Missing Financial Data[☆]

Svetlana Bryzgalova

London Business School, CEPR, sbryzgalova@london.edu

Sven Lerner

Stanford University, Institute for Computational and Mathematical Engineering, svenl@stanford.edu

Martin Lettau

University of California at Berkeley, Haas School of Business, NBER, CEPR, lettau@berkeley.edu

Markus Pelger

Stanford University, Department of Management Science & Engineering, mpelger@stanford.edu

Abstract

Missing data is a prevalent, yet often ignored, feature of company fundamentals. In this paper, we document the structure of missing financial data and show how to systematically deal with it. In a comprehensive empirical study we establish four key stylized facts. First, the issue of missing financial data is profound: it affects over 70% of firms that represent about half of the total market cap. Second, the problem becomes particularly severe when requiring multiple characteristics to be present. Third, firm fundamentals are not missing-at-random, invalidating traditional ad-hoc approaches to data imputation and sample selection. Fourth, stock returns themselves depend on missingness. We propose a novel imputation method to obtain a fully observed panel of firm fundamentals. It exploits both time-series and cross-sectional dependency of firm characteristics to impute their missing values, while allowing for general systematic patterns of missing data. Our approach provides a substantial improvement over the standard leading empirical procedures such as using cross-sectional averages or past observations. Our results have crucial implications for many areas of asset pricing.

Keywords: Missing data, firm characteristics, PCA, factor model, big data, asset pricing
JEL classification: C14, C38, C55, G12
This draft: January 23, 2023
First draft: March, 22, 2022

[☆]We thank Serhiy Kozak (discussant), Allan Timmermann, Michael Weber (discussant), Guofu Zhou (discussant), and seminar and conference participants at UC San Diego, EPFL, University of Lausanne, University of Maryland, Stanford, NBER Big Data and Securities Markets Conference, Annual Meeting of the American Finance Association, the AI & Big Data in Finance Research Forum, NBER Forecasting and Empirical Methods Summer Institute, NBER-NSF Time-Series Conference, the Annual Meeting of the American Economic Association, Machine Learning and Quantitative Finance Workshop at Oxford, RCEA Big Data and Machine Learning Conference, the California Econometric Conference, German Economists Abroad and the International Conference on Computational and Financial Econometrics.

1. Introduction

This paper studies a widespread yet little-researched phenomenon in finance: missing data in firm fundamentals. Firm characteristics are the cornerstone of academic research in asset pricing, investment, and corporate finance. Yet, the issue of missing data is usually ignored, and most studies simply exclude firms with missing observations. The standard source of fundamental firm-level data is the Compustat database, which includes over 1,000 individual variables. Most firm characteristics used in asset pricing combine Compustat variables with information in the Center for Research in Security Prices (CRSP) database. Many Compustat variables are sparsely populated; for example, Koh and Reeb (2015) report that R&D information of 42% of all firms is missing between 1980 and 2006.[1] The coverage of other important variables, such as current assets and liabilities, physical assets, investment, profits, taxes, among others, is also limited, while other variables are present for almost all firms.[2] As a result, the patterns of "missingness" vary substantially across characteristics.

Missing characteristic data has several potential effects for asset pricing. First, it reduces the number of stocks in portfolios that are constructed by sorts on characteristics. Second, the set of stocks in portfolios may vary by characteristic, which could make comparisons across factors difficult. Third, the performance of factor premia might be affected if firm fundamentals are not missing at random. For example, consider two characteristics, A and B. For characteristic A missing observations are distributed randomly and independently of other characteristics. However, observations of characteristic B are more likely to be missing for small stocks than for large stocks. If stock size is a priced factor, returns of portfolios based on univariate sorts on A versus B will yield biased results. In this simple example, double-sorting on size could partially rectify the bias, however, such solutions are infeasible if the distribution of missing observations is more complex in cross-section and time series.

This paper has three objectives: *(i)* provide a comprehensive analysis of missing data in 45 asset pricing characteristics, *(ii)* estimate an econometric model for imputing missing values, and, *(iii)* analyze how missingness affects returns conditional on characteristics. First, we find that the issue of missing data is profound in several dimensions. While the frequency of missing data is particularly severe until the early 1980s, missing data is still prevalent in more recent data. For example, through the 2000s, over 75% of all stocks, accounting for over half the market cap, have missing ob-

[1] We confirm their finding and find similar results in our updated sample.

[2] Compustat codes ACT, LCT, PPEGT, CAPX, GP, and variables starting with TX.

servations. Moreover, while the frequency of missing observations decreases with firm size, even the largest firms are affected. Second, the problem becomes particularly severe when requiring multiple present characteristics. Third, firm fundamentals are not missing-at-random, and have complicated dependency in both time series and cross-section; as a result, imputation based on simple cross-sectional averages or focusing on a fully balanced panel of observations, may lead to a significant bias in empirical findings. Fourth, stock returns depend on missingness and are different for the subset stock with fully observed characteristics.

Based on the documented structure of missing data, we propose a novel imputation method to obtain a fully observed panel of firm fundamentals. Our approach efficiently leverages the information available in the data, from both time-series and cross-section. Importantly, our imputation remains valid if the missingness depend on the dependency structure in characteristics. As a result, we show that it performs significantly better than the current standards in extensive empirical testing. Our approach to data imputation is easy to use in real-time, it is data-driven yet transparent, and could be naturally extended to other settings.

A comprehensive analysis of the issue of missing firm fundamentals should first and foremost answer the following questions: How widespread is the problem? What kind of firms are affected? What are the key empirical regularities? We establish the following stylized facts:

Fact #1: Missing financial data is very prevalent, being a feature of almost any characteristic. The number of missing fundamentals is large, both statistically and economically. Our dataset includes of the 45 the most popular and widely used characteristics in asset pricing. From the start of our sample period in 1967 until 1981, over 25% of observations across all stocks are missing, while 10% of observations are missing between 1990 and 2020. Until 1975, all stocks have at least one missing characteristic in any given year and only 25% of all stocks have no missing characteristics in any year since 2000. There is, of course, substantial heterogeneity in the cross-section and over time, with particular characteristics and time periods, for which over 90% of the data is missing. Missingness is a feature of firms, which are small and large, young and mature, those, which are profitable and financial distress.

Fact #2: The problem of missing data becomes substantially worse whenever one requires observations of multiple characteristics at the same time. A study of return predictability relying on a fully observed panel of 45 firm characteristics would have to omit over 70% of firms, representing about one half of the total market capitalization. The issue remains in subsets of characteristics.

Consider five of the most widely-studied characteristics: book-to-market (BM), earnings-to-price (EP), momentum (MOM), operating profitability (OP), and investment (INV). Between 1967 and 1980, only 50% of all stocks have a complete record of all five characteristics. The number increases to 80% towards the end of our sample, so that one-fifth of all stocks miss at least one of the five characteristic in a year. Hence, considering only firms with a fully observed set of fundamentals neglects a substantial amount of data and, as we show, leads to severe sample selection.

Fact #3: Data is not missing at random. There is strong heterogeneity and dependency in the distribution of missing observations, creating clusters both cross-sectionally and over time. Naturally, some of the missingness patterns arise mechanically, for example different fundamentals might require similar accounting variables, or young firms lack a prior history for constructing certain characteristics (e.g., momentum or long-term reversal). At the same time, there is a substantial number of characteristics missing during any stage of the life cycle of the firm. Other clusters arise because firms with missing data have a similar underlying latent structure. In particular, we note that small-cap companies generally have a higher propensity for missing data, and that more extreme realization of characteristics are often more likely to be unobserved.

Fact #4: Returns on their own depend on whether a firm has missing fundamentals. We show that returns of stocks with observed and unobserved characteristics are different, which drives a substantial selection bias, should one focus only on the data with observed characteristic values. On average, we find that stocks with a missing characteristic value have lower overall returns than their counterparts when the same variable is observed. Requiring the presence of multiple characteristics has a pronounced and complex effect on mean returns of characteristic sorted portfolios.

Our paper also provides a novel approach to the imputation of missing firm fundamentals. Any imputation method has two components. First, it requires a model for characteristics. Second, this model needs to be estimated from partially observed data. We provide a conceptual contribution to both components. First, we jointly model characteristics values in the three-dimensional space, reflecting time periods, individual firms, and the type of characteristics. This allows our characteristic model to leverage both the time-series and contemporaneous cross-sectional dependency in characteristics. Second, we can consistently estimate our model from sparsely observed data, while allowing for the complex patterns in missing data.

Imputing missing firm fundamentals is challenging for two reasons. First, characteristics are dependent, both in the cross-section and over time. For example, small stocks are more likely to be

also growth stocks, or given the strong persistence of book-to-market ratios, prior observed values contain information for future realizations. Hence, ad-hoc imputation methods like a simple cross-sectional median would incur an omitted variable bias. Omitting relevant information leads to an omitted variable bias even if observations would be missing at random. Second, characteristics are not missing completely randomly. For example, small stocks are more likely to have missing observations. Even if characteristics would not be predictable by cross-sectional information or their time-series, non-random missingness leads to a selection bias. This is a second reason why ad-hoc approaches like the median are invalid. The most challenging problem is that the latent information which can predict characteristics can also affect the missingess itself. This makes it very hard to learn a latent model for characteristics from the observed data. Flexible methods, that are estimated on the observed data, and do not account for this interplay, are also subject to a selection bias. Our approach provides a solution to all of these challenges.

First, we use a latent factor model to capture contemporaneous cross-sectional dependency in characteristics. The key benefit of our procedure is that it remains valid even in the presence of complex missing patterns. We can reliably recover the latent characteristic factor model when the probability of missing data varies over time, for different characteristics and for different stocks. In particular, we allow the missing data to depend on the factor model itself. For example, consistent with the data, our approach allows, that missing characteristic observations happen with a higher probability among smaller stocks, or stocks, whose underlying characteristic values are more extreme relative to other stocks. Our approach also allows for complex time-series patterns, including less observed values at the beginning of the sample, mixed-frequency observations and dependence on prior missing values. Second, we use a time-series model to capture the persistence in characteristics. Our model combines the cross-sectional factors and time-series observation, and hence can extract slow persistent movements from the time-series, while capturing fast changes from contemporaneous factor realizations.

We show that our imputation method strongly dominates leading conventional approaches. The most widely used imputation approach for firm characteristics is a simple cross-sectional median (of the whole market or the industry the firm belongs to). We show that our model allows to achieve a 50% reduction in the out-of-sample imputation error compared to using both types of medians. Another popular approach, especially for persistent characteristics, lies in simply using their last observed, stale values. This also leads to a subpar empirical performance, in particular, when there

are blocks of consecutively missing observations. Overall, we conclude that even though our imputation method is very simple, transparent, and easy-to-use, it uniformly dominates leading empirical approaches.

Modeling the joint dependence in characteristics also allows us to uncover new facts about the underlying structure of firm-specific characteristics. In particular, we show that they have a very pronounced cross-sectional dependence, which can be efficiently and parsimoniously captured by a six-factor model. Interestingly, this factor structure is stable over time, and the factors driving the underlying characteristic space are approximately the same for all the time periods, with a clear economic interpretation. Furthermore, our setting also provides new insights on the relative importance of time-series and cross-sectional dependency in the characteristic space. Our approach allows the data to speak, and endogenously provides predictions based on their relative information content. In particular, it shows that to effectively impute more volatile characteristics, one should put more weight on the contemporaneous cross-sectional information, while imputing observations for persistent characteristics should rely more on their prior history, whenever it is available.

Missing financial data can have a profound impact on asset pricing, depending on the application and extent of the problem. Missing firm characteristics can have two fundamental effects on asset pricing. The first effect is the selection bias. Asset pricing and investment results depends on which stock are included. Firms with missing characteristics are different from those with observed entries. Hence, using only the subsample of stocks with fully observed characteristics leads to a selection bias in asset pricing metrics. This is reflected in the substantially higher out-of-sample Sharpe ratio of the stochastic discount factor based on conditional latent factors that are extracted from all stocks instead of the non-representative subsample with fully observed data. In order to dissect and provide intuition for the selection bias we study the simplest possible object: univariate portfolio sorts. In our analysis the portfolios sorts either use a subset of stocks, which require certain characteristics to be observed, or the full set with imputed values. In many cases, requiring more characteristic observations lowers expected returns, while at the same time the lower diversification with less stocks increases the volatility, resulting in overall lower Sharpe ratios. However, as data is not missing at random, the effect can be complex.

The second effect is the imputation bias. Asset pricing pricing results depend on the imputation method. We study the fundamental problem of estimating the risk premium of characteristics from cross-sectional characteristic regressions. Biased imputation methods like the median impu-

tation lead to uniformly and substantially larger errors in asset pricing metrics compared to our more precise imputation approach. The imputation bias of the median values leads to wrong risk premia, correlations and variances of the characteristic mimicking factor portfolios. In contrast, our imputation method provides precise estimates of the risk premium and time-series of mimicking portfolios.

Closely Related Literature

There is vast literature on the topic of missing data in statistics and data science. Our review focuses only on the most closely related literature in economics and finance. The most widely recently used approaches to deal with missing data in firm fundamentals are a) cross-sectional median imputation (e.g., Kozak et al. (2020) and Gu et al. (2020)), and b) using only the subset of fully observed data (e.g., Freyberger et al. (2020) and Kelly et al. (2019)).

Naturally, our work is related to the econometrics literature on missing data in panels, with the most widespread solutions relying on the estimation of a low rank model, which is then used to impute missing values. The cross-sectional factor model, proposed in this paper, builds on the work of Xiong and Pelger (2019), who provide an all-purpose estimator for latent factors that allows for very general missing patterns. Importantly, their approach allows the missing pattern to depend on the latent factor model, which is crucial for our application. Bai and Ng (2021), Cahan et al. (2021), and Jin et al. (2021) develop alternative latent factor estimators with different assumptions on the missing pattern. The imputation of missing values in a panel is closely related to conducting causal inference in a panel, as discussed, among others, by Athey et al. (2021) and Xiong and Pelger (2019). The unobserved counterfactual outcomes can be modeled as missing values. Hence, the common challenge consists in uncovering a low-rank model that could be used to impute the missing data, when the missingness or treatment depends on unobserved confounders. In particular, a naive machine-learning prediction method is not appropriate for causal inference, if the treatment is not completely at random. Conversely, the same problem arises with imputation of the data, which needs to allow for various patterns of missingness in the estimation of the latent model.

Our empirical results and methods have direct implications for the multidimensional challenge raised by Cochrane (2011). A fast growing literature has studied asset pricing with a large number of predictors. Some representative work include Bryzgalova et al. (2019), Chen et al. (2019), Gu et al. (2020), Freyberger et al. (2020) and Kelly et al. (2019). The methods used in those papers require

the presence of multiple characteristics, and as such lead either to some form of data selection or data imputation. Our systematic study of missing firm fundamentals and the imputation tools that we provide help to further improve the work in this research direction. A noteworthy study is the work of Kaniel et al. (2021), which uses a version of our model to impute missing fundamentals in the holdings of mutual funds.

Our paper is also related to latent factor models in financial data. Usually, factor models are directly applied to a panel of returns. Representative works of estimating unconditional latent factors with some version of principal component analysis (PCA) include Connor and Korajczyk (1988), Pelger (2019) and Lettau and Pelger (2020a,b). Conditional latent factors can be estimated from returns that are either projected on characteristics in the case of Kelly et al. (2019) or on economic states in Pelger and Xiong (2021b). Our paper does not extract a factor structure in returns, but in fundamentals. Importantly, the factors are extracted from only partially observed data. Another distinguishing element is that we deal with a three-dimensional data set, instead of the conventional two-dimensional panel. This is related to Lettau (2022), who considers a fully observed three-dimensional mutual fund data set, from which he extracts a tensor factor model.

Unfortunately, there is very little work that directly addresses the problem of missing financial data. In a contemporaneous paper, Freyberger et al. (2021) also consider missing firm characteristics in asset pricing, and show how to adjust the general GMM estimation in the presence of missing data. Their work is focused on the estimation of conditional moments, with missingness modeled as a function of pre-specified cross-sectional covariates. Xiong and Pelger (2022) use methods for missing data imputation in the context of causal inference in finance. Their imputed values represent the counterfactual outcome for studying the publication effect in a panel of anomalies. Blanchet et al. (2022) analyze the trade-off between look-ahead-bias and variance in an imputation used for out-of-sample investment. The few contemporaneous papers that are closely related to our work, therefore, have very different goals and are complementary. Fundamentally, we provide a systematic study of missing data in finance, establish the magnitude of this phenomenon, its stylized features, and provides a "general purpose" solution to it, with a complete data set of firm fundamentals, which can then be used in any of the follow-up applications.

Table 1: Firm Characteristics by Category

	Past Returns				**Value**		
(1)	r2_1	Short-term momentum	Monthly	(25)	A2ME	Assets to market cap	Quarterly
(2)	r12_2	Momentum	Monthly	(26)	BEME	Book to Market Ratio	Quarterly
(3)	r12_7	Intermediate momentum	Monthly	(27)	C	Ratio of cash and short-term investments to total assets	Quarterly
(4)	r36_13	Long-term momentum	Monthly	(28)	CF	Free Cash Flow to Book Value	Quarterly
(5)	LT_Rev	Long-term reversal	Monthly	(29)	CF2P	Cashflow to price	Quarterly
				(30)	D2P	Dividend Yield	Monthly
				(31)	E2P	Earnings to price	Mixed Quart. & Monthly
	Investment			(32)	Q	Tobin's Q	Mixed Quart. & Monthly
(6)	Investment	Investment	Quarterly	(33)	S2P	Sales to price	Mixed Quart. & Monthly
(7)	NOA	Net operating assets	Quarterly	(34)	Lev	Leverage	Quarterly
(8)	DPI2A	Change in property, plants, equipment and inventory over assets	Quarterly				
(9)	NI	Net Share Issues	Quarterly		**Trading Frictions**		
				(35)	AT	Total Assets	Quarterly
	Profitability			(36)	Beta	CAPM Beta	Monthly
(10)	PROF	Profitability	Mixed Quart. & Yearly	(37)	IdioVol	Idiosyncratic volatility	Monthly
(11)	ATO	Net sales over lagged net operating assets	Quarterly	(38)	LME	Size	Monthly
(12)	CTO	Capital turnover	Quarterly	(39)	LTurnover	Turnover	Monthly
(13)	FC2Y	Fixed costs to sales	Mixed Quart. & Yearly	(40)	MktBeta	Market Beta	Monthly
(14)	OP	Operating profitability	Quarterly	(41)	Rel2High	Closeness to past year high	
(15)	PM	Profit margin	Quarterly	(42)	Resid_Var	Residual Variance	Monthly
(16)	RNA	Return on net operating assets	Quarterly	(43)	Spread	Bid-ask spread	Monthly
(17)	ROA	Return on assets	Quarterly	(44)	SUV	Standard unexplained volume	Monthly
(18)	ROE	Return on equity	Quarterly	(45)	Variance	Variance	Montly
(19)	SGA2S	Selling, general and administrative expenses to sales	Quarterly				
(20)	D2A	Capital intensity	Quarterly				
	Intangibles						
(21)	AC	Accrual	Quarterly				
(22)	OA	Operating accruals	Quarterly				
(23)	OL	Operating leverage	Quarterly				
(24)	PCM	Price to cost margin	Quarterly				

This table shows the 45 firm-specific characteristics sorted into six categories. More details on the construction are in Table B.11.

2. Missing values

2.1. Data

We obtain the data from the CRSP/Compustat universe with the usual filters for outliers and exchanges.[3] Our sample consists of 648 months from January 1967 to December 2020 and includes 22,630 individual stocks. We consider 45 characteristics related to value, investment, profitability, intangibles, past returns, and trading frictions, see Table 1. The raw characteristics are converted into centered rank quantiles and scaled to be in the $[-0.5, 0.5]$ interval.

