependent variable (market return) via Bayesian regression. The variables q_t are generated from the data by Gibbs sampler.[1]

The proxy is estimated at annual frequency for each stock and calendar year. Lower frequency than annual leads to severe deterioration of the proxy's performance.

[1] Note that there is an error in the original paper in *Journal of Finance*. The correct posterior distribution for σ_u^2 is $IG(\alpha_{prior} + \frac{n}{2}, \beta_{prior} + \frac{\sum u_t^2}{2})$.

Appendix J

Additional Results

Table J.1:

Performance of the Mispricing Strategy Estimated on Stocks Outside the US: Weighted Least Squares Regressions

The table shows returns of the mispricing strategy as described in Table 3.4 that is estimated on individual stocks from the US, US & Japan, US & Europe, or US & Japan & Europe. The returns are in percentage points per month.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Evidence from the US										
Mean	0.801	0.680	0.922	0.782	0.810	0.575	0.647	0.648	0.633	0.639
Sharpe Ratio	0.479	0.541	0.701	0.497	0.763	0.348	0.472	0.410	0.318	0.550
Skewness	-0.340	0.239	-0.425	-0.356	0.025	-0.121	-0.017	-0.688	-0.238	-0.041
Kurtosis	8.521	5.673	4.167	3.713	8.488	7.214	6.483	5.544	4.824	7.880
Max Drawdown	-64.70	-37.10	-43.36	-47.61	-43.51	-69.75	-34.16	-44.52	-49.86	-50.85
Evidence from the US & Japan										
Mean	0.694	0.648	0.796	0.748	0.722	0.607	0.703	0.602	0.569	0.639
Sharpe Ratio	0.404	0.485	0.543	0.450	0.642	0.337	0.485	0.339	0.302	0.498
Skewness	-0.215	0.296	-0.407	-0.373	0.055	-0.167	0.296	0.590	0.105	0.362
Kurtosis	7.929	5.619	5.593	4.571	7.920	7.716	5.202	11.16	3.816	9.129
Max Drawdown	-69.18	-42.84	-55.39	-48.93	-49.92	-73.56	-36.78	-53.55	-51.64	-54.34
Information Ratio	-0.181	-0.056	-0.189	-0.035	-0.218	0.042	0.098	-0.044	-0.045	0.000
Evidence from the US & Europe										
Mean	0.780	0.700	0.854	1.140	0.842	0.708	0.638	0.569	0.897	0.715
Sharpe Ratio	0.435	0.546	0.646	0.697	0.736	0.418	0.477	0.340	0.487	0.603
Skewness	-0.581	0.082	-0.590	-0.356	-0.397	-0.308	0.141	-0.617	-0.255	-0.221
Kurtosis	10.31	5.844	4.491	3.826	8.745	8.525	5.838	4.677	3.726	7.486
Max Drawdown	-63.84	-41.66	-41.74	-48.17	-41.68	-63.39	-36.58	-58.68	-49.51	-44.83
Information Ratio	-0.037	0.031	-0.116	0.442	0.082	0.173	-0.014	-0.098	0.274	0.142
Evidence from the US & Japan & Europe										
Mean	0.808	0.699	0.884	1.112	0.853	0.765	0.626	0.469	1.020	0.730
Sharpe Ratio	0.444	0.516	0.645	0.691	0.736	0.422	0.426	0.267	0.530	0.580
Skewness	-0.359	0.304	-0.267	-0.321	-0.096	-0.303	-0.038	0.134	-0.466	0.034
Kurtosis	8.281	5.359	4.212	4.241	7.407	7.961	5.712	7.329	5.093	7.344
Max Drawdown	-70.99	-45.05	-39.66	-42.05	-45.60	-72.31	-37.82	-63.48	-45.19	-50.12
Information Ratio	0.009	0.028	-0.057	0.346	0.096	0.221	-0.030	-0.167	0.320	0.155

Table J.2:

Is Marginal Return to Following New Anomalies Decreasing over Time?