We construct characteristics if the required variables are available in CRSP and COMPUSTAT. Otherwise, we consider a characteristic missing. Characteristics are either updated monthly or at

[3]The sample only includes stocks listed on the NYSE, NASDAQ, and AMEX exchanges (exchange codes 10, 11, 12) with share codes 1, 2, or 3 (common stock, foreign incorporated, ADR) and at least one entry in the Compustat accounting tables. We do not filter out stocks based on share price, nor do we filter out financial firms. However, we show in an extensive robustness study that our results are not affected by these choices, that is, the results are robust to including or excluding either of those subsets.

a lower frequency which is typically quarterly. For quarterly updated characteristics, we do not observe the monthly observations in-between the quarters, which are therefore mechanically missing. To avoid these mechanical effects, all our evaluation metrics for characteristics that are updated quarterly are based on quarterly data points. We are not "imputing" the months between the quarters with stale values, nor do we count those as missing values in our summary statistics.[4] However, our procedure will provide imputed values in-between the quarters and hence also provides a solution to mixed-frequency observations.

We use the most-up-to-date last observed values as current characteristics. For characteristics based on the ratio of variables with different updating frequencies, we use the most up-to-date information of each variable, and the variable with the slowest updating frequency determines the updating frequency of the characteristic. For example, the quarterly updated book-to-market ratio divides the book value from the most recent quarter by the last observed monthly market capitalization. Asset pricing applications, which condition on characteristics, usually lag characteristics by several months to ensure that the information is available to investors. Our data imputation uses the most recent information; however, we lag characteristics in asset pricing applications.

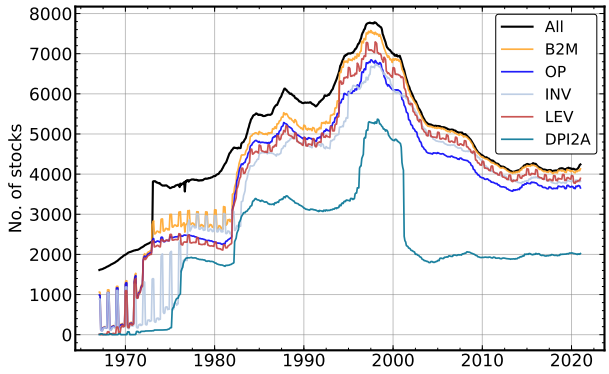
2.2. How much data is missing?

Missing financial data is prevalent, and almost all characteristics have missing observations. The number of missing fundamentals is large, both statistically and economically. Figures 1 and 2 summarize some patterns in missing values over time. The black line in Panel (a) of Figure 1 shows the number of firms in our sample over time. As is well-known, the number of listed stocks has declined over the last 25 years. At its peak in November 1997, our sample includes 7,784 stocks but only 4,241 in December 2020. The spike in January 1973 is due to the inclusion of the NASDAQ. The plot also shows the number of firms with observed values of five important characteristics: book-to-market (B2M), operating profitability (OP), investment (INV), and leverage (LEV). We also include the ratio of real investment to book value of assets (DPI2A, Lyandres et al. (2008)) since it has the most missing values among all 45 characteristics. Panel (b) shows the percentage of stocks with missing values for each of the five characteristics.

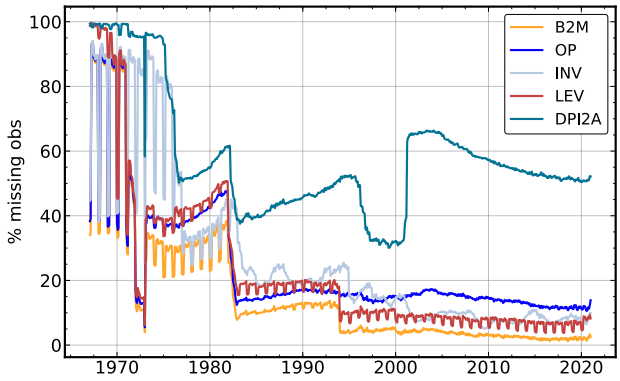
[4] Using stale values in-between observations of characteristics with low updating frequency is a form of data imputation itself. Using stale values as the actual monthly characteristics values would also lead to mechanical trivial predictability.

Figure 1: Missing Values over Time

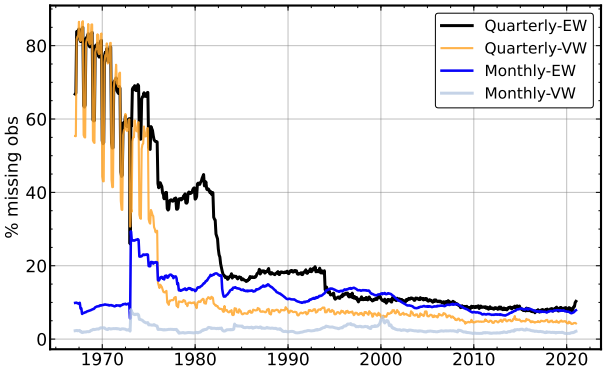
(a) Number of Stocks



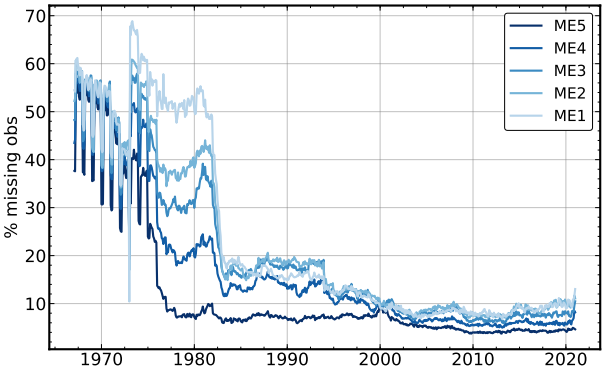
(b) Missing Percentage



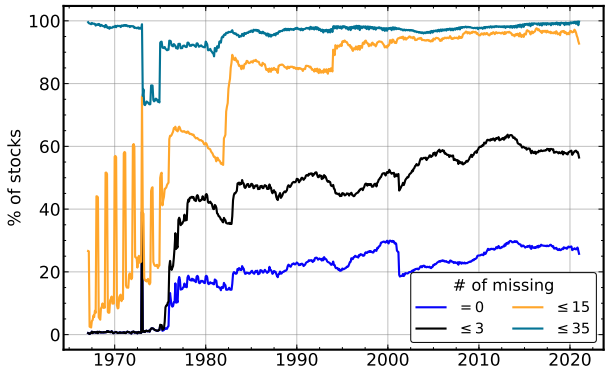
(c) Quarterly & Monthly



(d) Size Quintiles



(e) Multiple Chars.



Note: This figure summarizes missing values over time. Subfigure (a) shows the total number of stocks and those that have observed values for our five example characteristics book-to-market (B2M), operating profitability (OP), investment (INV, growth in total assets), leverage (LEV) and real investment (defined as the change in property, plants, equipment and inventory) over lagged total assets (DPI2A). Subfigure (b) shows the percentage of missing observations for the five example characteristics. Subfigure (c) plots the percentage of missing observations for quarterly and monthly updated characteristics based on equal and market capitalization-weighted averages. Subfigure (d) shows the percentage of missing observations by market capitalization quintiles. Subfigure (e) displays the proportion of missing stocks that have no missing observations or at most 3, 15 or 35 missing characteristics at a given point in time.

Panels (a) and (b) of Figure 1 show substantial cross-sectional and time variation in missingness. First, the proportion of missing values has, on average, decreased over time, which is not surprising since the coverage of COMPUSTAT has improved throughout the sample, and changes in regulations led to more comprehensive and more frequent disclosures of accounting information. Consider first the four accounting variables B2M, OP, INV, and LEV. Missing data is particularly prevalent throughout the early 1980s for all four characteristics. Between 30% and 95% of observations are missing between 1967 and 1981.[5] About 15% to 20% of observations are missing between 1982 and 1992 followed by a further decline throughout the 2000s. At the end of the sample in 2020, 14%/10%/8%/3% of OP, INV, LEV, and B2M data is missing, respectively. Fewer book-to-market observations are missing than of the other variables because its definition includes several alternatives and fall-back options if individual component variables are not in COMPUSTAT.[6]

The pattern of missing values of DPI2A (real investment-to-total assets) differs substantially from those of the other four variables. Until 1975, very few firms have real investment observations in COMPUSTAT, so DPI2A is virtually completely missing. In contrast to the other variables, the share of missing observations remains above 35% over the rest of the sample. In 2004, 67% firm observations were missing and more than half are missing in 2020. While DPI2A has the most missing observations, there are several other variables with more than 20% missing data in 2020: accruals (AC), fixed-costs-to-sales (FC2Y), operating accruals (OA), and SGAto-sales (SGA2S).

Figure 1(c) shows the time series of the share of missing values averaged across all characteristics. We form two groups of characteristics that are updated either monthly or quarterly. Price or return-based characteristics are available at a high frequency, while accounting variables are (at most) available quarterly. Consider first the equal-weighted averages in Panel (c). The time series of missingness of quarterly characteristics (black line) is similar to those found for B2M, OP, INV, and LEV in Panel (b). Before 1982, over 40% of observations are missing; between 1982 and 1992, about 20%, and between 8% and 14% afterward. Since the CRSP database has an (almost) complete record of prices and returns, there are, on average, fewer missing values for characteristics that are

[5]During this period, most stocks many report accounting variables only once per year, which accounts for the spikes in the plots. As the reporting frequency increases over the sample, this pattern largely vanishes.

[6]Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITC), minus preferred stock (PS). SH is shareholders' equity (SEQ). If missing, SH is the sum of common equity (CEQ) and preferred stock (PS). If missing, SH is the difference between total assets (AT) and total liabilities (LT). Depending on availability, PS is redemption value (item PSTKRV), liquidating value (item PSTKL), or par value (item PSTK). The market value of equity (PRC*SHROUT) is as of the current month.

updated monthly. However, many monthly characteristics require lags of prices or returns, and thus some observations are missing mechanically. For example, reversals require a return history of 60 months so that newly listed firms do not have any observations for the first five years. As a result, between 10% and 20% of monthly characteristics are missing throughout the sample. The exception is the period from 1973 to 1975 when the inclusion of the NASDAQ added many firms without a history of prices and returns.

Figure 2 shows the share of missing values of all characteristics over time in the form of heatmaps. Lighter (darker) shades correspond to lower (higher) shares of missing observations. The heatmaps reveal time-series variation as well as heterogeneity across characteristics. The frequency of missing data of most quarterly characteristics, shown in the top panel, decreases substantially in the early 1980s and again in the mid-1990s. There are several characteristics with many missing values throughout the sample: AC, DPI2A, FC2Y, OA, OP, and SGA2S. The frequency of missing values in monthly variables is directly linked to the number of lagged values that are required. The exceptions are SUV and TURN, which are based on trading volume, however, volume for many NASDAQ stocks is missing from CRSP between 1973 and 1983. Thus, the share of stock with missing values of SUV and TURN is particularly during this period, which is visible in the heatmap in Panel (b).

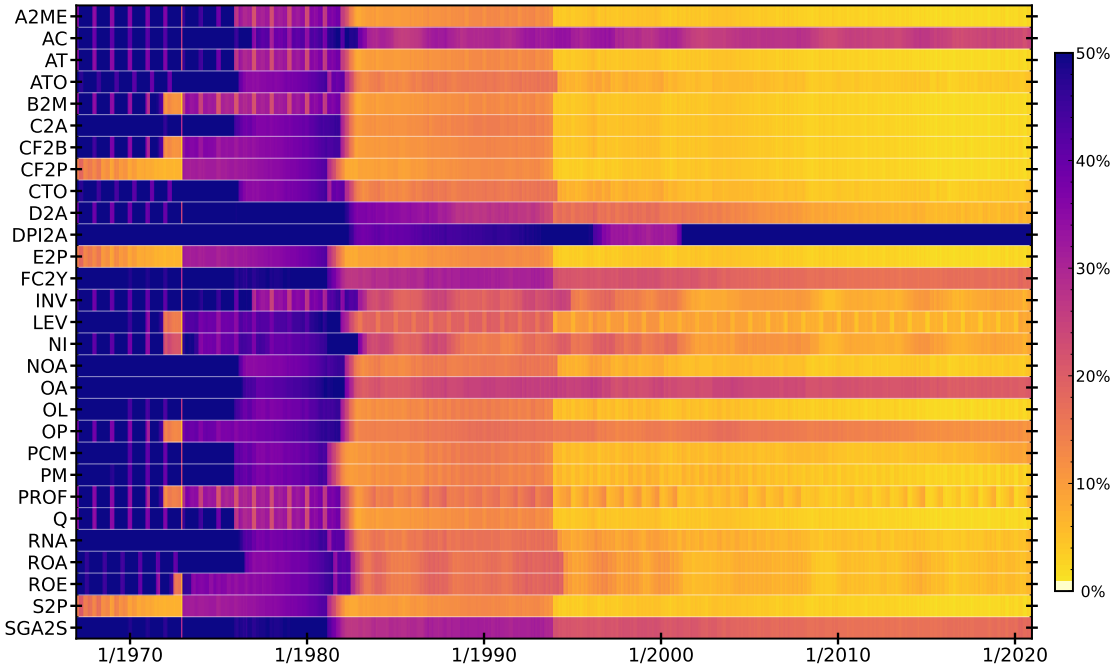
The evidence so far was based on firm counts without taking firm size into account. Figure 1(c) also shows the value-weighted percentage of missing observations for monthly (light blue) and quarterly (orange) characteristics. While the value-weighted percentage is lower than its equal-weighted counterpart, it is still substantial. In particular, quarterly updated characteristics are missing for over 10% of the market capitalization after 1977.

Figure 1(d) reports the percentage of missing observations for quintiles of market capitalization of companies. We observe that historically smaller companies used to have worse data coverage. However, in the last 20 years, small and large companies have shown similar degrees of missingness. Importantly, at no point in time is missing only due to small-cap companies.

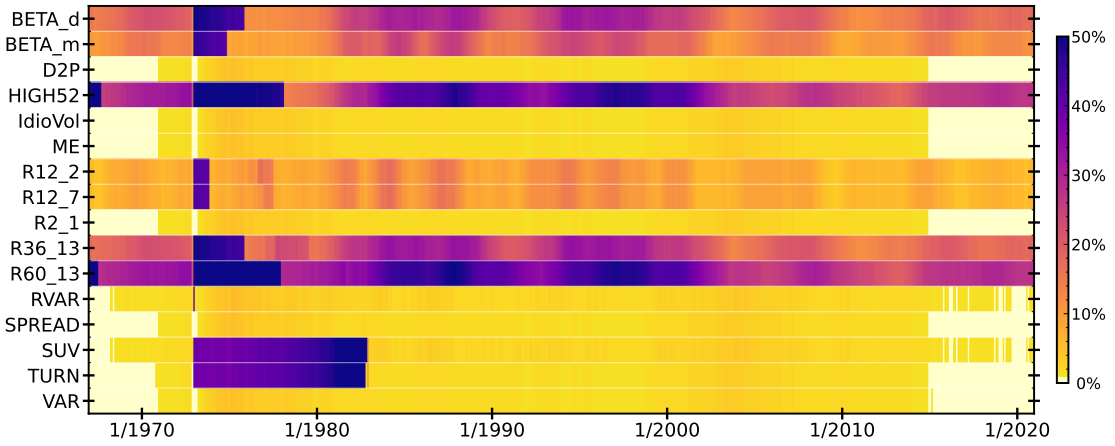
Missing data is a paramount problem whenever multiple characteristics are required. The missingness in individual characteristics largely underrepresents the severity of the problem. Figure 1(e) shows the percentage of stocks that have no characteristics, less than 3, less than 15 or less than 35 of the 45 characteristics missing. The results are striking. Over 70% of firms are missing at least some popular characteristic at any point in time. The total market cap corresponds to 48%. In other words, an application that requires all 45 characteristics to be observed neglects half of the market

Figure 2: Missing Observations over Time By Characteristics

(a) Quarterly Characteristics



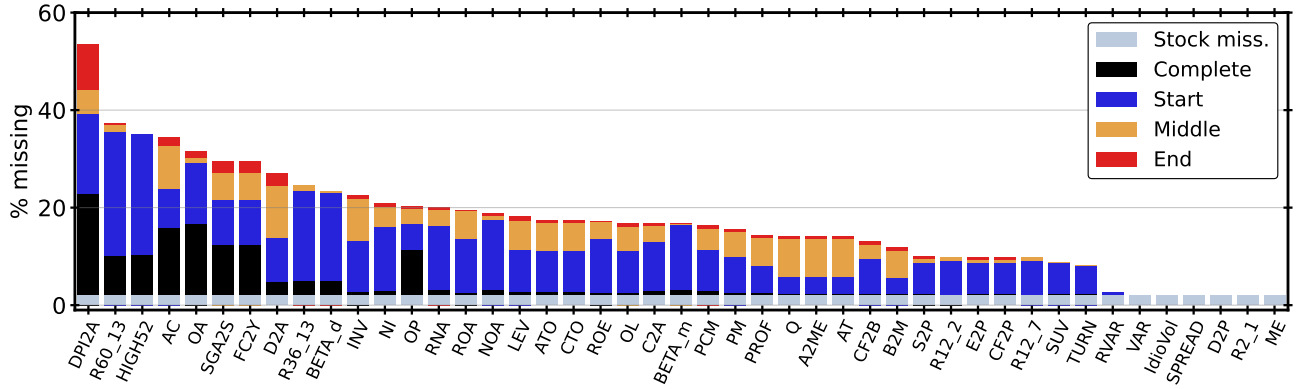
(b) Monthly Characteristics



Note: This figure is a heatmap of percentage of missing values for all 45 characteristics over time. Quarterly characteristics collect all characteristics that are updated at a frequency lower than monthly.

capitalization and 70% of the companies at any point in time. As we will show, using a fully observed panel of data may lead to severe sample selection. This can affect all applications that require a full panel of characteristics, which includes characteristic panel models, conditional latent factor models or machine learning applications.

Figure 3: Missing Observations by Characteristic



Note: This figure shows the average percentage of missing observations for each characteristic. We decompose the missing values in those missing at the start (no previous observations), the middle (some previous and future observations), the end (no further observations) and completely missing.

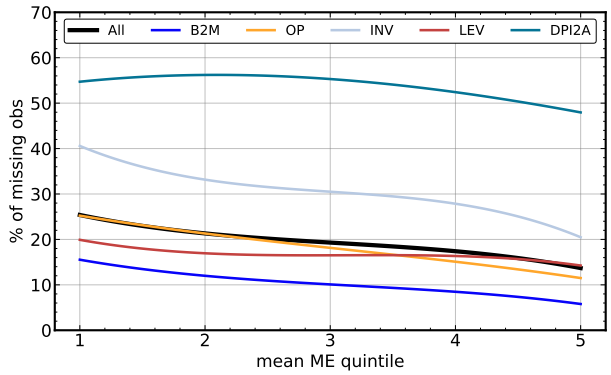
2.3. What is the structure of missingness?

In order to understand the structure of missingness, we study when, which, and for what values firm fundamentals are missing. Figure 3 displays the percentage of missing observations for each characteristic. We report if characteristics are missing at the start, at the end, or in the middle. Recall that we only include a stock in the sample when we observe its returns and at least one entry in Compustat in a given month. Missingness in the middle implies that we observe some previous and future values. Missingness at the start mechanically appears for younger firms, while missingness at the end can occur at the end of a company's life. We see that many accounting-based variables are missing after having been previously observed, which often occurs in missing time blocks. Overall we confirm that missingness is a problem for all characteristics.