The table shows returns of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. The set of anomalies in the estimation is restricted to those that were published before 1995, 2000, 2005, or 2010. Returns are reported in percentage points per month over the 2010-2016 period.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Weighted Least Squares										
Published by 1995										
Mean	0.189	0.408	0.324	0.343	0.301	0.190	0.300	1.028	0.424	0.337
Sharpe Ratio	0.210	0.408	0.306	0.255	0.440	0.208	0.263	0.926	0.295	0.442
Published by 2000										
Mean	0.193	0.777	0.577	0.055	0.409	0.129	0.588	1.063	-0.121	0.415
Sharpe Ratio	0.213	0.811	0.475	0.042	0.592	0.138	0.506	0.816	-0.090	0.540
Information Ratio	0.006	0.556	0.333	-0.306	0.253	-0.086	0.354	0.040	-0.421	0.150
Published by 2005										
Mean	0.896	0.780	0.873	0.517	0.793	0.807	0.476	1.376	0.010	0.860
Sharpe Ratio	1.031	0.812	0.760	0.386	1.181	1.001	0.392	1.091	0.006	1.189
Information Ratio	1.233	0.004	0.476	0.380	0.974	0.858	-0.107	0.376	0.081	0.686
Published by 2010										
Mean	0.997	0.868	0.978	1.283	0.994	0.797	0.569	0.934	1.247	0.892
Sharpe Ratio	1.240	0.997	0.978	0.937	1.589	0.907	0.561	0.830	0.863	1.262
Information Ratio	0.198	0.156	0.153	0.796	0.595	-0.015	0.130	-0.539	0.792	0.069
Gradient Boosting Regression Trees										
Published by 1995										
Mean	0.514	0.446	0.634	1.034	0.609	0.446	0.695	0.706	0.817	0.617
Sharpe Ratio	0.638	0.480	0.637	0.723	1.115	0.483	0.739	0.661	0.528	1.006
Published by 2000										
Mean	0.517	0.906	1.090	1.595	0.918	0.134	0.652	1.172	1.399	0.561
Sharpe Ratio	0.655	1.015	1.167	1.246	1.765	0.150	0.728	1.001	1.049	0.947
Information Ratio	0.006	0.658	0.487	0.546	0.752	-0.389	-0.054	0.348	0.432	-0.096
Published by 2005										
Mean	0.830	0.948	1.169	1.583	1.045	0.822	0.792	1.182	1.161	0.941
Sharpe Ratio	1.223	1.080	1.286	1.483	2.324	1.054	1.082	1.241	1.028	1.978
Information Ratio	0.483	0.057	0.095	-0.009	0.294	0.827	0.154	0.009	-0.152	0.645
Published by 2010										
Mean	1.011	1.140	0.817	1.943	1.121	0.509	1.085	0.911	1.765	0.898
Sharpe Ratio	1.443	1.264	0.803	1.562	2.039	0.585	1.221	0.733	1.435	1.530
Information Ratio	0.348	0.289	-0.447	0.369	0.216	-0.405	0.382	-0.277	0.485	-0.094

J.1 Monthly Updated Fundamental Anomalies

Anomalies based on annual financial statements have so far been updated annually every June. June was chosen so that firms with financial year ending in December have 6 months to publish their statements. The explicit assumption was that all the firms publish their statements within 6 months after their financial year has ended. The rule was originally devised on the US data where great majority of firms have their financial year ending in December. The usual financial year end is, however, different in the other regions. 78% of firms in Japan have financial year ending in March. The most frequent choice of financial year end in Asia Pacific region is either December or June both being about equally likely. Financial year end date outside December leads to the financial statements being older than 6 months in June and thus being less relevant. Bartram and Grinblatt (2018b) and Jacobs and Müller (2017c) circumvented this problem when working with international data by relying on point-in-time Reuters database that presents financial statements as they were published by a given date and creating the fundamental signals monthly. We do not have access to the point-in-time database but we will here create a pseudo point-in-time database and will also refresh the fundamental signals monthly rather than annually.

Table J.3 presents results from Table 3.4 based on the annual construction of fundamental signals along with their monthly construction. Everything remains the same as in Table 3.4 with the only difference being that the fundamental signals are updated every month with financial statement information from financial years ending at least 6 months prior. The explicit assumption again is that all the firms publish their statements within the 6 months after their financial year has ended. All the trade data information such as market cap is also updated monthly and taken the most recent. Market cap was previously taken from the previous calendar year end as in Fama and French (1992) and was therefore outdated by 6 months by June. Asness and Frazzini (2013) showed that market cap from June leads to better performance of value factor. There can therefore also be some benefit from shifting the trade data information.

Table J.3 documents that the lag in availability of the financial statements leads to some loss in performance in almost all the regions. Both mean returns and Sharpe ratios with the monthly updating of the fundamental signals are about 10% higher relative to when they are updated annually. To conclude, the monthly updating can slightly improve the performance of the mispricing strategy but it does not affect the main conclusions of this study.

Table J.3:
Performance of the Mispricing Strategy with Monthly Updated Fundamental Signals

The table shows returns of the mispricing strategy described in Table 3.4 that is estimated on the individual stocks from the US. The results labelled "Annually Updated Fundamental Signals" directly correspond to Table 3.4 where the fundamental signals are updated every June while the results labelled "Monthly Updated Fundamental Signals" are created using fundamental signals that are updated every month based on financial statements released more than six months prior. Panel A describes results from weighted least squares estimation method while Panel B reports results from gradient boosting regression trees method. The returns are reported in percentage points per month.

	Equal-weighted					Value-weighted				
	USA	Europe	Japan	AP	Global	USA	Europe	Japan	AP	Global
Panel A: Weighted Least Squares										
Annually Updated Fundamental Signals										
Mean	0.801	0.680	0.922	0.782	0.810	0.575	0.647	0.648	0.633	0.639
Sharpe Ratio	0.479	0.541	0.701	0.497	0.763	0.348	0.472	0.410	0.318	0.550
Skewness	-0.340	0.239	-0.425	-0.356	0.025	-0.121	-0.017	-0.688	-0.238	-0.041
Kurtosis	8.521	5.673	4.167	3.713	8.488	7.214	6.483	5.544	4.824	7.880
Max Drawdown	-64.70	-37.10	-43.36	-47.61	-43.51	-69.75	-34.16	-44.52	-49.86	-50.85
Monthly Updated Fundamental Signals										
Mean	0.889	0.750	0.869	1.078	0.883	0.736	0.585	0.640	0.676	0.696
Sharpe Ratio	0.537	0.644	0.690	0.663	0.867	0.447	0.440	0.386	0.360	0.634
Skewness	-0.203	0.246	-0.479	-0.307	0.008	-0.046	0.075	-0.139	-0.209	-0.063
Kurtosis	8.039	4.970	4.987	3.822	7.683	6.466	6.056	5.155	4.145	6.239
Max Drawdown	-63.40	-38.59	-39.29	-42.31	-37.78	-65.32	-26.37	-62.52	-41.22	-41.57
Information Ratio	0.173	0.133	-0.074	0.272	0.219	0.218	-0.105	-0.008	0.029	0.114
Panel B: Gradient Boosting Regression Trees										
Annually Updated Fundamental Signals										
Mean	1.165	0.870	1.173	1.650	1.163	1.391	0.591	1.011	1.415	1.033
Sharpe Ratio	0.720	0.644	0.766	1.005	1.146	0.831	0.412	0.525	0.800	0.870
Skewness	0.319	-1.160	0.682	-0.437	-0.449	0.561	-1.314	0.800	-0.112	-0.433
Kurtosis	6.653	10.17	8.274	5.575	6.812	9.287	12.03	8.611	4.718	7.797
Max Drawdown	-38.31	-48.25	-34.37	-36.65	-27.45	-43.93	-42.31	-41.79	-39.58	-35.62
Monthly Updated Fundamental Signals										
Mean	1.264	1.039	1.242	1.597	1.254	1.492	0.771	1.207	1.314	1.125
Sharpe Ratio	0.786	0.840	0.858	0.965	1.241	0.900	0.538	0.645	0.770	0.902
Skewness	0.260	-0.848	0.438	-0.099	-0.337	1.041	-1.332	1.057	0.203	0.518
Kurtosis	7.663	10.95	8.076	5.330	6.896	8.525	16.78	10.60	5.307	7.435
Max Drawdown	-47.13	-43.82	-30.79	-37.49	-27.12	-43.11	-43.00	-36.56	-35.74	-31.02
Information Ratio	0.142	0.270	0.091	-0.042	0.237	0.110	0.179	0.191	-0.065	0.141

Bibliography

Abarbanell, Jeffery S, and Brian J Bushee, 1998, Abnormal returns to a fundamental analysis strategy, *Accounting Review* 19–45.

Acharya, Viral V, and Lasse Heje Pedersen, 2005, Asset pricing with liquidity risk, *Journal of financial Economics* 77, 375–410.

Alwathainani, Abdulaziz M, 2009, Consistency of firms' past financial performance measures and future returns, *The British Accounting Review* 41, 184–196.

Amihud, Yakov, 2002, Illiquidity and stock returns: cross-section and time-series effects, *Journal of financial markets* 5, 31–56.

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, *Journal of financial Economics* 17, 223–249.

Anderson, Evan W, Eric Ghysels, and Jennifer L Juergens, 2005, Do heterogeneous beliefs matter for asset pricing?, *The Review of Financial Studies* 18, 875–924.

Andrews, Donald WK, and J Christopher Monahan, 1992, An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica: Journal of the Econometric Society* 953–966.

Andrikopoulos, Panagiotis, James Clunie, and Antonios Siganos, 2013, Short-selling constraints and quantitativeinvestment strategies, *The European Journal of Finance* 19, 19–35.