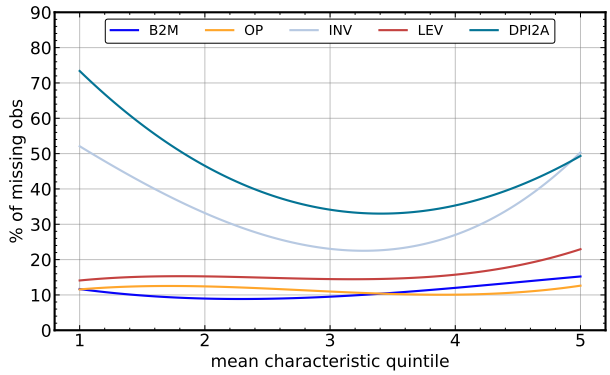
Some of the missingess patterns are purely mechanical and expected. For example, long-term reversal and momentum have by construction missing observations for a new firm without prior history. At the other extreme, market capitalization is always observed when there is a return in the prior month. Figure C.1 in the Appendix provides missing observations by characteristic pooled by stocks, which can be different than the overall averages if there is heterogeneity in the patterns for individual stocks. While the overall percentage and relative ranking seems to be quite similar, there are notable differences. Missing in the middle is less pronounced in the pooled averages, which implies that there is a smaller subset of stocks for which observations are primarily missing in the middle. Those characteristics that are based on past return observations, also constitute a larger percentage for the pooled averages. The lower panel in Figure C.1 shows the value-weighted pooled

Figure 4: Missing Observations by Characteristic Quintiles

(a) Mean ME Quintiles



(b) Mean Char. Quintiles



Note: This figure shows the percentage of missing observations for different characteristic quintiles. The left subfigure displays the missing observations for all characteristics and the example characteristics book-to-market, operating profitability, investment, leverage and change in property, plants, equipment and inventory over lagged total assets for the five size quintiles of stocks. The right subfigure presents the proportion of missing values for the five example characteristics for their corresponding characteristic quintile. The characteristic quintiles are based on the average observed characteristic value of the corresponding stock.

averages with similar findings.

Next, we investigate how values of characteristics interact with the frequency of missing values. We sort stocks into quintiles of a characteristic and compute the share of missing values among stocks in each quintile. Figure 4(a) shows the percentage of missing observations by size quintiles. The black line shows the average share of missing values across all 45 characteristics and shows that smaller stocks have more missing values than large stocks, however, the difference is modest. Even in the largest size quintile 15% of the characteristics are missing. The downward slope is present in most characteristics, but the dependency on size is heterogeneous. The size effect on leverage and DPI2A is almost flat, while it is more pronounced for investment.

Next, we compute how the frequency of missing values of a characteristic depends on characteristic values themselves. Obviously, we do not observe the actual characteristic realizations when they are missing. Hence, we study the patterns of missingness for firms that are on average in a certain characteristic quintile. In more detail, for each characteristic, we sort stocks into quintiles based on their observed values and compute the proportion of missing values of the characteristic of the stocks in each quintile. The results are shown in Panel (b) of Figure 4. The black line shows the mean across all characteristics. Its convex shape implies that stocks with low and high characteristics have more missing values than stocks with average characteristics. The difference

Table 2: Logistic regressions explaining missingess

D2P	IdioVol	ME	R2_1	SPREAD	TURN	VAR	FE	Last Val	Missing Gap	train AUC	test AUC
Missing at the beginning											
1.85*** [239.78]	2.29*** [33.52]	-1.25*** [-143.73]	0.06*** [9.99]	0.6*** [62.78]	0.68*** [113.83]	-1.74*** [-25.76]	F	F	F	0.49	0.50
1.96*** [181.16]	0.66*** [8.59]	-0.58*** [-56.17]	-0.08*** [-11.63]	0.61*** [53.7]	0.86*** [122.41]	-0.38*** [-5.05]	F	F	0.06 [450.24]	0.61	0.63
							T	F	F	0.69	0.73
							T	F	0.01 [153.85]	0.69	0.72
0.47*** [37.55]	-1.30*** [-13.75]	-0.64*** [-51.01]	0.11*** [13.78]	-0.02*** [-1.69]	-0.10*** [-11.08]	0.91*** [9.79]	T	F	0.01 [146.23]	**0.71**	**0.74**
Missing in the middle											
0.59*** [268.86]	0.63*** [28.28]	-0.44*** [-141.07]	0.04*** [18.04]	0.52*** [151.52]	0.27*** [118.95]	-0.82*** [-37.19]	F	F	F	0.55	0.52
							T	F	F	0.78	0.82
							T	5.37 [961.19]	F	0.92	0.96
							T	0.06 [137.87]	-4.74 [-279.65]	0.93	0.96
0.3*** [26.89]	-0.4*** [-3.3]	-0.65*** [-42.21]	0.07*** [7.09]	0.39*** [24.39]	-0.26*** [-24.38]	0.49*** [4.06]	T	0.06 [139.69]	-4.9 [-270.68]	**0.94**	**0.97**
Missing at the end											
0.63*** [395.91]	0.48*** [29.86]	-0.58*** [-258.87]	0.03*** [18.63]	0.44*** [178.27]	0.06*** [38.04]	-0.63*** [-39.56]	F	F	F	0.61	0.55
							T	F	F	0.80	0.83
1.52*** [461.06]	0.98*** [27.55]	-0.89*** [-196.39]	0.06*** [19.65]	0.43*** [88.27]	-0.17*** [-49.96]	-1.07*** [-30.44]	T	F	F	**0.82**	**0.83**

This table shows the results of logistic regressions to predict the missingness of individual stock characteristics. We report the results for different sets of explanatory variables for characteristics missing at the beginning, in the middle or at the end. The values of the seven characteristics D2P, IdioVol, ME, R2_1, SPREAD, TURN and VAR are always observed and hence can be included in the regressions. We also include characteristic fixed effects (FE), an indictor variable if the last characteristic value was observed, and the length of a missingness if the last value is not observed. The area under the curve (AUC) measures the accuracy of the logistic regression. The regression is pooled over time, stocks and characteristics. The model is estimated on the training data (1988-1998) and evaluated out-of-sample on the test data (1999-2020). We also include the z-scores of the regression coefficients. Stars indicate the statistical significance, where *** corresponds to 1% significance.

is economically large; missing values of stocks at the extreme of the characteristic distributions are twice as frequent as for stocks at the center (29% vs. 14%). This pattern is true for the majority of individual characteristics, see DPI2A, INV, and to a lesser extent, B2M in Panel (b). These results suggest that missing values are not distributed randomly and depend on characteristics themselves. In this sense, missingness is endogenous.

In order to better understand the structure of missingness, we predict missigness of individual firm characteristics with logistic regressions. Table 2 shows the results for different sets of explana-

tory variables. We report separate regressions for characteristics missing at the beginning, in the middle or at the end, as, for example, missingness at the end of a company's life can be more related to firm fundamentals than mechanical missingness for new firms.[7] We explain missingness with the seven characteristics that are always observed, an indicator if the last observation was missing, and the length of the missingness if the last observation was missing. We also allow for characteristics fixed effect. The category missing in the middle is the most important for our analysis and represents the largest part of this sample. It contains all observed and missing characteristic values that have at least one prior observation and a last observation. The area under the curve (AUC) measures the accuracy of the prediction. Our best models achieve an out-of-sample AUC of 0.97, which means that we explain a large part of the missing pattern and that the logistic regression captures important features.

First, characteristic fixed effects are crucial in the prediction, confirming our previous finding that missingness is heterogeneous. Second, the realization of contemporaneous characteristics is highly significant in predicting missingness. As we will show, characteristics are cross-sectionally correlated, which confirms the endogeneity in missingess. Last but not least, missingness is correlated over time. The negative sign on the length of a missing gap indicates that missing data is likely to appear in blocks. Table B.2 reports the number of missing blocks and their mean and median length. Indeed, most missing values cluster together and have an average length of around one to two years.

The structure of missingness has also important implications for how to impute missing values. First, imputation methods need to allow for different information sets. If no prior values are observed, it is obviously not possible to condition on prior observations in the imputation method. Second, stocks with different fundamentals can be more likely to have missing values. Hence, an imputation methods needs to allow the probability of missingness to be heterogenous and depend on fundamentals. If we model characteristics with a factor model, this implies that the joint distribution

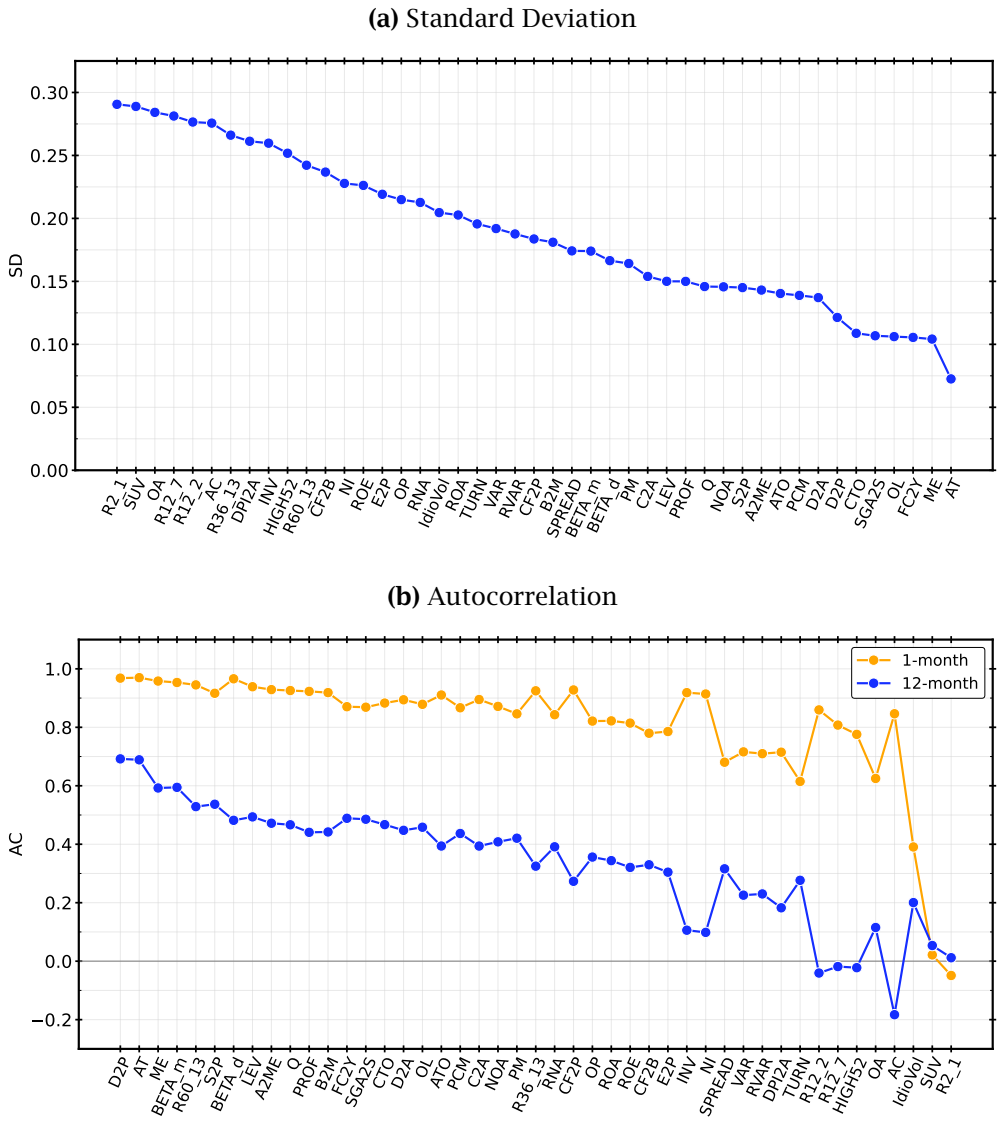
[7] For missing at the beginning, we consider the set of all characteristic observations that are missing at the beginning of the sample and include the first time a characteristic is observed. Hence, the results for missing at the beginning essentially predict the change from missing at the beginning to being observed for the first time. For missing at the end, we include only the set of characteristic observations that end in terminal missingness. In more detail, we include the set of only observed values (after potentially missing values) and the first terminal missing value. Thus, the results for missing at the end predict the change from being observed to be missing completely. Missing in the middle excludes the subset of characteristic observations that are missing at the beginning (no prior observations) and at the end (no further observations after missingness). Note that this means that the same stock for the same characteristic can have part of its time-series included in missing at the beginning (first set of observtions), missing in the middle (all observed and missing values in the middle) and missing at the end (last block of observations).

of missingness can depend on the factor model itself.

2.4. Characteristics Dependency

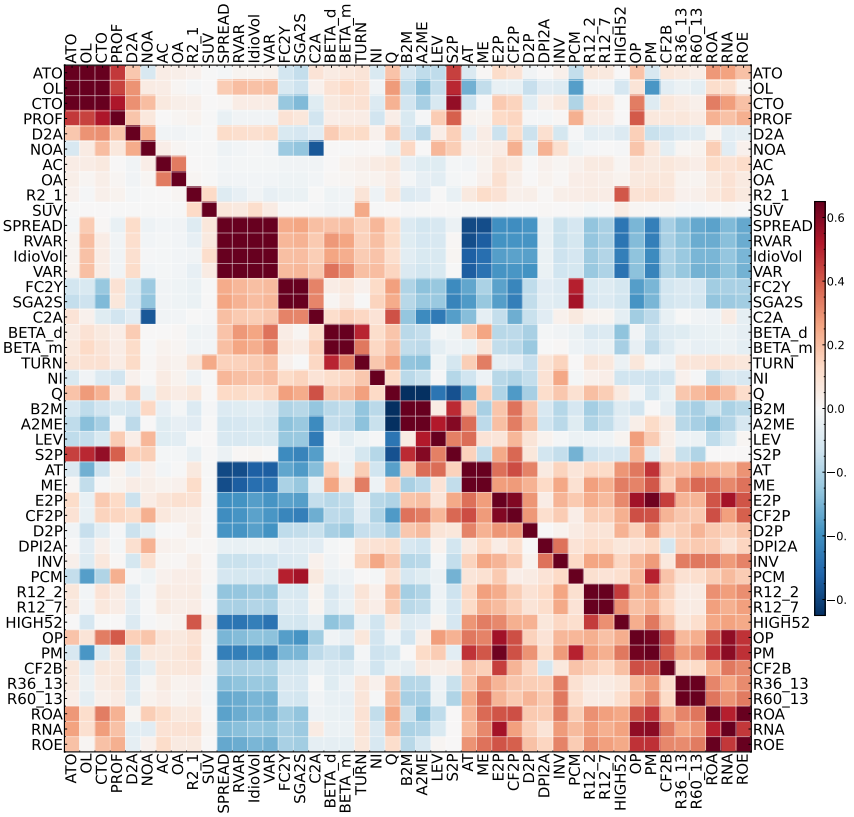
Characteristics are dependent over time and cross-sectionally on other characteristics. This dependency establishes the foundation of any method that tries to model or predict characteristics. It implies that observing the realizations of other characteristics or prior values allows us to predict the realizations of unobserved characteristics.

Figure 5: Time-series Variation and Dependency of Characteristic Ranks



Note: This figures presents the time-series variation and dependency of characteristic ranks. The top figure shows the sorted standard deviation over time for each characteristic. The bottom figures summarizes the 1-month and 12-months autocorrelation coefficients for each characteristic.

Figure 6: Heatmap of Pairwise Correlation



Note: This figure shows the pairwise correlations across time and stocks for each characteristic. The time period is the sample from 1977-2020.

Many characteristics are very persistent. Figure 11 shows the 45 characteristics sorted by their standard deviation and autocorrelation. As expected, many characteristics, for example market capitalization and total assets, are very slowly moving and highly serially correlated. This implies that the previous values of these persistent characteristics have information for their future realizations. In fact, the autocorrelation of several characteristics is close to one, implying that their previous value would be a good predictor. This predictability persists over longer horizons. Indeed, the 12-months autocorrelation is still over 0.4 for around half of the characteristics. However, we also find that a number of characteristics, primarily based on prior returns like short-term momentum or idiosyncratic volatility, are highly volatile and seem to show negligible time-series predictability. Hence, the persistence is quite heterogenous. Overall, we conjecture that disregarding time dependency when imputing missing values might lead to an omitted variable bias.

Characteristics are cross-sectionally correlated. Figure 6 shows pairwise correlations in charac-

teristics averaged over time and stocks. We observe obvious clusters of correlations. These could be interpreted as exposure to common characteristic factors. Hence, disregarding observed values of other characteristics when imputing missing values could lead to an additional omitted variable bias. For example, small stocks are more likely to be growth stocks. Therefore, imputing a missing book-to-market value of a small company with a market median, would inherently lead to a bias. The clusters of cross-sectional dependence seem to form around different groups of characteristics. Not surprisingly, characteristics based on past returns exhibit correlations. Similarly, we observe a dependency cluster among trading friction or value characteristics. However, the dependency is complex and requires a sophisticated tool to capture it from the data.

The general dependency patterns between characteristics seem to be stable over time. We have observed in Figure 1 that the frequency of missing characteristics changes drastically around the year 1977. Figure C.2 shows the pairwise correlations in characteristics averaged over time and stocks from 1967 to 1976, while Figure 6 is based on 1977-2020. While the strength of the dependency seems to vary, the location of correlation clusters stays the same. This would be consistent with a factor model in the characteristic space, where the factors stay the same, but the scale of the exposure to those factors can vary.

3. Model

The estimation of a model for the imputation of missing values faces two fundamental challenges. First, it should take advantage of all available information. An ad-hoc imputation method, such as the cross-sectional median, would incur an omitted variable bias. Omitting relevant latent information also leads to an omitted variable bias, even if observations were missing at random. Our solution to the problem is to extract all latent cross-sectional information from the data instead of pre-specifying a set of covariates. In other words, we let the data speak about which contemporaneous information can best predict a given characteristic. Second, the model for characteristics, that is estimated on the observed data, needs to be valid on the unobserved data as well. This is the crucial aspect where our approach stands out from the related literature. Even when the missingness depends in a complex way on latent information extracted with our model, our predictions provide correct imputed values for the unobserved entries. Flexible methods that are estimated on the observed data and do not account for the dependency between missingness and the information that predict characteristics are subject to a selection bias. In particular, as data is not missing randomly, ad-hoc approaches

suffer from a selection bias in addition to the omitted variable bias.

Our data set of month/stock/characteristic observations forms a three-dimensional vector space:

$$C_{i,t,l} \qquad \text{with } i = 1, ..., N_t, t = 1, ..., T \text{ and } l = 1, ..., L.$$

The data have a cross-sectional dimension of N_t stocks, a time-series dimension T, and the number of different characteristics L. The typical dimensions are around $N_t = 6,000$, $T = 600$ and $L = 45$. The notation of an upper index selects a matrix of this three dimensional array. For example, we denote by

$$C^t_{i,l} \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L$$

the $N_t \times L$ matrix of characteristics at time t.

Based on our empirical findings above, we use the time-series dependency and cross-sectional dependencies in characteristics to predict missing values. The fundamental problem is to estimate a low dimensional model to predict a characteristic value with past, (possibly) future, and other contemporaneous cross-sectional information. The prediction model is used to impute missing values. We use an estimation approach that allows us to estimate the parameters of the prediction model in the presence of missing values.