Ang, Andrew, Joseph Chen, and Yuhang Xing, 2006a, Downside risk, *The Review of Financial Studies* 19, 1191–1239.

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006b, The cross-section of volatility and expected returns, *The Journal of Finance* 61, 259–299.

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2009, High idiosyncratic volatility and low returns: International and further us evidence, *Journal of Financial Economics* 91, 1–23.

Asness, Clifford, and Andrea Frazzini, 2013, The devil in hmls details, *The Journal of Portfolio Management* 39, 49–68.

Asparouhova, Elena, Hendrik Bessembinder, and Ivalina Kalcheva, 2010, Liquidity biases in asset pricing tests, *Journal of Financial Economics* 96, 215–237.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transaction costs, *Journal of Financial Economics* 106, 473–491.

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.

Bali, Turan G, Lin Peng, Yannan Shen, and Yi Tang, 2013, Liquidity shocks and stock market reactions, *The Review of Financial Studies* 27, 1434–1485.

Ball, Ray, Joseph Gerakos, Juhani T Linnainmaa, and Valeri Nikolaev, 2016, Accruals, cash flows, and operating profitability in the cross section of stock returns, *Journal of Financial Economics* 121, 28–45.

Banz, Rolf W, 1981, The relationship between return and market value of common stocks, *Journal of financial economics* 9, 3–18.

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.

Barber, Brad, Reuven Lehavy, Maureen McNichols, and Brett Trueman, 2001, Can investors profit from the prophets? security analyst recommendations and stock returns, *The Journal of Finance* 56, 531–563.

Barber, Brad M, Emmanuel T De George, Reuven Lehavy, and Brett Trueman, 2013, The earnings announcement premium around the globe, *Journal of Financial Economics* 108, 118–138.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *The journal of finance* 65, 179–216.

Barry, Christopher B, and Stephen J Brown, 1984, Differential information and the small firm effect, *Journal of Financial Economics* 13, 283–294.

Barth, Mary E, and Amy P Hutton, 2004, Analyst earnings forecast revisions and the pricing of accruals, *Review of accounting studies* 9, 59–96.

Bartov, Eli, and Myungsun Kim, 2004, Risk, mispricing, and value investing, *Review of Quantitative Finance and Accounting* 23, 353–376.

Bartram, Söhnke M, and Mark Grinblatt, 2018a, Agnostic fundamental analysis works, *Journal of Financial Economics* 128, 125–147.

Bartram, Söhnke M, and Mark Grinblatt, 2018b, Global market inefficiencies .

Basu, Sanjoy, 1977, Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis, *The journal of Finance* 32, 663–682.

Belo, Frederico, and Xiaoji Lin, 2011, The inventory growth spread, *The Review of Financial Studies* 25, 278–313.

Belo, Frederico, Xiaoji Lin, and Santiago Bazdresch, 2014, Labor hiring, investment, and stock return predictability in the cross section, *Journal of Political Economy* 122, 129–177.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The control of the false discovery rate in multiple testing under dependency, *Annals of statistics* 1165–1188.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al., 2013, Valid post-selection inference, *The Annals of Statistics* 41, 802–837.

Bhandari, Laxmi Chand, 1988, Debt/equity ratio and expected common stock returns: Empirical evidence, *The journal of finance* 43, 507–528.

Blitz, David, Joop Huij, and Martin Martens, 2011, Residual momentum, *Journal of Empirical Finance* 18, 506–521.

Blume, Marshall E, and Frank Husic, 1973, Price, beta, and exchange listing, *The Journal of Finance* 28, 283–299.

Bondt, Werner FM, and Richard Thaler, 1985, Does the stock market overreact?, *The Journal of finance* 40, 793–805.

Boudoukh, Jacob, Roni Michaely, Matthew Richardson, and Michael R Roberts, 2007, On the importance of measuring payout yield: Implications for empirical asset pricing, *The Journal of Finance* 62, 877–915.

Bradshaw, Mark T, Scott A Richardson, and Richard G Sloan, 2006, The relation between corporate financing activities, analysts forecasts and stock returns, *Journal of Accounting and Economics* 42, 53–85.

Chan, Louis KC, Josef Lakonishok, and Theodore Sougiannis, 2001, The stock market valuation of research and development expenditures, *The Journal of Finance* 56, 2431–2456.

Chen, Tianqi, and Tong He, 2017, xgboost: extreme gradient boosting .

Chordia, Tarun, Amit Goyal, and Alessio Saretto, 2017, p-hacking: Evidence from two million trading strategies .

Chordia, Tarun, Avanidhar Subrahmanyam, and V Ravi Anshuman, 2001, Trading activity and expected stock returns, *Journal of Financial Economics* 59, 3–32.

Chui, Andy CW, Sheridan Titman, and KC John Wei, 2010, Individualism and momentum around the world, *The Journal of Finance* 65, 361–392.

Chung, Kee H, and Hao Zhang, 2014, A simple approximation of intraday spreads using daily data, *Journal of Financial Markets* 17, 94–120.

Cooper, Michael, and Huseyin Gulen, 2006, Is time-series-based predictability evident in real time?, *The Journal of Business* 79, 1263–1292.

Cooper, Michael J, Huseyin Gulen, and Michael J Schill, 2008, Asset growth and the cross-section of stock returns, *The Journal of Finance* 63, 1609–1651.

Da, Zhi, and Mitch Warachka, 2011, The disparity between long-term and short-term forecasted earnings growth, *Journal of Financial Economics* 100, 424–442.

Dai, Rui, 2012, International accounting databases on wrds: Comparative analysis .

Daniel, Kent, and Sheridan Titman, 2006, Market reactions to tangible and intangible information, *The Journal of Finance* 61, 1605–1643.

Datar, Vinay T, Narayan Y Naik, and Robert Radcliffe, 1998, Liquidity and stock returns: An alternative test, *Journal of Financial Markets* 1, 203–219.

Davison, AC, and Peter Hall, 1993, On studentizing and blocking methods for implementing the bootstrap with dependent data, *Australian & New Zealand Journal of Statistics* 35, 215–224.

Dechow, Patricia M, Richard G Sloan, and Mark T Soliman, 2004, Implied equity duration: A new measure of equity risk, *Review of Accounting Studies* 9, 197–228.

Dechow, Patricia M, Richard G Sloan, and Amy P Sweeney, 1995, Detecting earnings management, *Accounting review* 193–225.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2007, Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?, *The review of Financial studies* 22, 1915–1953.

Dichev, Ilia D, 1998, Is the risk of bankruptcy a systematic risk?, *the Journal of Finance* 53, 1131–1147.

Diebold, Francis X, and Roberto S Mariano, 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* .

Diether, Karl B, Christopher J Malloy, and Anna Scherbina, 2002, Differences of opinion and the cross section of stock returns, *The Journal of Finance* 57, 2113–2141.

Driscoll, John C, and Aart C Kraay, 1998, Consistent covariance matrix estimation with spatially dependent panel data, *The Review of Economics and Statistics* 80, 549–560.

Eberhart, Allan C, William F Maxwell, and Akhtar R Siddique, 2004, An examination of long-term abnormal stock returns and operating performance following r&d increases, *The Journal of Finance* 59, 623–650.

Eisfeldt, Andrea L, and Dimitris Papanikolaou, 2013, Organization capital and the cross-section of expected returns, *The Journal of Finance* 68, 1365–1406.

Elgers, Pieter T, May H Lo, and Ray J Pfeiffer Jr, 2001, Delayed security price adjustments to financial analysts' forecasts of annual earnings, *The Accounting Review* 76, 613–632.

Fairfield, Patricia M, J Scott Whisenant, and Teri Lombardi Yohn, 2003, Accrued earnings and growth: Implications for future profitability and market mispricing, *The accounting review* 78, 353–371.

Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *the Journal of Finance* 47, 427–465.

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

Fama, Eugene F, and Kenneth R French, 2010, Luck versus skill in the cross-section of mutual fund returns, *The journal of finance* 65, 1915–1947.

Fama, Eugene F, and Kenneth R French, 2012, Size, value, and momentum in international stock returns, *Journal of financial economics* 105, 457–472.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1 – 22.

Fama, Eugene F, and Kenneth R French, 2016, Choosing factors .

Fama, Eugene F, and Kenneth R French, 2017, International tests of a five-factor asset pricing model, *Journal of Financial Economics* 123, 441–463.

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

Fama, Eugene F, and James D MacBeth, 1973b, Risk, return, and equilibrium: Empirical tests, *The journal of political economy* 607–636.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2017, Taming the factor zoo .

Fong, Kingsley, Craig Holden, and Ondrej Tobek, 2017, Are volatility over volume liquidity proxies useful for global or us research? .

Foster, F Douglas, Tom Smith, and Robert E Whaley, 1997, Assessing goodness-of-fit of asset pricing models: The distribution of the maximal r 2, *The Journal of Finance* 52, 591–607.

Francis, Jennifer, Ryan LaFond, Per M Olsson, and Katherine Schipper, 2004, Costs of equity and earnings attributes, *The accounting review* 79, 967–1010.

Frankel, Richard, and Charles MC Lee, 1998, Accounting valuation, market expectation, and cross-sectional stock returns, *Journal of Accounting and economics* 25, 283–319.