3.1. Cross-Sectional Information

An essential building block for our model is based on a cross-sectional factor model. We start by estimating a low-dimensional cross-sectional factor model by PCA for each month t:

$$C^t_{i,l} = F^t_i {\Lambda^t_l}^\top + e^t_{i,l} \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

The upper index t indicates that we can have separate factor models for each time t. We assume a K factor model, i.e. $F^t \in \mathbb{R}^{N_t \times K}$ and $\Lambda^t \in \mathbb{R}^{L \times K}$. Without missing values, we can estimate F^t and Λ^t as the singular values of C^t, i.e. we apply a simple PCA to ${C^t}{C^t}^\top$. More specifically, we obtain $F^t \in \mathbb{R}^{N_t \times K}$ as the eigenvectors of the K largest eigenvalues of the $N_t \times N_t$ matrix

$$\frac{1}{L} \sum_{l=1}^{L} C^t_l {C^t_l}^\top.$$

The different entries in this "characteristic covariance" matrix indicate how close two different stocks are. Two stocks with very similar characteristics have a high "characteristic covariance". In the presence of missing values, we use the approach of Xiong and Pelger (2019) and estimate F^t as the eigenvectors of the K largest eigenvalues of

$$\tilde{\Sigma}_{i,j}^{\text{XS},t} = \frac{1}{|Q_{i,j}^t|} \sum_{l \in Q_{i,j}^t} C_{i,l}^t C_{j,l}^t,$$

where $Q_{l,p}^t$ is the set of all characteristics which are observed for the two stocks i and j at time t. By construction $|Q_{i,j}^t| \leq L$. The characteristic loadings follow from a regression on the estimated $\hat{F}^t$,

$$\hat{\Lambda}_l^t = \left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t \hat{F}_i^{t\top} \right)^{-1} \left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t C_{i,l}^t \right),$$

where $W_{i,l}^t = 1$ if characteristic l is observed for stock i at time t and $W_{i,l}^t = 0$ otherwise. Hence, this is simply a linear regression using only observed values. Xiong and Pelger (2019) provide the formal theory and show that this estimator is consistent under general assumptions on the approximate factor model and the missing pattern. The setup is a large dimensional panel, that is, both N_t and L go to infinity, but at general and possibly different rates. An approximate factor model assumes that asymptotically most of the dependency is captured by the factors, while the "idiosyncratic" characteristic errors $e_{i,l}^t$ are only weakly dependent. This setup allows for a different factor model at each time t and hence is a *local* model.

Based on our empirical findings, the "loadings" Λ are close to constant over time, which results in the model

$$C_{i,l}^t = F_i^t \Lambda_l^\top + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

Under the assumption of constant "loadings", we can estimate Λ from a pooled regression

$$\hat{\Lambda}_l = \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^t F_i^t F_i^{t\top} \right) \right)^{-1} \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W_{i,l}^t F_i^t C_{i,l}^t \right) \right).$$

While, in principle, the factors can be estimated as in the local model, they need to be appropriately rotated to represent the same factors over time. Appendix A discusses the implementation. The global $\hat{\Lambda}$ can be interpreted as characteristic "portfolio weights" to construct the latent factors $\hat{F}_i^t$.

As this estimation uses the full data, it represents a *global* model. If the loadings are constant over time, the global model is more precise as it uses substantially more data.

3.2. Time-Series Information

We combine the XS (cross-sectional) information with TS (time-series) information. Given an estimate of the contemporaneous XS factors $\hat{F}^t \in \mathbb{R}^{N_t \times K}$, we combine those with past and (possibly) future time-series information to predict contemporaneous characteristics. We consider a backward-cross-sectional model (B-XS) with only the past observed information and a backward-forward-cross-sectional model (BF-XS), which combines past and future information. Both models are based on regressions to estimate either $\beta^{l,\text{B-XS}} \in \mathbb{R}^{K+1}$ or $\beta^{l,\text{BF-XS}} \in \mathbb{R}^{K+2}$:

B-XS Model:

$$\hat{C}_{i,t}^{l,\text{B-XS}} = \beta^{l,\text{B-XS}\top} \begin{pmatrix} C_{i,t-1}^{l} & \hat{F}_{i,1}^{t} & \cdots & \hat{F}_{i,K}^{t} \end{pmatrix}$$

BF-XS Model:

$$\hat{C}_{i,t}^{l,\text{BF-XS}} = \beta^{l,\text{BF-XS}\top} \begin{pmatrix} C_{i,t-1}^{l} & C_{i,t+1}^{l} & \hat{F}_{i,1}^{t} & \cdots & \hat{F}_{i,K}^{t} \end{pmatrix}.$$

The framework includes several important special cases:

(a) Time-series AR(1) model (B): $\beta^{l,\text{B-XS}} = \begin{pmatrix} \beta^{\text{B}} & 0 & \cdots & 0 \end{pmatrix}$.

(b) Last observed value (PV): $\beta^{l,\text{B-XS}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$.

(c) Cross-sectional median: $\beta^{l,\text{B-XS}} = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}$ (as we have centered the rank quantiles at 0).

We estimate the β vectors in a regression using the stacked observed values. This means that we use all $C_{i,t}^{l}$ with observed $C_{i,t-1}^{l}$ (respectively $C_{i,t-1}^{l}$ and $C_{i,t+1}^{l}$) and stack them together in a large vector. Without missing values, this vector would have the dimension $\sum_{t=1}^{T} N_t$. For each characteristics l, we obtain the vector $\beta^{l,\text{B-XS}} \in \mathbb{R}^{K+1}$ and $\beta^{l,\text{BF-XS}} \in \mathbb{R}^{K+2}$. In the local model, we use the local factors and the observed characteristics for the time t to obtain the local $\hat{\beta}^t$, while the global model uses globally estimated factors in a regression that stacks all characteristics over time. For a given set of cross-sectional and time-series information in the vector $X_i^{l,t}$ we obtain the local model from the

Table 3: Different Imputation Methods

Method	Estimation
Backward-Forward-XS (BF-XS)	$\hat{C}^{\text{BF-XS}}_{i,t} = (\hat{\beta}^{\text{BF-XS}})^\top \begin{pmatrix} C^l_{i,t-1} & C^l_{i,t+1} & \hat{F}^l_{i,1} & \cdots & \hat{F}^l_{i,K} \end{pmatrix}$
Backward-XS (B-XS)	$\hat{C}^{\text{B-XS}}_{i,t} = (\hat{\beta}^{\text{B-XS}})^\top \begin{pmatrix} C^l_{i,t-1} & \hat{F}^l_{i,1} & \cdots & \hat{F}^l_{i,K} \end{pmatrix}$
Forward-XS (F-XS)	$\hat{C}^{\text{F-XS}}_{i,t} = (\hat{\beta}^{\text{F-XS}})^\top \begin{pmatrix} C^l_{i,t+1} & \hat{F}^l_{i,1} & \cdots & \hat{F}^l_{i,K} \end{pmatrix}$
Cross-sectional (XS)	$\hat{C}^{\text{XS}}_{i,t} = (\hat{\beta}^{\text{XS}})^\top \begin{pmatrix} \hat{F}^l_{i,1} & \cdots & \hat{F}^l_{i,K} \end{pmatrix}$
Time-series (B)	$\hat{C}^{\text{B}}_{i,t} = (\hat{\beta}^{\text{B}})^\top \begin{pmatrix} C^l_{i,t-1} \end{pmatrix}$
Previous value (PV)	$\hat{C}^{\text{PV}}_{i,t} = C^l_{i,t-1}$
Cross-sectional median	$\hat{C}^{\text{median}}_{i,t} = 0$

Note: This table summarizes the different estimation approaches. Each estimation approach has a local and global version.

local regression

$$\hat{\beta}^{l,t} = \left(\sum_{i=1}^{N_t} W^t_{i,l} X^{l,t}_i X^{l,t\top}_i \right)^{-1} \left(\sum_{i=1}^{N_t} W^t_{i,l} X^{l,t}_i C^l_{i,t} \right),$$

and the global model from a global regression

$$\hat{\beta}^{l} = \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W^t_{i,l} X^{l,t}_i X^{l,t\top}_i \right) \right)^{-1} \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} W^t_{i,l} X^{l,t}_i C^l_{i,t} \right) \right).$$

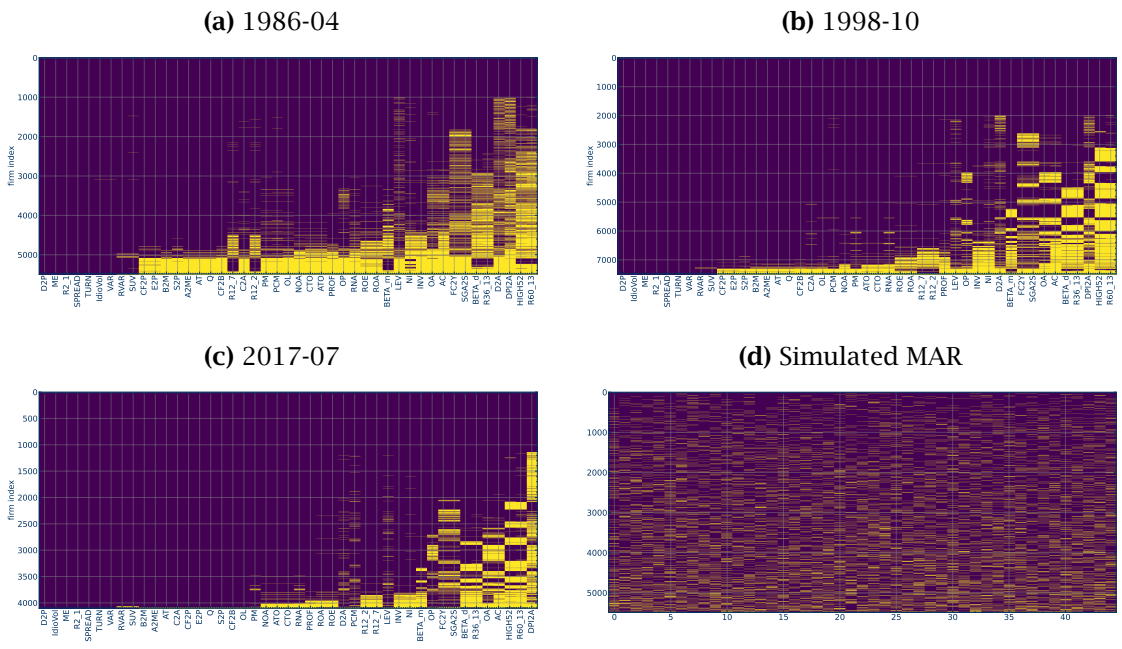
Table 3 summarizes the different estimation approaches. For each estimator we have a local version that only uses information at time t and a global version that uses the full time-series.

3.3. Distribution of Missingness

The fact that characteristics are not missing at random has implications for how to correctly impute missing values. A straightforward attempt would be to use a parametric or non-parametric model to predict characteristics either based on their own past and/or given the contemporaneous realizations of other characteristics. If such a model is estimated by masking characteristics at random, then it would only be appropriate to impute characteristics, which are missing completely at random. However, as we have already documented in the previous sections, characteristics are not missing at random. Therefore, a machine learning application with random masking on the training data, could lead to a bias in imputed values.

The missingess in characteristics is complex, as illustrated in Figure 7. We show the joint distribution of missing patterns on three representative example months. These plots show the missing

Figure 7: Joint Distribution of Missing Patterns



Note: This figure shows the heatmaps of missing data for each stock for three representative example days. Both axis are sorted by the missing percentage, where we first order by firms and then characteristics. Missing data is indicated in yellow. The three representative example months are 1986-04, 1998-10 and 2017-07. For illustration we also include simulated missing-at-random (MAR) data, which we sort in the same way.

entries for each firm, where the characteristics are sorted by their missing percentage. Obviously, the missingness is heterogenous and dependent between characteristics. The dependency is also expected as many characteristics depend on similar CRSP or Compustat variables in their construction, as summarized in Table B.12. For illustration, we also include a plot that shows the simulated pattern for missing at random (MAR). The missing-at-random assumption is clearly violated in the data.

Our imputation method is particularly well suited for this problem as it allows for general missing patterns. We allow missingness to be heterogenous, time-varying, stock-specific and to depend on the latent factor model. These general results follow from the theory provided in Xiong and Pelger (2019), which correspond to our local cross-sectional model (XS). As the generality of the missing pattern in the Xiong and Pelger (2019) approach is of particular importance for our application, we discuss it in more detail.

In our local cross-sectional model (XS), the probability of missingness, $\mathbb{P}\left(W_{i,l}^t = 0\right) =: p_{i,l}^t$ can depend on the specific stock i, the characteristic l and the time t. First, note that our setup allows

for a different factor model at each time t, and hence imposes no assumptions on the temporal structure of $p_{i,l}^t$. This means that the missingness can vary in a completely general way over time, which includes periods of more unobserved data like at the beginning of our sample, block-missing patterns, mixed-frequency observations or missingness because prior values are unobserved. The probability of missingness is also very general in the characteristic dimension and can be different for each characteristic. This allows for characteristic-specific heterogeneity, for example DPI2A has a higher probability to be unobserved than book-to-market ratios. Another case is group-specific heterogeneity, where for example there are less observations when characteristics are updated quarterly or when a group of characteristics relies on the same accounting variable as an input. Last, and most importantly, the probability of missingness can in an extremely general way depend on the features of each stock. More precisely, the probability can be a general, time-varying and characteristic-specific function of any vector of stock specific information $S_i^t \in \mathbb{R}^r$ and the stock-specific factors F_i^t, that is, $p_{i,l}^t = f_{i,t}(F_i^t, S_i^t)$. For example, the characteristics of small stocks or more extreme characteristic realizations are more likely to be unobserved, which we can account for. In this sense, we allow for an endogenous missing pattern.

However, for the purpose of identification, we need to impose some assumptions on the missingness, which cannot be further relaxed. The random variable $W_{i,l}^t$ has to be independent of Λ_l^t and $e_{i,l}^t$. Essentially, the "characteristic covariance" $\tilde{\Sigma}_{i,j}^{\text{XS},t}$ should be asymptotically the same if estimated from the partially observed data or the infeasible complete data. In other words, we can learn from the partially observed data which stocks are similar to each other. This is a reasonable assumption. Overall, our model is extremely general and accounts for all empirical features of missing characteristics.

The results extend to the global models and the B-XS, F-XS and BF-XS models. The estimation step of the loadings in the global models can be formulated as a "local" model with a larger number of stocks by stacking together the different time periods of individual stocks. The factors are the same for the local and global models. Hence, once we show that Λ^t is close to a global Λ, all the results of the local model carry over. The models that combine cross-sectional and time-series observations use the same type of cross-sectional regression weighted by observed values in the second step. These regressions are key for the generality of our results as they do not impose any further assumptions on the missing pattern besides that $\left(\sum_{i=1}^{N_t} W_{i,l}^t X_i^{l,t} X_i^{l,t\top}\right)$ is asymptotically of full rank and that the error in the combined regression is independent of the missing pattern.

We want to emphasize that the complex missing patterns are one of the reasons why it is so challenging to correctly impute missing values. The imputation of missing values is closely related to problems in causal inference as discussed among others in Athey et al. (2021) and Xiong and Pelger (2019). A naive machine learning prediction method is not appropriate for causal inference if treatments are not completely random. The same problem arises with imputation, which needs to account for patterns in the missingness. This is done with our approach.

3.4. Discussion

3.4.1. Look-ahead bias

The choice of imputation method has implications for the follow-up application. Using more data, either in the form of a global model or by incorporating future information, generally improves the quality of the imputation. However, some of the most important use cases of the characteristics data, including out-of-sample asset pricing and investment, need to avoid a look-ahead-bias. This means future information cannot be used in the imputation, as it could make the performance of an investment strategy appear to be better than what it is actually achievable. Blanchet et al. (2022) discuss the tradeoff between look-ahead-bias and the precision of the imputation.

In our empirical study, the model that uses the most information while avoiding any look-ahead-bias is the local Backward (B-XS) model. The model that uses the most information overall, but also "peaks" into the future, is the global Backward-Forward (BF-XS) model. These two benchmark models allow us to study the tradeoff between using more data and using future information. There are other modifications of our models that could avoid a look-ahead-bias, while using more data. Instead of using only the current month for the local B-XS model, we could use a rolling window for a "locally" global version of the B-XS model. However, as we will show in our analysis below, the factor structure of the cross-sectional factor model is very stable over time. Hence, the global XS and B-XS model are very close to a rolling window look-ahead-bias-free version. The more serious look-ahead bias can arise from directly using future information as an input for imputation, that is, in the Forward models.

3.5. Rank normalization vs. raw characteristics

We model rank normalized data, which can easily be mapped back into raw characteristics. In order to obtain a statistical model for characteristics, we need to appropriately normalize them. Fundamentally, this relates to the conceptual question about how we model dependency. Centered

rank normalized characteristics are the natural choice. By using ranks, we deal with the outliers in the raw characteristics, and also achieve stationarity in the cross-section and over time.

There is a simple mapping between the rank quantiles and raw values through the empirical density function of each characteristics. Therefore, after estimating the density functions, the imputed rank quantiles also provide imputed values for the raw characteristics. We will include these results in our empirical study. We will estimate the density function non-parametrically and also parametrically assuming a normal distribution. In both cases, we do not assume that there is a linear dependency between raw characteristics, but only between their relative ranks. As a further robustness result, we also include the results for a factor model which is directly applied to the characteristic space. This requires us to normalize the raw characteristic values by their cross-sectional median and cross-sectional standard deviation after winsorizing the extreme outliers.

We center our ranks at zero, i.e., we report characteristic quantiles between $[-0.5, 0.5]$, which is without loss of generality. Hence, the cross-sectional median corresponds to the value zero. Using uncentered rank quantiles between $[0, 1]$ simply adds an additional latent cross-sectional factor, that captures the median and is similar to a "market" or "level" factor.

3.6. Evaluation metrics

We evaluate the different models based on their RMSE (root mean squared errors). The aggregated RMSE for the model implied characteristic $\hat{C}_{i,t,l}$ is averaged over all stocks, time-periods and characteristics:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

We also consider the RMSE for each characteristic separately

$$\text{RMSE}_l = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2},$$

as well as over time

$$\text{RMSE}_t = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}.$$

All our results are reported in-sample and out-of-sample. The in-sample results evaluate how well a low dimensional model can approximate the characteristics. As these results can be biased

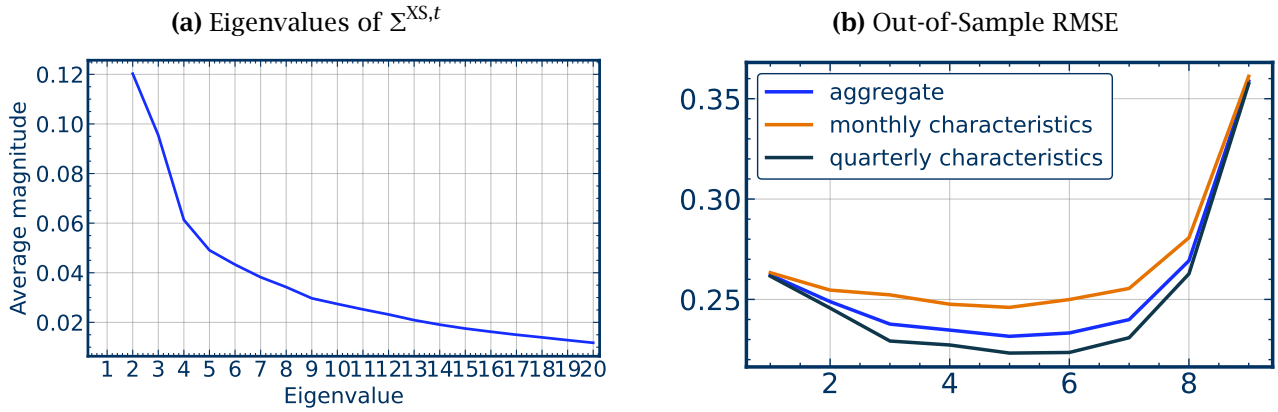
upwards due to overfitting, we also need to conduct an out-of-sample analysis (OOS). The OOS analysis masks observed entries before we estimate the model on the remaining data. The OOS RMSE compares the masked observed entries with the model implied values. We consider three different missing patterns for the out-of-sample analysis. The first case is OOS missing-at-random, where we mask 10% of the observed characteristics completely randomly. The second case is OOS block-missing, where we mask 10% of characteristics in blocks of 1 year. The second case accounts for the empirically observed temporal dependency in missing patterns. It is important to include this case, as for very persistent characteristics the last observed value can provide a very good prediction, but empirically it is often not available. Third case uses the logistic regression model from Table 2 with all covariates and fixed effects to mask entries. The propensity of the logistic regression captures important features of missing patterns. In particular, the probability of missingness is heterogeneous, appears in blocks over time and in the cross-section and depends on the realization of observed characteristics.