Frazzini, Andrea, Ronen Israel, and Tobias J Moskowitz, 2012, Trading costs of asset pricing anomalies, *Fama-Miller working paper* 14–05.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2017, Dissecting characteristics nonparametrically, Technical report, National Bureau of Economic Research.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, 2001, *The elements of statistical learning*, volume 1 (Springer series in statistics New York, NY, USA:).

Friedman, Jerome H, 2001, Greedy function approximation: a gradient boosting machine, *Annals of statistics* 1189–1232.

George, Thomas J, and Chuan-Yang Hwang, 2004, The 52-week high and momentum investing, *The Journal of Finance* 59, 2145–2176.

Goncalves, Silvia, and Timothy J Vogelsang, 2011, Block bootstrap hac robust tests: The sophistication of the naive bootstrap, *Econometric Theory* 27, 745–791.

Götze, Friedrich, Hans R Künsch, et al., 1996, Second-order correctness of the blockwise bootstrap for stationary observations, *The Annals of Statistics* 24, 1914–1933.

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* hhx019.

Griffin, John M, 2002, Are the fama and french factors global or country specific?, *The Review of Financial Studies* 15, 783–803.

Griffin, John M, Patrick J Kelly, and Federico Nardari, 2010, Do market efficiency measures yield correct inferences? a comparison of developed and emerging markets, *The Review of Financial Studies* 23, 3225–3277.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu, 2018, Empirical asset pricing via machine learning .

Hafzalla, Nader, Russell Lundholm, and E Matthew Van Winkle, 2011, Percent accruals, *The Accounting Review* 86, 209–236.

Hahn, Jaehoon, and Hangyong Lee, 2009, Financial constraints, debt capacity, and the cross-section of stock returns, *The Journal of Finance* 64, 891–921.

Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.

Harvey, Campbell R, and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *The Journal of Finance* 55, 1263–1295.

Hasbrouck, Joel, 2009, Trading costs and returns for us equities: Estimating effective costs from daily data, *The Journal of Finance* 64, 1445–1477.

Haugen, Robert A, and Nardin L Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

Hawkins, Eugene H, Stanley C Chamberlin, and Wayne E Daniel, 1984, Earnings expectations and security prices, *Financial Analysts Journal* 40, 24–38.

Heston, Steven L, and Ronnie Sadka, 2008, Seasonality in the cross-section of stock returns, *Journal of Financial Economics* 87, 418–445.

Hirshleifer, David, Kewei Hou, Siew Hong Teoh, and Yinglei Zhang, 2004, Do investors overvalue firms with bloated balance sheets?, *Journal of Accounting and Economics* 38, 297–331.

Holm, Sture, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics* 65–70.

Hou, Kewei, G Andrew Karolyi, and Bong-Chan Kho, 2011a, What factors drive global stock returns?, *The Review of Financial Studies* 24, 2527–2574.

Hou, Kewei, G Andrew Karolyi, and Bong-Chan Kho, 2011b, What factors drive global stock returns?, *The Review of Financial Studies* 24, 2527–2574.

Hou, Kewei, and David T Robinson, 2006, Industry concentration and average stock returns, *The Journal of Finance* 61, 1927–1956.

Hou, Kewei, Chen Xue, and Lu Zhang, 2017, Replicating anomalies, Technical report, National Bureau of Economic Research.

Ikenberry, David, Josef Lakonishok, and Theo Vermaelen, 1995, Market underreaction to open market share repurchases, *Journal of financial economics* 39, 181–208.

Ince, Ozgur S, and R Burt Porter, 2006, Individual equity return data from thomson datastream: Handle with care!, *Journal of Financial Research* 29, 463–479.

Jacobs, Heiko, 2016, Market maturity and mispricing, *Journal of Financial Economics* 122, 270–287.

Jacobs, Heiko, and Sebastian Müller, 2017a, ... and nothing else matters? on the dimensionality and predictability of international stock returns .

Jacobs, Heiko, and Sebastian Müller, 2017b, Anomalies across the globe: Once public, no longer existent? .

Jacobs, Heiko, and Sebastian Müller, 2017c, Measuring mispricing .

Jegadeesh, Narasimhan, 1990, Evidence of predictable behavior of security returns, *The Journal of finance* 45, 881–898.

Jegadeesh, Narasimhan, Joonghyuk Kim, Susan D Krische, and Charles Lee, 2004, Analyzing the analysts: When do recommendations add value?, *The journal of finance* 59, 1083–1124.

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.

Kelly, Bryan, and Hao Jiang, 2014, Tail risk and asset prices, *The Review of Financial Studies* 27, 2841–2871.