As we work with rank-quantiles, the characteristics are normalized and the RMSE provides an interpretable measure of the deviation from the true value. In addition, we report the R^2 that measures the explained variation relative to the cross-sectional median imputation.

4. Factor Structure in Characteristics

Empirically, firm characteristics are well described by a parsimonious factor model. Before conducting an extensive comparison between different imputation methods, we study the properties of a cross-sectional latent factor model. We discuss the choice of the number of factors, their economic interpretation and variation over time. Estimating a cross-sectional latent factor model requires that at least some characteristics are observed for each stock. Moving forward, our analysis focusses on the data set of all stocks that have at least ten characteristics observed at each point in time. As shown in Figure 4(e) this requirement imposes almost no restrictions, and on average 97% of all stocks have ≤ 35 of the 45 characteristics missing after 1977. The second restriction is that we focus on the data after 1977, which is more homogenous and more widely used in empirical applications. We have confirmed that our general results are robust to these two choices.

Figure 8: Number of Latent Factors



Note: In this figure we determine the number of latent factors. The left subplot shows the magnitude of eigenvalues of the characteristic covariance matrix relative to the sum of all eigenvalues averaged over time. The right subplot displays the out-of-sample imputation RMSE as function of the number of cross-sectional factors using the block-missing masking.

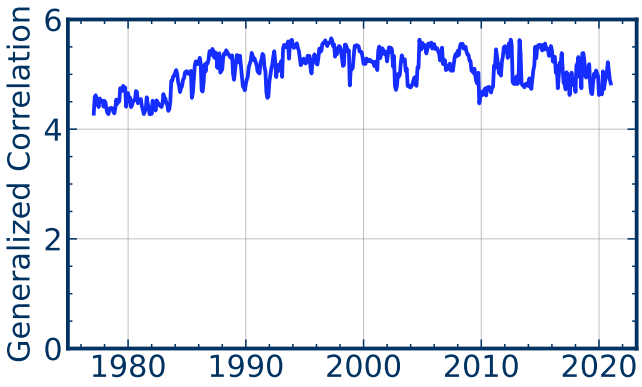
4.1. Number of factors

The number of systematic cross-sectional characteristic factors is directly linked to the eigenvalues of the characteristic "covariance" matrix $\tilde{\Sigma}_{l,p}^{XS,t}$. Figure 8(a) plots the magnitude of eigenvalues of $\tilde{\Sigma}_{l,p}^{XS,t}$ relative to the sum of all eigenvalues averaged over time. These eigenvalues can be interpreted as the amount of variation explained by a small number of global factors. The first four factors explain the most variation in the data. It seems that the factors five to nine also contribute a non-negligible amount. Overall, we find strong evidence for a factor structure.

We select the number of factors by minimizing the out-of-sample RMSE. Figure 8(b) shows the OOS RMSE for the block-missing masking as a function of the number of factors. We consider a global cross-sectional model (XS) and report the RMSE for monthly, quarterly and the all characteristics. The OOS RMSE of quarterly updated characteristics is minimized for six latent factors, while monthly updated characteristics are well described by a five-factor model. The aggregated RMSE reflects these two findings and six latent factors are very close to the optimum. We select six factors as our parsimonious baseline model.

The results of the cross-sectional model (XS) carry over to the models that also include time-series information. Table B.3 in the Appendix shows the OOS RMSE for block-missing patterns for different number of factors for the local B-XS, global B-XS and local XS. The optimal number depends on the type of characteristic and method, but seems to be between six and eight factors. The benefit

Figure 9: Generalized Correlation of Global and Local Factor Weights



Note: This figure shows the time series of the generalized correlation of the constant global Λ with the time-varying local Λ^t estimated each month. We consider a six-factor model.

of including more than six factors seems to be only marginal and hence we opt for the parsimonious six-factor model.

4.2. Local vs. global factors

The loading structure of the cross-sectional factor model is relatively stable over time. A global factor model assumes a constant loading matrix Λ, while a local factor model allows for time-varying loadings Λ^t. We show that the loading structure is relatively stable over time and hence justifies the use of constant loadings. Figure 9 plots the generalized correlations between the global loadings Λ and local loadings Λ^t for the first six factors over time. A generalized correlation equal to six would imply that the two loading matrices span the same space. While there is some variation, the generalized correlation is close to the maximum. We conclude that it is meaningful to analyze the composition of the global factors.

4.3. Structure of factors

The characteristic factors have a meaningful economic interpretation. The loadings Λ can be interpreted as weights to construct the characteristic factors. We focus on the global model as it is described by only one set of weights, which are closely related to the local weights. Figure C.3 in the Appendix plots the composition of the six latent factors, which are described by Λ. The characteristics are grouped together by categories.

Some of the latent factors can be linked to characteristic categories. The second factor seems to load heavily on value characteristics. The third factor has large weights for profitability characteris-

tics. The fourth factor seems to be a trading friction factor. The sixth factor has positive positions in past returns and investment and negative positions in the other categories. Some of the structure seems also to be related to the updating frequency of characteristics and their volatility. Figure C.4 shows the composition based on the updating frequency. Factor one has large weights on monthly updated characteristics, and in particular on those that have a high volatility. In this sense we can label it a high volatility characteristic factor. On the other hand, factor five loads more on slowly moving characteristics.

4.4. Rank normalization vs. raw characteristics

Our main analysis reports the results for rank quantiles, but the results carry over to raw characteristics. Table B.4 in the Appendix shows the out-of-sample imputation RMSE in the original characteristic space without transforming characteristics into ranks. We consider OOS block-missing for different number of cross-sectional factors. The raw characteristics are normalized by their cross-sectional mean and variance.[8] The RMSE are further normalized by the RMSE of a simple median imputation. The first model is our baseline factor model estimated on ranks and transformed back into the characteristic space with the empirically estimated density function of each characteristic. We estimate the density function with the machine learning method k-nearest neighbor. The second and third model estimates the factor model directly on the characteristics. In the fourth and fifth case, we estimate the factor model in the kernel transformed space with a Gaussian kernel and revert it back to the raw characteristics.

We observe that a factor model estimated on rank quantiles and inverted back to raw characteristics outperforms a factor model directly applied to raw characteristics. If we use a normal distribution instead of a non-parametric density function to invert the model into the raw characteristic space, we perform slightly worse, but still substantially better than directly estimating a factor model in the raw characteristic space. A local model with locally estimated normal density function can perform better than the empirical non-parametric density. We conclude that the rank quantile space is appropriate for the latent factor model and provides better results than a factor model in the raw characteristic space.

[8] Because of the outliers we need to winsorize the data. In more detail, we first estimate the cross-sectional mean and standard deviation of each raw characteristics for each day. Then, we winsorize the values that deviate more than five standard deviations from the cross-sectional mean. After winsorizing, we reestimate the mean and standard deviation, which we use to finalize the normalization of the raw characteristics.

5. Imputation

5.1. Aggregate comparison between methods

In an extensive comparison study we compare the quality of different imputation approaches. We include the different variations of our model framework and the most widely used conventional ways to deal with missing data. The baseline model without look-ahead bias (that is, without using future information) is the local B-XS. The baseline model using as much information as possible is the global BF-XS. All cross-sectional models use six latent factors based on the analysis in the previous section. We consider the global and local versions of our models and different combinations of time-series information, that is backward, forward or none. Another special case would be to drop the cross-sectional model and only run an AR(1) model. The popular conventional approaches encompasses using only the previous value, a cross-sectional median or the industry-specific median for imputation. In total, have the following 11 models: global BF-XS, global B-XS, global F-XS, global XS, global B, local B-XS, local XS, local B, previous value (PV), XS median and industry median.

The main results are summarized in Table 4, which shows the imputation errors for these different imputation methods. We report the in-sample, OOS missing-at-random, OOS block-missing and OOS logit results for all characteristics and separated by their updating frequency. The first striking observation is that cross-sectional median or industry median results in roughly twice as large imputation errors compared to our baseline models local B-XS and global BF-XS. These results are robust to the updating frequency and the in- or out-of-sample analysis. We conclude that the current standard of ignoring the time-series and cross-sectional dependency is strongly suboptimal. The local and global versions of our model are relatively close, but the global version seems to lead to slightly smaller imputation errors. We will revisit this aspect in more detail in Section 5.3.

Our baseline models are the best within their categories. Within the global models the global BF-XS dominates the alternative approaches. This is not surprising as using future information should be beneficial. However, the difference between the global BF-XS and global B-XS for the out-of-sample data is much smaller compared to using only a cross-sectional model (XS). This is expected as very persistent characteristics should be well predicted by their past observations. Using simply the previous value performs worse than using an AR(1) time-series model as characteristics are usually not stale, but only autocorrelated. A simple backward time-series model, labeled as B, performs surprisingly well. However these results depend crucially on the availability of previous observations. Around one third of values are missing at the beginning and cannot be imputed with

Table 4: Imputation Error for Different Imputation Methods

	In-Sample			OOS MAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
	Imputation RMSE											
global BF-XS	**0.09**	**0.08**	**0.12**	**0.13**	**0.13**	**0.13**	**0.10**	**0.08**	**0.13**	**0.10**	**0.09**	**0.13**
global F-XS	0.09	0.06	0.13	0.15	0.15	0.14	0.10	0.06	0.14	0.18	0.16	0.23
global B-XS	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.13	0.12	0.15
global XS	0.19	0.18	0.21	0.22	0.21	0.24	0.23	0.22	0.24	0.25	0.24	0.27
global B	0.15	0.15	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.14	0.13	0.16
local B-XS	**0.14**	**0.14**	**0.13**	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.13**	**0.12**	**0.15**
local XS	0.21	0.20	0.21	0.23	0.22	0.24	0.23	0.23	0.24	0.25	0.24	0.27
local B	0.15	0.15	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.14	0.13	0.16
prev	0.17	0.16	0.18	0.17	0.16	0.18	0.17	0.16	0.19	0.15	0.14	0.19
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.31
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.30	0.31
	Explained Variation R^2											
global BF-XS	**0.85**	**0.94**	**0.80**	**0.80**	**0.83**	**0.79**	**0.83**	**0.94**	**0.77**	**0.93**	**0.94**	**0.55**
global F-XS	0.85	0.98	0.77	0.75	0.77	0.74	0.81	0.97	0.71	0.49	0.74	0.06
global B-XS	0.78	0.81	0.77	0.76	0.79	0.74	0.75	0.81	0.71	0.87	0.87	0.48
global XS	0.57	0.61	0.54	0.42	0.47	0.39	0.38	0.43	0.36	0.23	0.35	0.11
global B	0.76	0.79	0.74	0.75	0.78	0.73	0.74	0.79	0.71	0.85	0.86	0.45
local B-XS	**0.79**	**0.82**	**0.78**	**0.77**	**0.80**	**0.75**	**0.76**	**0.81**	**0.73**	**0.87**	**0.87**	**0.49**
local XS	0.50	0.52	0.50	0.40	0.43	0.38	0.37	0.38	0.35	0.25	0.34	0.11
local B	0.76	0.80	0.74	0.75	0.78	0.73	0.74	0.80	0.71	0.85	0.86	0.45
prev	0.66	0.76	0.60	0.64	0.75	0.58	0.63	0.76	0.56	0.84	0.85	0.01
XS-median	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ind-median	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00

Note: This table shows imputation RMSE and R^2 by imputation method averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data either missing at random or missing in time-series blocks for 12 consecutive months. The logit masking is based on the logistic regression model with all covariates and fixed effects as estimated in Table 2. The R^2 is the explained variation relative to a cross-sectional median imputation.

the backward time-series model, but require contemporaneous cross-sectional information. In the case of block-missing patterns, which is empirically more relevant, the global and local B-XS model outperforms the global and local B. We conclude that using both information sets, the time-series and cross-sectional dependency, seems to be beneficial and that our baseline models, local B-XS and global BF-XS, are the best imputation methods.

The general ordering of imputation methods holds among all masking mechanisms. The outperformance of the baseline models, local B-XS and global BF-XS, is even more pronounced for the logit masking. In this case, most masking occurs for quarterly characteristics. When monthly characteristics are masked, it is likely that a very large number of characteristics is masked simultaneously and/or the block of missing data is very long, which provides a challenge for all imputation methods.

The in-sample results can be interpreted as an evaluation of the parsimonious characteristic model, while the out-of-sample results also test how well the parsimonious can be estimated from the partially observed data. The fact that the in-sample and out-of-sample results are very close is evidence that our characteristic models do not overfit, but provide a good description for characteristics.

The lower part of Table 4 reports the R^2 which measures the explained variation relative to a cross-sectional median imputation. It clarifies how substantial the improvements are for our baseline models. The global BF-XS can achieve an out-of-sample R^2 of 0.93 for logit masking, while the local B-XS achieves an impressive 0.87. This means in terms of explained variation these methods achieve an out-of-sample improvement of over (or close to) 90%. The median imputation has by definition an R^2 of zero.

Many applications use only the subset of largest or smallest characteristic values. One prominent example are portfolio sorting strategies based on the extreme quantiles of characteristics. These applications depend on a precise imputation of the extreme characteristic quantiles, but are less affected by the imputation quality in the center of the distribution. The outperformance of our baseline models relative to naive imputation is even more pronounced for these values.

Table 5 reports the RMSE for the masked characteristic values which are in the first or fifth characteristic quintile. By construction the median imputation performs particularly badly. The local B-XS has less than half of the median RMSE while the global BF-XS has around one third. This confirms that our baseline models provide the preferred imputed values even for extreme realizations.

In order to provide some intuition, we illustrate the model implied and imputed time-series for representative examples. Figure 10 shows characteristic time-series for Microsoft and Hasbro, two representative companies in different industries and hence with different fundamentals. We show their characteristic time-series for three characteristics with different levels of persistence. The most persistent is market capitalization. Tobin's Q has a medium level of persistence, while the local variance is a fast fluctuating characteristic. These three examples are relatively representative as they capture stylized features of other characteristics. We show the model implied values in-sample and also the imputation results for out-of-sample missing blocks of 12 months.

The most obvious observation is that the median value creates very large errors in observed and imputed values. Importantly, if we would use the median imputed values for the missing blocks, we would also distort the time-series of the characteristics. For example the centered rank quantile for

Table 5: Imputation Error for Extreme Characteristic Quintiles

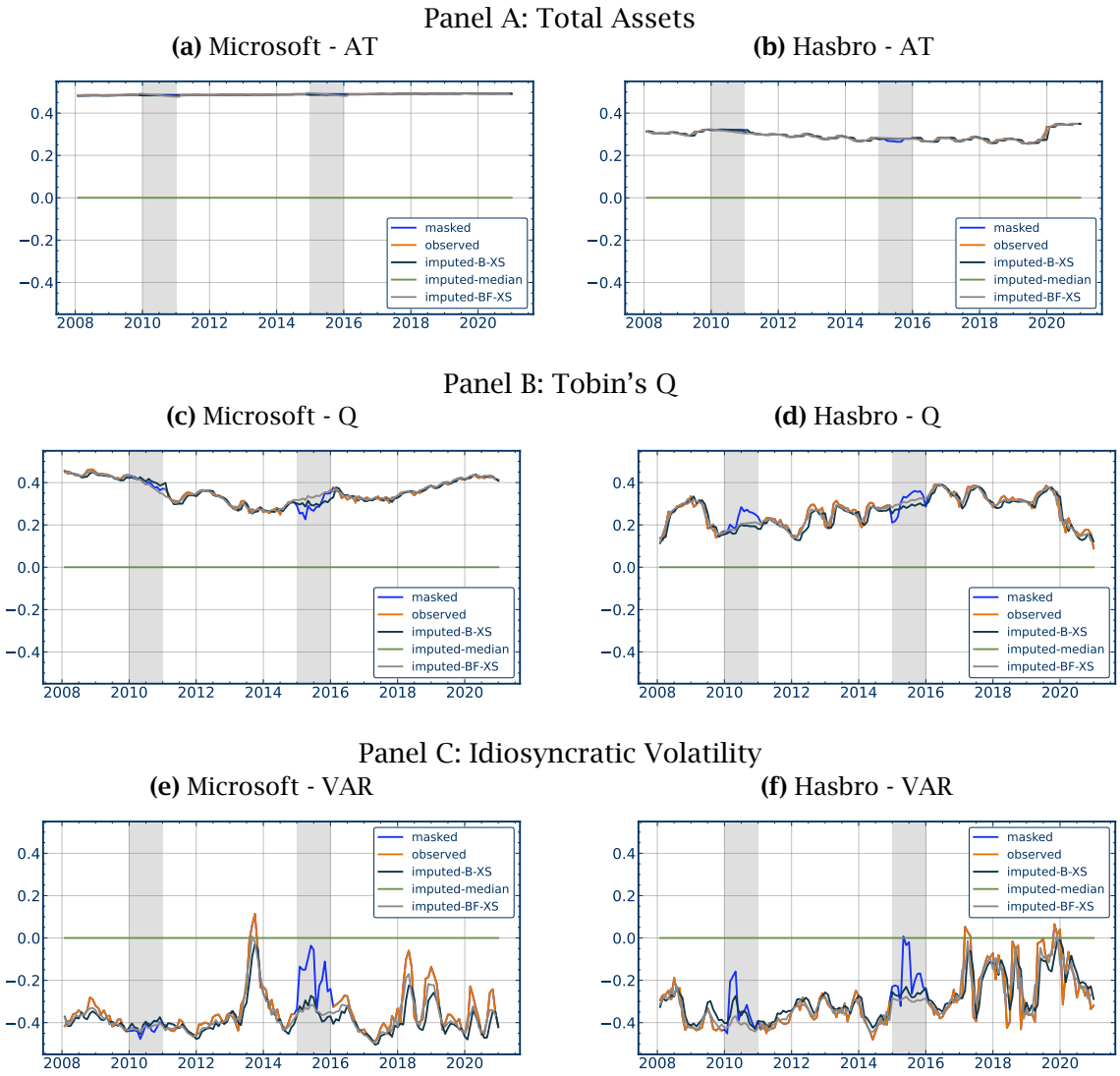
	In-Sample			OOS MAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
First characteristic quintile												
global BF-XS	**0.12**	**0.09**	**0.15**	**0.16**	**0.16**	**0.16**	**0.12**	**0.10**	**0.16**	**0.13**	**0.11**	**0.17**
global F-XS	0.11	0.06	0.16	0.18	0.18	0.17	0.12	0.08	0.17	0.21	0.18	0.26
global B-XS	0.17	0.17	0.16	0.17	0.18	0.17	0.17	0.17	0.18	0.16	0.15	0.19
global XS	0.24	0.24	0.26	0.28	0.27	0.29	0.29	0.28	0.30	0.32	0.31	0.35
global B	0.18	0.18	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.16	0.19
local B-XS	**0.17**	**0.17**	**0.16**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	**0.16**	**0.15**	**0.18**
local XS	0.26	0.26	0.27	0.29	0.28	0.30	0.30	0.30	0.30	0.32	0.31	0.35
local B	0.18	0.18	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.16	0.15	0.19
prev	0.19	0.19	0.20	0.19	0.19	0.20	0.20	0.19	0.21	0.18	0.16	0.22
XS-median	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
ind-median	0.40	0.40	0.41	0.40	0.40	0.41	0.40	0.40	0.41	0.41	0.41	0.41
Fifth characteristic quintile												
global BF-XS	**0.12**	**0.10**	**0.16**	**0.17**	**0.16**	**0.17**	**0.13**	**0.10**	**0.17**	**0.13**	**0.11**	**0.18**
global F-XS	0.11	0.06	0.17	0.18	0.18	0.18	0.12	0.07	0.19	0.22	0.19	0.30
global B-XS	0.17	0.17	0.17	0.18	0.17	0.18	0.18	0.17	0.18	0.17	0.15	0.20
global XS	0.25	0.23	0.27	0.28	0.26	0.31	0.29	0.28	0.31	0.32	0.30	0.36
global B	0.18	0.18	0.18	0.19	0.18	0.19	0.19	0.18	0.19	0.17	0.16	0.21
local B-XS	**0.17**	**0.17**	**0.17**	**0.18**	**0.17**	**0.18**	**0.18**	**0.17**	**0.18**	**0.16**	**0.15**	**0.20**
local XS	0.27	0.26	0.28	0.29	0.28	0.31	0.30	0.29	0.31	0.33	0.31	0.36
local B	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.19	0.17	0.16	0.21
prev	0.20	0.19	0.21	0.20	0.19	0.22	0.20	0.19	0.22	0.19	0.17	0.24
XS-median	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
ind-median	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41

Note: This table shows imputation RMSE by imputation method for different types of missingness. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data either missing at random, missing in time-series blocks for 12 consecutive months, or with the logistic regression model. We report the RMSE for the subset of masked values which are in the first or fifth characteristic quintile.

the total assets (AT) of Microsoft would jump from about 0.5 to 0 and back to 0.5. In contrast, the imputed values with our methods reflect substantially better the level and dynamics of characteristics. Second, our two baseline models are very exact on the in-sample data. Obviously, the imputation is more challenging on the out-of-sample data. Third, our models reflect dynamic changes in the out-of-sample data, which are captured by the cross-sectional factor component. As we will see in Section 5.3, this cross-sectional component is more relevant for fast changing characteristics like the variance. Last but not least, the BF-XS seems to "connect" the two end points of the missing data, while the B-XS model is for obvious reasons "anchored" at the starting point of the missing block.

The aggregated comparison results are robust over time and with respect to the market capitalization of the stocks. Figures C.13, C.15 and C.17 in the Appendix show the RMSE for each month.

Figure 10: Illustrative Model-Implied and Imputed Time-Series



Note: This figure shows illustrative realized and model-implied characteristic time-series for Microsoft and Hasbro. We plot the realized characteristic rank over time, and the model implied values with the B-XS, BF-XS and median model. The gray shaded areas indicate missing blocks of one-year which are not part of the estimation, and hence serve as out-of-sample evaluation. We consider Total Assets, Tobin's Q and variance, which are three representative characteristics of decreasing persistence.

The relative ordering of the different methods is very stable over time. Table B.8 in the Appendix reports the RMSE for different size deciles. While the errors are larger in magnitude among smaller stocks, the relative comparison between the models stays the same. Importantly, even the largest size decile accounts for a substantial part of the imputation errors, and hence the results are not driven by fitting only small cap stocks.[9]

[9]Tables B.8 and B.9 show that the results are also robust with respect to filters based on share prices and to excluding

5.2. Imputation results for different types of missingness

As a next step we want to understand how the imputation results are affected by the type of missingness. Hence we show all the results of the previous subsection for data missing at the start, the middle and the end of the sample. Table 6 collects the in-sample and out-of-sample RMSE results. Note that the type of missingness restricts which models can be used. For example, when observations are missing at the beginning of the sample, we can obviously not use any of the models that require prior observations. Similarly, for observations at the end, the forward models are excluded. Only missingness in the middle of the sample allows us to use all models. Our aggregated results in the previous subsection only reported the errors for observations where a model was applicable. Here we separate those effects.

The best model for missing observations at the beginning of the sample are the global F-XS when using all possible information and the local XS when avoiding a look-ahead bias. These are the special cases of our baseline models that exclude the prior information. Importantly, the difference to the median imputation is even more pronounced than for the aggregated results. Therefore, we recommend to use these two baseline models for imputing the missing values at the start.

The best model for missing observations in the middle are the global BF-XS for full observations and the local B-XS among the look-ahead-bias free models. The magnitude of the RMSE and relative ordering is very close to the aggregate results in Table 4. Overall our baseline models dominate the other approaches. Last but not least, we show that the global B-XS and local B-XS are the best model for missingness at the end of the sample. While the relative ordering of methods stays the same, the magnitude of errors seems to be higher.

We conclude that the best model avoiding future information is the local B-XS, and, if data is missing at the beginning, we replace it by the local XS. The best global model is the global BF-XS, which we replace by the global F-XS for missingness at the beginning and the global B-XS for missingness at the end.

5.3. Which information matters?

Which characteristics are hard to predict and what information is the most useful? In order to answer these questions we compare the imputation errors for each characteristic. In the main text we focus on the out-of-sample results with block-missing pattern, while the Appendix collects the

financial institutions.

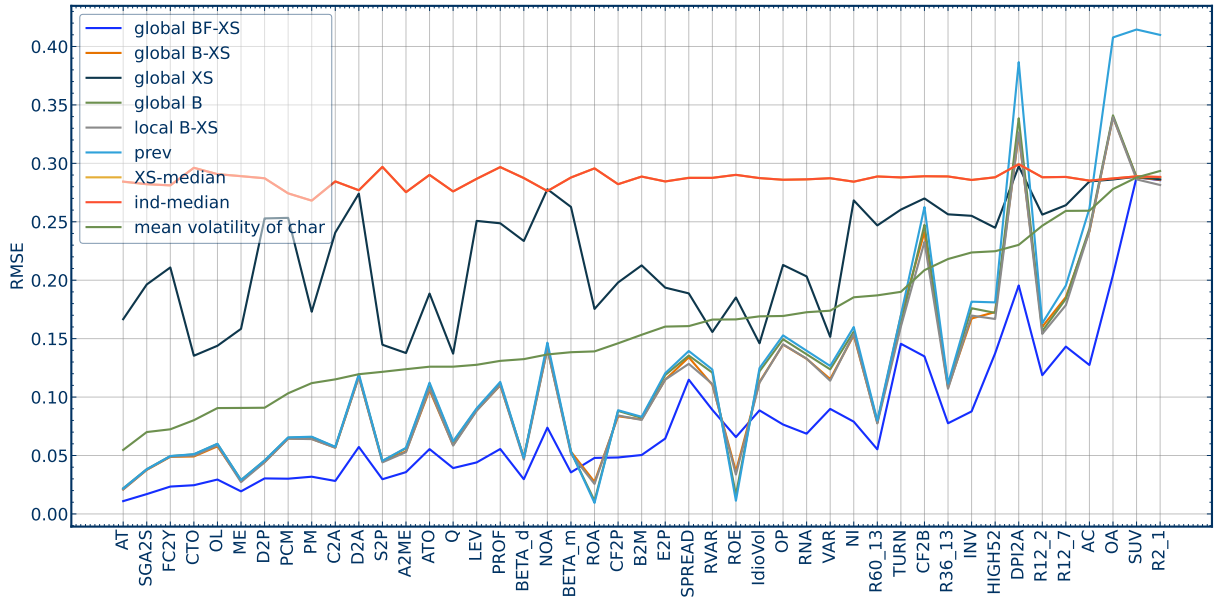
Table 6: Imputation Error for Types of Missingness

	In-Sample			OOS MAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
Start of the sample												
global BF-XS	-	-	-	-	-	-	-	-	-	-	-	-
global F-XS	**0.10**	**0.05**	**0.16**	**0.17**	**0.17**	**0.18**	**0.12**	**0.07**	**0.17**	**0.22**	**0.20**	**0.26**
global B-XS	-	-	-	-	-	-	-	-	-	-	-	-
global XS	0.22	0.21	0.24	0.26	0.24	0.28	0.27	0.26	0.28	0.29	0.29	0.29
global B	-	-	-	-	-	-	-	-	-	-	-	-
local B-XS	-	-	-	-	-	-	-	-	-	-	-	-
local XS	**0.24**	**0.23**	**0.25**	**0.26**	**0.25**	**0.28**	**0.27**	**0.26**	**0.27**	**0.29**	**0.29**	**0.29**
local B	-	-	-	-	-	-	-	-	-	-	-	-
prev	-	-	-	-	-	-	-	-	-	-	-	-
XS-median	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.32	0.32	0.31
ind-median	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.32	0.32	0.31
Middle of the sample												
global BF-XS	**0.09**	**0.08**	**0.12**	**0.13**	**0.13**	**0.13**	**0.10**	**0.08**	**0.13**	**0.10**	**0.09**	**0.13**
global F-XS	0.09	0.06	0.13	0.14	0.15	0.14	0.1	0.06	0.14	0.13	0.12	0.15
global B-XS	0.13	0.14	0.13	0.14	0.14	0.14	0.14	0.14	0.15	0.13	0.12	0.15
global XS	0.19	0.18	0.21	0.22	0.21	0.23	0.22	0.21	0.24	0.22	0.21	0.24
global B	0.14	0.15	0.14	0.15	0.15	0.15	0.15	0.143	0.15	0.14	0.13	0.16
local B-XS	**0.13**	**0.14**	**0.13**	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.14**	**0.13**	**0.12**	**0.15**
local XS	0.20	0.12	0.22	0.22	0.22	0.23	0.23	0.22	0.24	0.23	0.22	0.24
local B	0.14	0.15	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.14	0.13	0.16
prev	0.16	0.16	0.18	0.17	0.16	0.18	0.17	0.16	0.18	0.15	0.14	0.19
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
End of the sample												
global BF-XS	-	-	-	-	-	-	-	-	-	-	-	-
global F-XS	-	-	-	-	-	-	-	-	-	-	-	-
global B-XS	**0.16**	**0.15**	**0.17**	**0.18**	**0.18**	**0.17**	**0.18**	**0.18**	**0.17**	**0.12**	**0.12**	**0.14**
global XS	0.23	0.23	0.24	0.26	0.25	0.267	0.27	0.27	0.27	0.25	0.25	0.27
global B	0.19	0.19	0.19	0.19	0.19	0.18	0.19	0.19	0.18	0.13	0.13	0.15
local B-XS	**0.16**	**0.15**	**0.17**	**0.18**	**0.18**	**0.17**	**0.18**	**0.18**	**0.17**	**0.12**	**0.12**	**0.14**
local XS	0.25	0.25	0.25	0.26	0.26	0.27	0.27	0.28	0.27	0.26	0.26	0.27
local B	0.19	0.19	0.19	0.18	0.19	0.18	0.19	0.19	0.18	0.13	0.13	0.15
prev	0.21	0.20	0.22	0.20	0.20	0.22	0.21	0.19	0.22	0.14	0.13	0.17
XS-median	0.35	0.37	0.34	0.34	0.33	0.33	0.35	0.37	0.33	0.32	0.32	0.33
ind-median	0.35	0.37	0.34	0.34	0.33	0.33	0.35	0.37	0.33	0.32	0.32	0.33

Note: This table shows imputation RMSE by imputation method for different types of missingness. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data either missing at random or missing in time-series blocks for 12 consecutive months. The out-of-sample logit masking is based on the logistic regression model with all covariates and fixed effects as estimated in Table 2.

in-sample and out-of-sample missing-at-random and logit masking results. Figure 11 plots the out-of-sample block-missing imputation errors for individual characteristics sorted in ascending order based on their time-series volatility. Characteristics on the right, for example short-term momentum, fluctuate the most and hence might be harder to predict from the time-series, while the characteris-

Figure 11: Imputation Error For Individual Characteristics



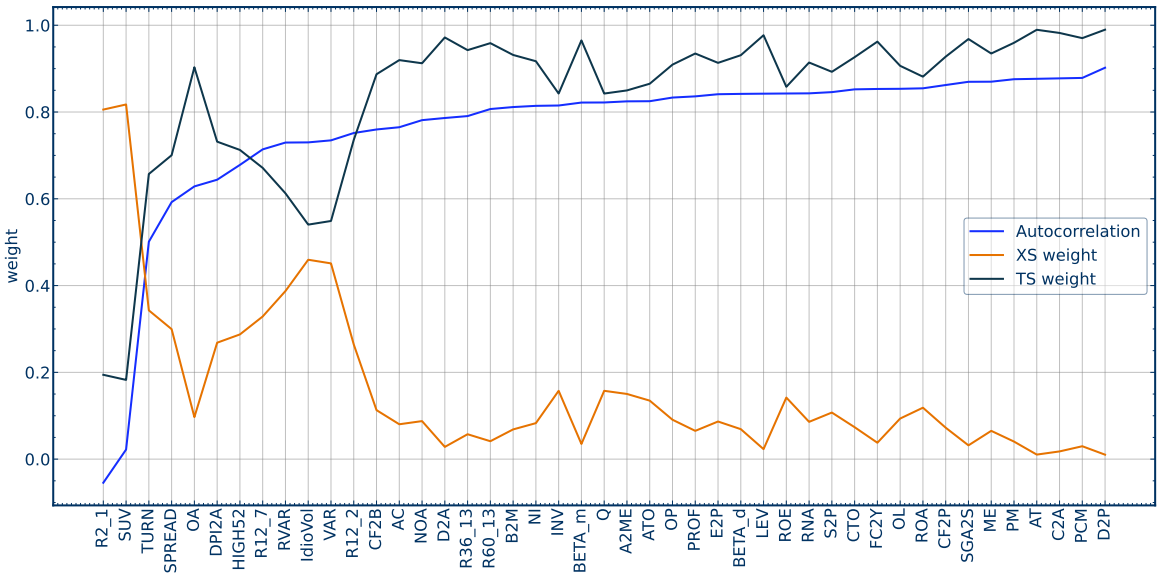
Note: This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data missing in time-series blocks for 12 consecutive months.

tics on the left, for example total assets, are more persistent.

The median or industry median are in almost all cases the worst possible models. The pure cross-sectional model, which includes the median as a special case for a zero factor model, strictly dominates the median imputation. The imputation of more volatile characteristics seems to benefit more from cross-sectional information. On the other hand, the more persistent characteristics seem to rely more on time-series information. A pure time-series or pure cross-sectional model is not uniformly better, and in almost all cases a combination of both information leads to superior results. The global BF-XS model has the smallest errors except for return in net-operating assets (ROA) and return on equity (ROE). The local B-XS is for almost all characteristics the best local model. The results are comparable for the logit masking as shown in Figure C.7 in the Appendix.

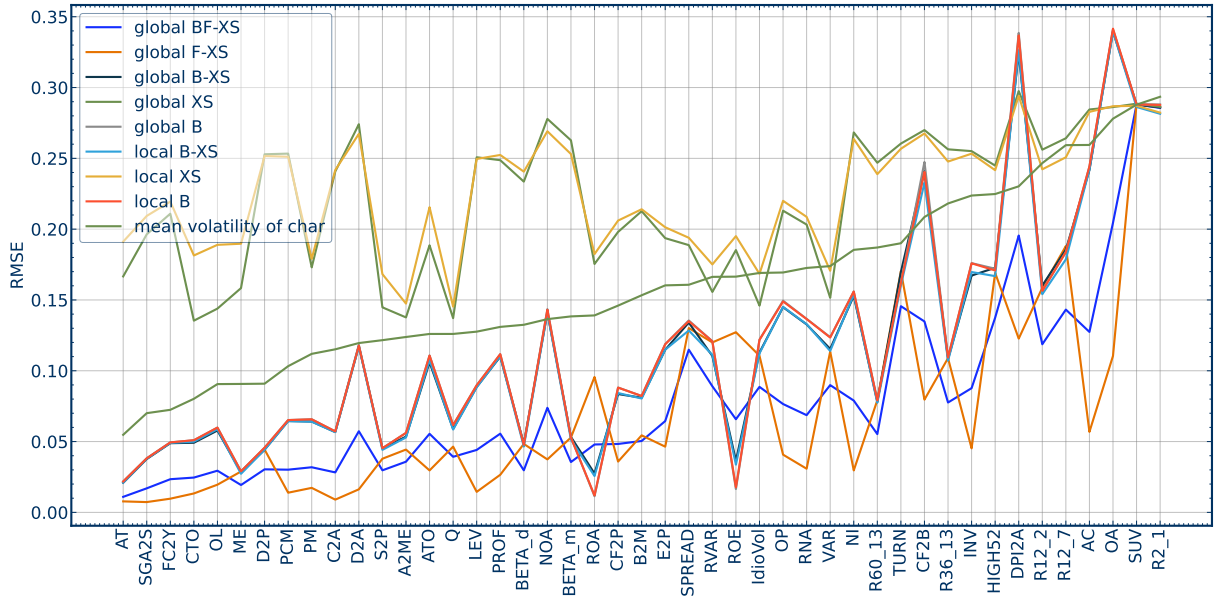
The results are qualitatively similar for missing-at-random as shown in Figure C.5 in the Appendix. Overall, the benefit of cross-sectional information for more persistent information seems to shrink. This is expected, as there are only very few missing points in a row and hence the last observed values can be very informative. However, the relative ranking stays the same. The results are comparable for the in-sample analysis.

Figure 12: Information used for Imputation



Note: This figure shows the regression coefficients on the cross-sectional factor model and the time-series information. The XS weight denotes the sum of absolute values of the coefficients on the cross-sectional factor model. The characteristics are sorted in ascending order based on their autocorrelations.

Figure 13: Global and Local Imputation For Individual Characteristics



Note: This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data missing in time-series blocks for 12 consecutive months.

In order to assess the relative importance of the time-series and cross-sectional information, we compare the relative weights in the regressions of the B-XS and BF-XS models. Figure 12 shows the regression coefficients on the cross-sectional factor model and the time-series information for the B-XS model. The XS weight denotes the sum of absolute values of the coefficients on the cross-sectional factor model. The characteristics are sorted in ascending order based on their autocorrelation. As expected, the time-series weight follows closely the autocorrelation. This means that the most persistent characteristics use primarily time-series information for the imputation. In contrast, highly volatile and only weakly serially correlated characteristics put larger weights on the cross-sectional factor model. For example, unexplained volume (SUV) puts 80% of its weight on cross-sectional factors. Figure C.8 shows that the BF-XS exhibits exactly the same pattern. Interestingly, the weights on past and future information are essentially symmetric. It seems that the weights on past and future information in the BF-XS add up to the time-series weights in the B-XS model, that is, the relative weights on the overall time-series is the same in both models. We provide the detailed weights on the individual factors in Tables B.6 and B.7 in the Appendix, which are in line with our interpretation of the cross-sectional factors.