Kiefer, Nicholas M, and Timothy J Vogelsang, 2005, A new asymptotic theory for heteroskedasticity-autocorrelation robust tests, *Econometric Theory* 21, 1130–1164.

Kiefer, Nicholas M, Timothy J Vogelsang, and Helle Bunzel, 2000, Simple robust testing of regression hypotheses, *Econometrica* 68, 695–714.

Kosowski, Robert, Narayan Y Naik, and Melvyn Teo, 2007, Do hedge funds deliver alpha? a bayesian and bootstrap analysis, *Journal of Financial Economics* 84, 229–264.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can mutual fund stars really pick stocks? new evidence from a bootstrap analysis, *The Journal of finance* 61, 2551–2595.

Kot, Hung Wan, and Kalok Chan, 2006, Can contrarian strategies improve momentum profits .

Kyle, Albert S, and Anna A Obizhaeva, 2016, Market microstructure invariance: Empirical hypotheses, *Econometrica* 84, 1345–1404.

La Porta, Rafael, 1996, Expectations and the cross-section of stock returns, *The Journal of Finance* 51, 1715–1742.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny, 1994, Contrarian investment, extrapolation, and risk, *The journal of finance* 49, 1541–1578.

Lam, FY Eric C, and KC John Wei, 2011, Limits-to-arbitrage, investment frictions, and the asset growth anomaly, *Journal of Financial Economics* 102, 127–149.

Lee, Charles, and Bhaskaran Swaminathan, 2000, Price momentum and trading volume, *the Journal of Finance* 55, 2017–2069.

Lee, Jason D, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al., 2016, Exact post-selection inference, with application to the lasso, *The Annals of Statistics* 44, 907–927.

Lee, Kuan-Hui, 2011, The world price of liquidity risk, *Journal of Financial Economics* 99, 136–161.

Lewellen, Jonathan, et al., 2015, The cross-section of expected stock returns, *Critical Finance Review* 4, 1–44.

Li, Dongmei, 2011, Financial constraints, r&d investment, and stock returns, *The Review* *of Financial Studies* 24, 2974–3007.

Lo, Andrew W, and A Craig MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *The Review of Financial Studies* 3, 431–467.

Lockwood, Larry, and Wikrom Prombutr, 2010, Sustainable growth and stock returns, *Journal of Financial Research* 33, 519–538.

Loughran, Tim, and Jay W Wellman, 2011, New evidence on the relation between the enterprise multiple and average stock returns, *Journal of Financial and Quantitative Analysis* 46, 1629–1650.

Lyandres, Evgeny, Le Sun, and Lu Zhang, 2007, The new issues puzzle: Testing the investment-based explanation, *The Review of Financial Studies* 21, 2825–2855.

MacKinlay, A Craig, 1995, Multifactor models do not explain deviations from the capm, *Journal of Financial Economics* 38, 3–28.

McLean, R David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability?, *The Journal of Finance* 71, 5–32.

McLean, R David, Jeffrey Pontiff, and Akiko Watanabe, 2009, Share issuance and cross-sectional returns: International evidence, *Journal of Financial Economics* 94, 1–17.

McLeod, Allan I, and William K Li, 1983, Diagnostic checking arma time series models using squared-residual autocorrelations, *Journal of time series analysis* 4, 269–273.

Moskowitz, Tobias J, and Mark Grinblatt, 1999, Do industries explain momentum?, *The Journal of Finance* 54, 1249–1290.

Nagelkerke, Nico JD, et al., 1991, A note on a general definition of the coefficient of determination, *Biometrika* 78, 691–692.

Newey, Whitney K, and Kenneth D West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica: Journal of the Econometric Society* 703–708.

Newey, Whitney K, and Kenneth D West, 1994, Automatic lag selection in covariance matrix estimation, *The Review of Economic Studies* 61, 631–653.

Novy-Marx, Robert, 2010, Operating leverage, *Review of Finance* 15, 103–134.

Novy-Marx, Robert, 2012, Is momentum really momentum?, *Journal of Financial Economics* 103, 429–453.

Novy-Marx, Robert, 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics* 108, 1–28.

Novy-Marx, Robert, and Mihail Velikov, 2015, A taxonomy of anomalies and their trading costs, *The Review of Financial Studies* 29, 104–147.

Ortiz-Molina, Hernán, and Gordon M Phillips, 2014, Real asset illiquidity and the cost of capital, *Journal of Financial and Quantitative Analysis* 49, 1–32.

Palazzo, Berardino, 2012, Cash holdings, risk, and expected returns, *Journal of Financial Economics* 104, 162–185.