Last but not least, we compare the global and local models in more detail. Figures 13 and C.6 show a comparison of imputation RMSE for local and global method across individual characteristics. As before, the characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. As expected by the aggregate statistics, the global models are slightly better than their local counterpart. However, highly volatile characteristics can benefit from local models. This is for example visible for the pure cross-sectional models. This implies that the models are relatively stable over time for most characteristics, but there can be some time variation among the more volatile characteristics.

6. Asset Pricing

6.1. Selection bias - Firms with missing characteristics are different

Firm characteristics are the most widespread conditioning drivers of expected returns in asset pricing. Missing financial data can have a profound impact on asset pricing, depending on the application and extent of the problem.

Missing values in firm characteristics can have two fundamental effects on asset pricing. The first effect is the selection bias studied in this section. Asset pricing and investment results depends on

which stock are included. Firms with missing characteristics are different from those with observed entries. Hence, using only the subsample of stocks with fully observed characteristics leads to a selection bias in asset pricing metrics. This is reflected in the substantially better out-of-sample investment performance of including all stocks instead of the non-representative subsample with fully observed data. The second effect is the imputation bias studied in Section 6.2. Asset pricing pricing results depend on the imputation method. Biased imputation methods like the median imputation lead to uniformly and substantially larger errors in asset pricing metrics compared to our more precise imputation approach. In all the empirical applications, in order to ensure that the characteristic information is available to an investor in real time, we use the values of observed or imputed characteristics lagged by six months.[10]

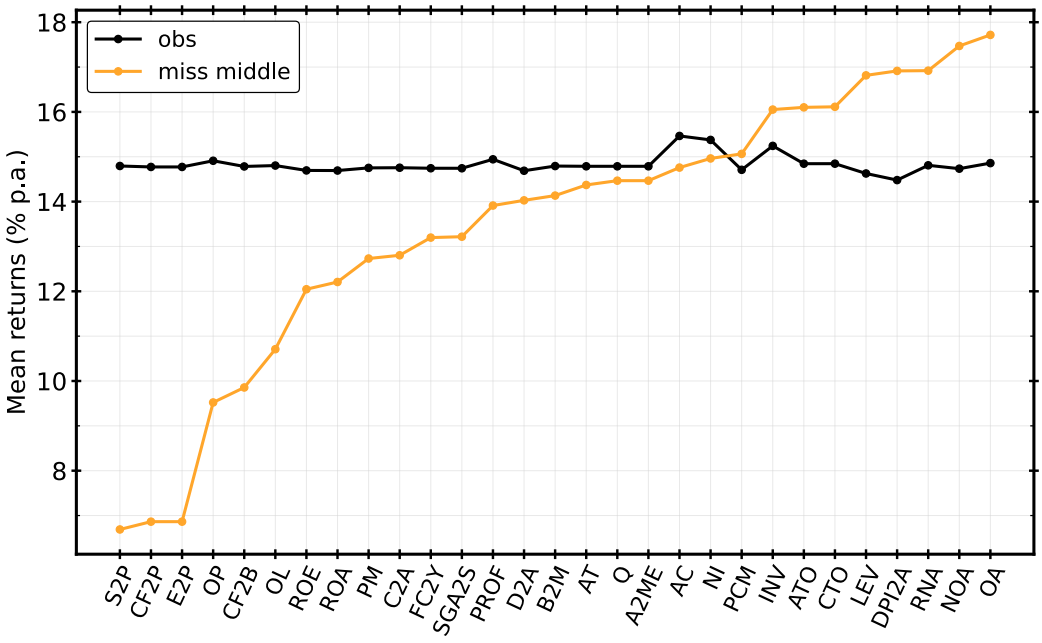
6.1.1. Market strategy with observables

We begin this section by documenting a very simple empirical result: even the average return on a market-style long-only portfolio of stocks depends on whether the portfolio is constructed of stocks that have particular characteristics observed. In other words, even simply having – or not – observable values for popular firm characteristics like book-to-market ratio, or Tobin's Q, on its own have an impact of asset returns, separate from its value.

Figure 14 shows average returns of stocks with observed or missing characteristics. In each month, we compute mean returns of all stocks with observed data of a particular characteristics as well as mean returns of stocks for which the characteristic is missing in the middle of the stock sample. The figure plots means across time for quarterly characteristics. The presence of many firm-specific fundamentals seems to have an impact on asset returns – in part, due to the selection of firms with certain characteristics into the observable set. For some characteristics, stocks with missing data have lower returns than stocks with observed data, while the reverse is true for other characteristics. Stocks with missing price-ratios have lower returns than stocks with observed data, in particular for the sales-to-price, cash-flow-to-price, and earnings-to-price ratios. This pattern is reversed for investment-related characteristics. The differences in mean returns are economically large ranging from 8.10% (p.a.) for the sales-to-price ratio to -2.86% for operating accruals. Clearly, estimating expected returns on only partially observed data can suffer from a selection bias.

[10]Our results are qualitatively the same if we use a lag of three month or longer lags. Note, our focus is not on the optimal lag horizon for investments, but to clarify that it has an impact how we deal with missing data. Investors could also use different lag horizons for different characteristics, yet even in that setting our results largely remain unchanged.

Figure 14: Market-wide investment strtagy



Note: This figure depicts the average annual return of stocks with observed or missing (in middle of a stock sample) characteristics. Means are taken by month and then averaged across all monts in the sample.

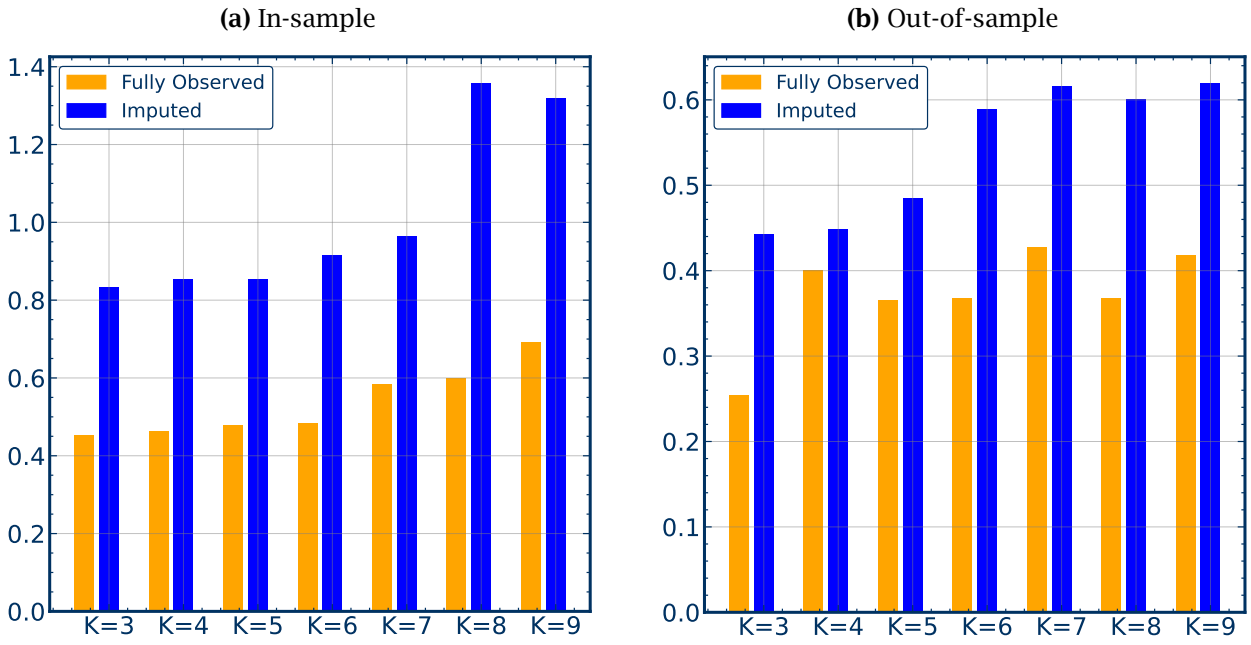
6.1.2. Investment with missing values

Neglecting firms with missing firm fundamentals leads to suboptimal investment decisions. We show that the out-of-sample performance of a conditional latent factor model is substantially better if it includes the larger set of companies with imputed characteristics.

We estimate the conditional latent factor model of Kelly et al. (2019) on the subset of stocks with fully observed characteristics and the larger set of stocks with imputed characteristics using the local B-XS imputation. The Instrumented Principal Component Analysis (IPCA) models the exposure to latent factors as a function of characteristics. Intuitively, the IPCA factors are obtained as PCA factors of characteristic managed portfolios. IPCA can only include stocks for the time periods when they have a complete set of characteristics. Hence, we either have to take a small subset of fully observed data, or need to impute the missing values.

We evaluate the performance of the IPCA factors based on the Sharpe ratio of the implied pricing kernel. Hence, we first obtain the mean-variance efficient combination the latent IPCA factors and report the Sharpe ratio of this investment strategy. A higher Sharpe ratio implies that the latent factors are a better approximation of the true pricing kernel. We show the in-sample and out-of-sample results for different numbers of latent factors. The out-of-sample analysis estimates the

Figure 15: Sharpe ratios with IPCA factors



Note: This figure shows the in- and out-of-sample Sharpe ratios of mean-variance efficient combination for different number of IPCA factors. We estimate a conditional latent factor model with the Instrumented Principal Component Analysis of Kelly et al. (2019). The estimation is either on the small subset of fully observed or the large set of all imputed stocks. The in-sample analysis is estimated on the full time period, while the out-of-sample analysis estimates the loadings and mean-variance efficient weights on the first half of the time-series and evaluates the portfolios on the second half.

IPCA model and mean-variance efficient combination on the first half of the sample, and reports the out-of-sample results for the second half of the panel.

Figure 15 shows the in-and out-of-sample Sharpe ratios for IPCA factors. Not surprisingly, the in-sample Sharpe ratios with more data are higher. This by itself is of limited value, as an in-sample analysis can overfit the data. Importantly, the out-of-sample Sharpe ratios with all stocks are also substantially higher than with the subset of fully observed stocks. This findings holds uniformly for any number of latent factors. In fact, a 3-factor model based on all stocks outperforms even a 9-factor model based on the subset of fully observed data.

Our finding has important implications. First, the stochastic discount factor (SDF) estimated on all stocks seems to be closer to the true SDF than the one estimated on only the subset of stocks with fully observed characteristics. Second an investor, who only invests in a non-representative subset of firms, foregoes profits. These results do not depend on the method used to extract the SDF. We obtain similar results for characteristic mimicking factors obtain from cross-sectional regressions

or machine learning prediction of returns with neural networks.[11]

In order to understand where this difference comes from, we now turn to the simpler, but very popular, way to construct cross-sectional strategies based on characteristics, namely decile-sorted portfolios.

6.1.3. Conditional sorts

In this section, we show how the selection bias with missing data affects conditional expected returns. We specifically focus on the simplest asset pricing application to dissect the implications for different characteristics and for conditional means and variances.

Most multivariate asset pricing applications, including multiple sorts, panel regressions on multiple characteristics or IPCA from the last subsection, require the presence of multiple characteristics. In order to illustrate the effect of requiring the presence of multiple characteristics, we focus on the properties of the most basic investment strategies, deciles sorts, and study how they are affected by the requirements of observing data for additional characteristics. Following the usual convention, the decile cutoff values are based on NYSE breakpoints, similar to Fama and French (1993).

First, we study the empirical effect of data selection and imputation on conditional returns for some of the most widely used characteristics, size (ME), book-to-market ratio (B2M), investment (INV), operating profitability (OP), momentum (R12_2) and long-term reversal (R60_12). In addition, we also consider the accounting based characteristics net share issues (NI) and expenses to sales (SGA2S), since those seem to be strongly affected by missing values. We construct value weighted decile sorted portfolios for the main characteristics, size, value, investment and operating profitability. In order to understand the effect of requiring the presence of multiple characteristics, we study the asset pricing implications for the first and last deciles of these four characteristic sorts, when requiring that additional characteristics are observed. In more detail, we first include only stocks that have the sorting characteristic available. Then, we take the subset of stocks for which also size is available. We continue stepwise, by incrementally requiring that in addition INV, OP, NI, SGA2S, R12_2, R60_13 or all 45 characteristics are available. The decile cutoff points remain the same NYSE breakpoints.

Figure 16 shows the Sharpe ratio, mean return, standard deviation and percentage of stocks used in the first and tenth decile. At first, we use the least restrictive sample of stocks that requires

[11] The results are available upon request.

Figure 16: Univariate Sorts With and Without Missing Values



Note: This figure shows the Sharpe ratio, average return, standard deviation and percentage of stocks for the univariate first and tenth value weighted characteristic sorted deciles for different subset of stock with and without imputation. We sequentially restrict the set of stocks to those that multiple characteristics available. First, we include all stocks for which only the sorting characteristic is available, then in addition we require in addition the availability of size (ME). In the next step, the sorting characteristics, size and investment (INV) need to be observed. We continue with operating profitability (OP), Net Share Issues (NI), Selling, general and Administrative expenses to sales (SGA2S, momentum (R12_2) and long-term reversal (R60_13). We sort based on book-to-market, size, investment and operating profitability. We impute missing values with our baseline local BW-XS model.

only that leading characteristic to be observed, progressively requiring more and more additional fundamentals to be observed. We also include the strategies with characteristic values imputed with

the local BW model, which is free of the look-ahead bias and could be easily used by investors in real time. The first obvious observation is that using all the stock with imputed values or all stocks for which we only require the availability of a single sorting variable, lead to essentially the same means and Sharpe ratios. This is reassuring, since it further confirms the validity of our imputation approach even for the firms that have fairly extreme values of the fundamentals. This result is in direct contrast with the situation where we simply require additional characteristics to be observed, which changes the composition of the decile portfolios, its rate of return and Sharpe ratio.

Requiring more characteristics drastically reduces the number of stocks that are included in portfolios sorts. In the case of size, the number of small stocks (decile 10, based on NYSE breakpoints) drops from almost 50% of all the tradable companies to less than 10%, whenever all the characteristics are required to be observed. Restricting stocks to have contemporaneous observations for the book-to-market (B2M), investment (INV), and operating profitability (OP), removes 15% of the overall sample. These results are even more extreme for portfolios sorted by the book-to-market ratio (see Panel A in Figure 16). In this case the number of available stocks for the extreme growth and value deciles drops from above 10% to about 2% of the sample, whenever all the characteristics are required to be observed. The requirement to observe ME, INV and OP (in addition to the book-to-market) already leads to a relative reduction of 10-20% of the initial number of firms, available for the strategy. The smaller number of stocks has an expected effect on the volatilities of the portfolio sorts, since one would expect having fewer stocks to lead to less diversified portfolios, and hence, higher overall volatility. Indeed, we observe that in most cases volatility increases. Note, however, that in general this does not have to yield a monotonic effect: since characteristics are not missing at random, both lower degree of diversification and firm selection contributes to the overall effect on volatility, making it difficult to predict the overall sign of the effect.

Importantly, the systematic structure in missing data creates a selection bias in mean returns. The mean returns of extreme deciles on investment, size, value, and operating profitability are already affected by requiring the presence of only three additional characteristics. Once again, due to the non-random nature of missingness, it can have an ambiguous effect on the risk premia. In all four cases, requiring the presence of all the characteristics leads overall to lower average returns. As the average returns tend to decrease in the more restrictive subsample of stocks, while the volatility effect increases in many cases, the Sharpe ratios tend to decrease as well. However, the exact effect on the Sharpe ratio and the corresponding t-statistics can be fairly complex.

Figure 17: Top and Bottom Deciles With and Without Missing Values



Note: This figure shows the Sharpe ratios and average returns for value weighted decile sorted portfolios, formed from stocks with observed single or full panel of characteristics. The left set of plots shows the Sharpe ratios of the top and bottom deciles, while the right set of plots shows the mean returns. The light blue and green bars correspond to the first and last deciles, comprised of a fully observed panel of stocks with all the characteristics. The dark blue and green bars correspond to the return on the extreme deciles formed by stocks required to have only the characteristic available used in sorting.

Our results extend to the conditional mean based on the majority of characteristics. Figure 17 shows the Sharpe ratios and mean returns for the top and bottom deciles of stocks, sorted by a given characteristic for two types of samples: first, requiring only that a single characteristic is observed, and second, requiring all 45 characteristics to be observed at the same time. We group firm-specific variables by their type, and report both the Sharpe ratio and average return of the corresponding deciles.

For most characteristics, the Sharpe ratios on the fully observed panel are lower than on the larger panel of firms with missing information. Consider for example, the case of sorting based on operating leverage (characteristic OL in the intangible category in Figure 17). In a fully observed panel, the Sharpe ratio of the bottom decile decile based on OL, is 25% lower compared to the case of a simple univariate sort that requires only a single observed characteristic. Similar patterns can be observed for dividend-to-price (D2P), momentum (R12_7), expenses-to-assets (DPI2A), spread (SPREAD), return on assets and equity (ROA/ROE), and many others. Hence, the combination of possible lower expected returns and/or higher volatility on a restricted sample can create a negative selection bias for simple asset pricing statistics. The directional effect on mean returns is more complex than for Sharpe ratios, emphasizing again the complex interaction between the sorting characteristics and missingness. It seems that in many cases, where mean returns are larger on the restricted sample, the increase in volatility dominates, thus resulting in a lower Sharpe ratio. The corresponding Sharpe ratios and mean returns of deciles with imputed data are very close to the sorts that require only a single characteristic to be observed.

The systematic selection bias in the expected returns of decile-sorted portfolios carries over to univariate long-short factors. Table B.10 in the Appendix reports the mean, standard deviation, Sharpe ratio, percentage and market value of missing characteristics for univariate long-short decile factors. As in the case of case of decile sorts, these factors are constructed with NYSE breakpoints. We compare the results when using (1) only stocks with fully observed 45 characteristics, (2) stocks with at least 10 characteristics observed and imputed data, (3) only the specific sorting characteristic observed, the combination of (2) and (3), and the difference between (2) and (3). The selection of stocks has obviously a strong effect on risk premia and Sharpe ratios, even for simple univariate long-short factors. As a long-short factor combines the impact of selection and imputation in the two separate legs, the effects can be complex and more or less pronounced than for the individual legs.

6.2. Imputation bias - Median imputation distorts asset pricing

Asset pricing results can depend on the imputation method. We show that median imputation substantially distorts the estimation of risk premia. Having established in the last subsection, that researchers should use all data to avoid the selection bias, we now compare the implications of different imputation methods. We study the fundamental problem of estimating the risk premium of characteristics from cross-sectional characteristic regressions. This is among the most widely used applications of characteristics. We follow the standard practice of running cross-sectional regressions of excess returns on lagged characteristics. These cross-sectional regressions require a complete vector of characteristics for each stock, and hence necessitate imputation.

We evaluate the risk premium estimates and correlation of factor mimicking portfolios with realistic masking of observed data. Simply comparing the risk premia estimates based on characteristics imputed either with the local B-XS or median values, reveals that we obtain quite different numbers, but it does not tell us which values are better. Hence, we take our full data set with missing values and mask additional characteristic values based on the logistic regression model in Table 2. This logistic regression propensity describes the empirical missingness pattern very well, and creates a realistic reference data set.[12] We impute these masked values either with the local B-XS or median values. The characteristic mimicking factor portfolios and their risk premia based on the observed entries without masking are the reference, and represent the true values.