Penman, Stephen H, Scott A Richardson, and Irem Tuna, 2007, The book-to-price effect in stock returns: accounting for leverage, *Journal of Accounting Research* 45, 427–467.

Piotroski, Joseph D, 2000, Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research* 1–41.

Politis, Dimitris N, and Joseph P Romano, 1992, A circular block-resampling procedure for stationary data, *Exploring the limits of bootstrap* 263–270.

Politis, Dimitris N, and Joseph P Romano, 1994, The stationary bootstrap, *Journal of the American Statistical association* 89, 1303–1313.

Pontiff, Jeffrey, and Artemiza Woodgate, 2008, Share issuance and cross-sectional returns, *The Journal of Finance* 63, 921–945.

Richardson, Scott A, Richard G Sloan, Mark T Soliman, and Irem Tuna, 2006, The implications of accounting distortions and growth for accruals and profitability, *The Accounting Review* 81, 713–743.

Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *The Journal of finance* 39, 1127–1139.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf, 2008, Formalized data snooping based on generalized error rates, *Econometric Theory* 24, 404–447.

Romano, Joseph P, and Michael Wolf, 2006, Improved nonparametric confidence intervals in time series regressions, *Nonparametric Statistics* 18, 199–214.

Rouwenhorst, K Geert, 1998, International momentum strategies, *The journal of finance* 53, 267–284.

Rouwenhorst, K Geert, 1999, Local return factors and turnover in emerging stock markets, *The journal of finance* 54, 1439–1464.

Schmidt, Peter Steffen, Urs Von Arx, Andreas Schrimpf, Alexander F Wagner, and Andreas Ziegler, 2017, On the construction of common size, value and momentum factors in international stock markets: A guide with applications, *Swiss Finance Institute Research Paper* .

Shao, Xiaofeng, and Dimitris N Politis, 2013, Fixed b subsampling and the block bootstrap: improved confidence sets based on p-value calibration, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 161–184.

Shumway, Tyler, 1997, The delisting bias in crsp data, *The Journal of Finance* 52, 327–340.

Sloan, Allan, 1996, Create an account or log in, *Accounting review* 71, 289–315.

Soliman, Mark T, 2008, The use of dupont analysis by market participants, *The Accounting Review* 83, 823–853.

Spiess, D Katherine, and John Affleck-Graves, 1995, Underperformance in long-run stock returns following seasoned equity offerings, *Journal of Financial Economics* 38, 243–267.

Storey, John D, 2002, A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 479–498.

Sullivan, Ryan, Allan Timmermann, and Halbert White, 1999, Data-snooping, technical trading rule performance, and the bootstrap, *The journal of Finance* 54, 1647–1691.

Sullivan, Ryan, Allan Timmermann, and Halbert White, 2001, Dangers of data mining: The case of calendar effects in stock returns, *Journal of Econometrics* 105, 249–286.

Sun, Yixiao, Peter CB Phillips, and Sainan Jin, 2008, Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing, *Econometrica* 76, 175–194.

Thomas, Jacob K, and Huai Zhang, 2002, Inventory changes and future returns, *Review of Accounting Studies* 7, 163–187.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tibshirani, Ryan J, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani, 2016, Exact post-selection inference for sequential regression procedures, *Journal of the American Statistical Association* 111, 600–620.

Titman, Sheridan, KC John Wei, and Feixue Xie, 2004, Capital investments and stock returns, *Journal of financial and Quantitative Analysis* 39, 677–700.

Titman, Sheridan, KC John Wei, and Feixue Xie, 2013, Market development and the asset growth effect: International evidence, *Journal of Financial and Quantitative Analysis* 48, 1405–1432.

Tuzel, Selale, 2010, Corporate real estate holdings and the cross-section of stock returns, *The Review of Financial Studies* 23, 2268–2302.

Ulbricht, Niels, and Christian Weiner, 2005, Worldscope meets compustat: A comparison of financial databases .

Watanabe, Akiko, Yan Xu, Tong Yao, and Tong Yu, 2013, The asset growth effect: Insights from international equity markets, *Journal of Financial Economics* 108, 529–563.

White, Halbert, 2000, A reality check for data snooping, *Econometrica* 68, 1097–1126.

Whited, Toni M, and Guojun Wu, 2006, Financial constraints risk, *The Review of Financial Studies* 19, 531–559.

Yan, Xuemin, and Lingling Zheng, 2017, Fundamental analysis and the cross-section of stock returns: A data-mining approach, *The Review of Financial Studies* 30, 1382–1423.

Zhang, X, 2006, Information uncertainty and stock returns, *The Journal of Finance* 61, 105–137.