The characteristic regressions yield time-series of characteristic mimicking factor portfolios:

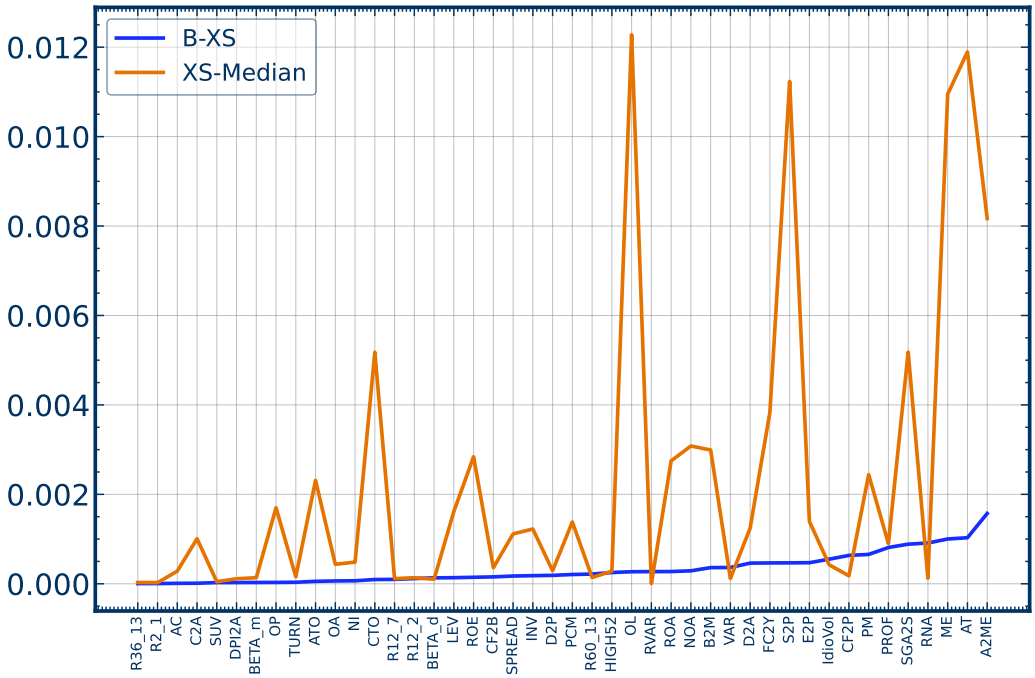
$$F_t^{\text{mimick}} = \left(\sum_{i=1}^{N_t} R_{i,t} C_i^{t-1} \right) \left(\sum_{i=1}^{N_t} C_i^{t-1} C_i^{t-1\top} \right)^{-1} .$$

The mean of these factor portfolio time-series corresponds to the risk premium for characteristics in the presence of other characteristics. We report the absolute error in characteristic risk premia of B-XS and median imputed values relative to true observed values. We also report the correlation in the time-series of the mimicking factor portfolios without masking and those with imputed values. Note that the characteristic regressions depend on all characteristics jointly. Hence, the imputation of some characteristics can affect even the return results for those characteristics that are fully observed.

Figure 18 shows the absolute errors in risk premia from cross-sectional regressions. The B-

[12] We obtain qualitatively similar results for block masking, which are available upon request.

Figure 18: Absolute error in risk premium from cross-sectional regressions

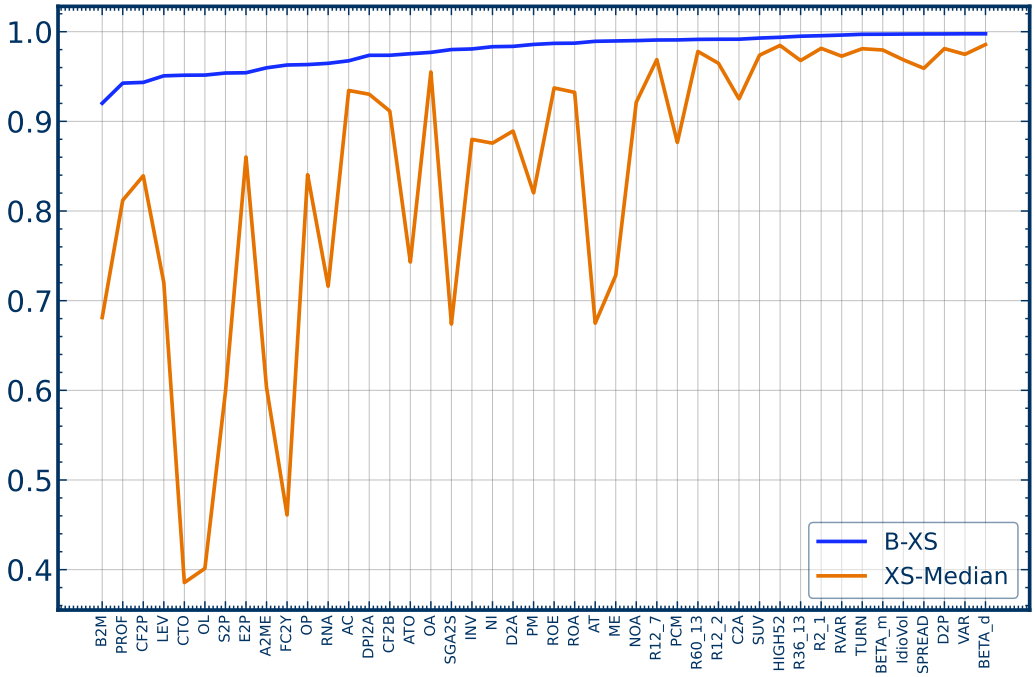


Note: This figure compares the absolute error in characteristic risk premia of B-XS and median imputed values relative to true observed values. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. The risk premium equals the mean of the mimicking factor portfolios. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed with the local B-XS or median value. The risk premium with observed entries is the reference.

XS imputed values have uniformly and substantially smaller risk premium errors compared to the median imputation. For some characteristics like Total Assets (AT) or Operating Leverage (OL), the error is around four to five times larger for the median imputation. We conclude that asset pricing metrics can be severely biased from using a naive median imputation.

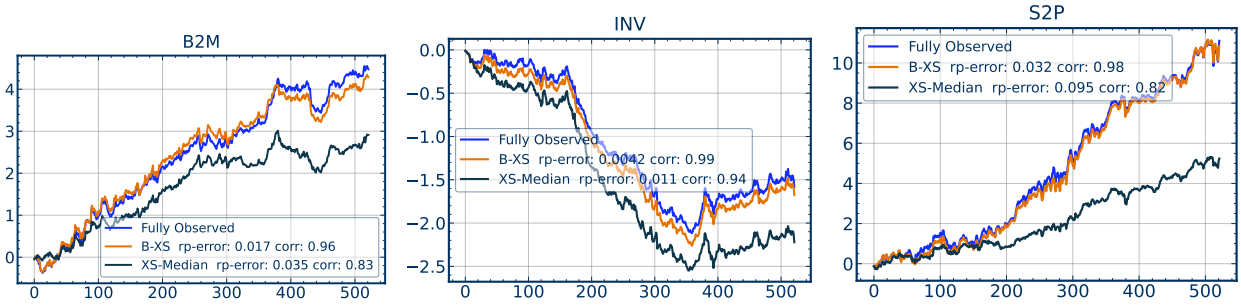
Figure 19 compares the R^2 of B-XS and median imputed values relative to observed values of characteristic projected portfolios. The R^2 measures the correlation in the time-series of the mimicking factor portfolios. The differences between the B-XS and median imputation are even more pronounced. The mimicking portfolio time-series are very close to the reference value for B-XS with correlations over 92% for all characteristics. In contrast, the median imputed time-series provide a poor approximation that is worse for all characteristics. In some cases, for example OL, the correlation is below 40%. This implies that that not only the first moment of the time-series, which measures the risk premium, is more precisely estimated with B-XS, but also the second moments. This matters for statistical inference, which is based on the covariances of the mimicking portfolios,

Figure 19: R^2 of factor mimicking portfolios from cross-sectional regressions



Note: This figure compares the R^2 of B-XS and median imputed values relative to true observed values of characteristic projected portfolios. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The R^2 measures the correlation in the time-series of the mimicking factor portfolios without masking and those with imputed values.

Figure 20: Characteristic mimicking factor portfolios



Note: These figures show the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference. We report the correlation and absolute error in characteristic risk premia.

and applications that require the time-series of the mimicking factor portfolios.

Figure 20 shows the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We highlight the results for book-to-market ratios (B2M), profitability (PROF) and sales-to-price ratios (S2P), while Figures C.9 to C.12 collect the results for the remaining characteristics with the same findings. The figures illustrate the precise approximation with the B-XS imputation. In contrast, we observe a substantial bias in the time-series for median imputation. This bias leads to wrong means, correlations and variances of the resulting characteristic mimicking factor portfolios.

7. Conclusion

This paper focuses on a very widespread yet rarely recognized issue of missing data in firm-specific characteristics. First, we document the systematic feature of missing data: it is pervasive and widespread among the overwhelming majority of firms. In our representative data set of the 45 most often used characteristics, more than 70% of firms are missing at least one of them at any given point of time. We show that firm fundamentals are not missing-at-random, but display complex systematic patterns. We leverage the complicated cross-sectional and time-series dependence in firm characteristics to propose a new imputation method, which is easy to use, and substantially outperforms existing alternatives.

Our findings are relevant for numerous applications in asset pricing, since, as we demonstrated, asset returns are affected by missing observations of the firm characteristics. The effects are particularly pronounced when requiring a large set of characteristics to be observed. While, for the sake of clarity, we demonstrate our findings with widely used univariate portfolio sorts, cross-sectional regressions and conditional latent factor models, we suspect it to have a first order effect in return predictability regressions of more complex models (including machine learning), as well as all the recently proposed advanced frameworks of stock returns that typically require a large balanced panel of stock characteristics.

Naturally, the problem of missing data does not just apply to stock-specific characteristics, and is encountered universally in various applications in finance: I/B/E/S forecast data, ESG ratings of firms, and many others. Given the Big Data environment, and new sources of information being available with an increasing speed, we suspect that the issue of missing data will become even more paramount going forward. We hope that our paper laid out the foundations and general guidelines for imputing missing data that could be applied in many different settings in the follow up work.

References

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, 2021, Matrix completion methods for causal panel data models, *Journal of the American Statistical Association* forthcoming.

Bai, J., and S. Ng, 2021, Matrix completion, counterfactuals, and factor analysis of missing data, *Journal of the American Statistical Association* 1746–1763.

Blanchet, J., F. Hernandez, V. A. Nguyen, M. Pelger, and X. Zhang, 2022, Bayesian imputation of missing data with optimal look-ahead-bias and variance tradeoff, *Working paper* .

Bryzgalova, S., M. Pelger, and J. Zhu, 2019, Forest through the trees: Building cross-sections of stock returns, *Working paper* .

Cahan, E., J. Bai, and Serena Ng, 2021, Factor-based imputation of missing values and covariances in panel data of large dimensions, *Working paper* .

Chen, L., M. Pelger, and J. Zhu, 2019, Deep learning in asset pricing, *Working paper* .

Connor, G., and R. Korajczyk, 1988, Risk and return in an equilibrium apt: Application to a new test methodology, *Journal of Financial Economics* 21, 255–289.

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.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber, 2021, Missing data in asset pricing panels, *Working paper* .

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

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

Jin, S., K. Miao, and L. Su, 2021, On factor models with random missing: Em estimation, inference, and cross validation, *Journal of Econometrics* 222, 745–777.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh, 2021, Machine-learning the skill of mutual fund managers, *Working paper* .

Kelly, B.T., S. Pruitt, and Y. Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Koh, Ping Sheng, and David M. Reeb, 2015, Missing R&D, *Journal of Accounting and Economics* 60, 73–94.

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

Lettau, M., 2022, High dimensional factor models with an application to mutual fund characteristics, *Working paper* .

Lettau, M., and M. Pelger, 2020a, Estimating latent asset pricing factors, *Journal of Econometrics* 218, 1–31.

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

Lyandres, Evgeny, Le Sun, and Lu Zhang, 2008, The new issues puzzle: Testing the investment-based explanation, *The Review of Financial Studies* 21, 2825–2855.

Pelger, M., 2019, Understanding systematic risk: A high-frequency approach, *Journal of Finance, forthcoming* .

Pelger, M., and R. Xiong, 2021a, Interpretable sparse proximate factors for large dimensions, *Journal of Business & Economic Statistics* 1–23.

Pelger, M., and R. Xiong, 2021b, State-varying factor models of large dimensions, *Journal of Business & Economic Statistics, forthcoming* .

Xiong, R., and M. Pelger, 2019, Large dimensional latent factor modeling with missing observations and applications to causal inference, *Journal of Econometrics, forthcoming* .

Xiong, R., and M. Pelger, 2022, The causal effect of publication on the cross-section of stock returns, *Working paper* .

Appendix A. Model

Implementation

In this Appendix we provide a modification of our latent factor model estimator in Section 3.1, which is faster and easier to implement. The estimation of the eigenvectors of the $N_t \times N_t$ dimensional characteristic covariance matrix $\tilde{\Sigma}^{\text{XS},t}$ are computationally expensive. For fully observed data, up to some normalization, the PCA estimation is "symmetric" in the two dimensions and we could base our analysis on the eigenvectors of the $L \times L$ matrix $\frac{1}{N_t}\sum_{i=1}^{N_t} C_i^t C_i^{t\top}$. However, in the presence of missing data, this would impose different assumptions on the missing pattern.

Here we propose a modification of the estimator in Section 3.1, that empirically results in essentially the same estimated model. First, we estimate "noisy" loadings $\tilde{\Lambda}^t \in \mathbb{R}^{L\times K}$ as the eigenvectors of the K largest eigenvalues of the $L \times L$ matrix

$$\frac{1}{|O_{l,p}^t|}\sum_{i\in O_{l,p}^t} C_{i,l}^t C_{i,p}^t,$$

where $O_{l,p}^t$ is the set of all stocks that have the characteristics l and p observed at time t. By construction $|O_{l,p}^t| \leq N_t$. The characteristic factors follow from a regression on the estimated $\tilde{\Lambda}$:

$$\hat{F}_i^t = \left(\sum_{l=1}^{L} W_{i,l}^t \tilde{\Lambda}_l^t \tilde{\Lambda}_l^{t\top}\right)^{-1}\left(\sum_{l=1}^{L} W_{i,l}^t \tilde{\Lambda}_l^t C_{i,l}^t\right). \tag{A.1}$$

In a last step, we obtain the loadings from a regression that accounts for missing observations:

$$\hat{\Lambda}_l^t = \left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t \hat{F}_i^{t\top}\right)^{-1}\left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t C_{i,l}^t\right).$$

The regressions, weighted by observed values, provide valid estimates even when the missing pattern depends on the factors. The first matrix, whose eigenvectors are used to extract the noisy loadings, imposes some restriction on the missing pattern. However, as long as the noisy loadings are correlated with the actual loadings, the third regression corrects for the complex missing pattern structure. The advantage of this second approach is that it is much faster to implement. It is motivated by the iterative PCA estimation, which is discussed among others in Xiong and Pelger (2019) for missing values and Pelger and Xiong (2021a) for noisy loadings in the case of fully observed data. This alternative implementation also motivates the interpretation of the loadings Λ as "characteristic portfolio weights", which provides insights into the economic meaning of the characteristic

factors.

Under the assumptions of constant loadings, the estimation is modified as follows. First, we estimate noisy loadings $\tilde{\Lambda}$ as the eigenvectors of the K largest eigenvalues of $\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_t}\sum_{i=1}^{N_t} C_i^t C_i^{t\top}\right)$. The second step of the factor estimation $\hat{F}_i^t$ is the same, and in the third step we use the pooled regression

$$\hat{\Lambda}_l = \left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t \hat{F}_i^{t\top}\right)\right)^{-1}\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{N_t} W_{i,l}^t \hat{F}_i^t C_{i,l}^t\right)\right).$$

Robust Finite Sample Performance with Adaptive Regularization

The finite sample performance of the latent factor model can be improved through adaptive ridge regularization without affecting the asymptotic inferential theory. The latent characteristic factors $\hat{F}_i^t$ are weighted averages of observed characteristics. Hence, the common component $\hat{C}_{i,l}^t = \hat{F}_i^t \hat{\Lambda}_l^{t\top}$ can be interpreted as a weighted average of observed characteristics. The asymptotic theory underlying the estimation for the latent characteristic factors assumes that the number of observed characteristics for a specific stock is large, but is allowed to grow at a lower rate than the number of all characteristics.

In our finite sample, there are cases where some stocks only have a small number of observed characteristics. Hence, the characteristic factors $\hat{F}_i^t$ are averages over only a few entries. This can become an issue if the characteristics with the largest factor weights in Λ_l^t are unobserved, that is, the average is taken only over entries that would have negligible weights in the case of fully observed characteristics.

We solve this problem with an adaptive ridge regression. In more detail, we add a ridge penalty to the regression A.1 of the observed characteristics on the estimated loadings $\tilde{\Lambda}$:

$$\hat{F}_i^t = \left(\sum_{l=1}^{L} W_{i,l}^t \tilde{\Lambda}_l^t \tilde{\Lambda}_l^{t\top} + \gamma_{i,t} I_K\right)^{-1}\left(\sum_{l=1}^{L} W_{i,l}^t \tilde{\Lambda}_l^t C_{i,l}^t\right).$$

The ridge penalty $\gamma_{i,t}$ shrinks characteristic factors with observed characteristics, that have only small factor weights, towards a cross-sectional median. For example, the first latent factor loads heavily on profitability characteristics. If a particular stocks at a specific time has all profitability characteristics missing, then its latent factor $\hat{F}_i^t$ would overweight the less relevant observed characteristics. A larger ridge penalty would shrink this specific factor realization towards zero.

The adaptive ridge penalty only applies shrinkage when it is needed. If sufficiently many charac-

teristics are observed, it is suboptimal to apply shrinkage. In our example of the first latent factor, that loads heavily on profitability characteristics, it is sufficient to observe some of the profitability characteristics for a specific stock to approximate this factor well. Hence, the adaptive ridge penalty takes into account the amount of observed entries. More specifically, $\gamma_{i,t}$ is an exponentially decaying function in the number of observed characteristics of stock i at time t. The decay exponent is a universal constant selected by cross-validation. Hence, for a sufficiently many observed entries, the penalty converges to zero and no shrinkage is applied. However, for stocks that have only very few observed characteristics, the penalty has an effect. The exponentially fast decay also implies that the asymptotic theory in Xiong and Pelger (2019) is not affected.

The adaptive regularization is beneficial for logistic masking results. In that case, it is possible that for some stocks a very large number of entries in the same characteristic group is masked. For those stocks, the regularization provides robust out-of-sample results. In case of missing-at-random or block-masking, the out-of-sample results with and without regularization are very similar. Hence, we suggest to include the adaptive regularization for robustness in finite samples and include it in our benchmark XS models.

Multihorizon Forecasts

In this appendix we discuss the prediction for longer horizons. Our pure cross-sectional models (XS) only use contemporaneous information and as such do not impose any assumptions on the time-series dynamics. In the main text we estimate models that incorporate the time-series dynamics of characteristics. The estimated models (B-XS, BF-XS, B) estimate one-step ahead forecasts. However, for blocks of missing time-series obervations we face the issue of a longer horizon forecast. Using a time-series model for a multi-step prediction requires to make further assumptions on the dynamics of the cross-sectional factors and the non-systematic component.

The implementations of our baseline models for B-XS, BF-XS and B, that we use in the main text, estimate the parameters based on one-step ahead forecasts, and plug in the last observed value for the multi-step forecast. This means that for our implementation of the B-XS model, the prediction for s periods into the future, given that $C^l_{i,t-1}$ is the last observed value, is

$$\hat{C}^{l,\text{B-XS}}_{i,t+s-1} = \left(\hat{\beta}^{l,\text{B-XS}}\right)^\top \begin{pmatrix} C^l_{i,t-1} & \hat{F}^t_{i,1} & \cdots & \hat{F}^t_{i,K} \end{pmatrix}.$$

This is different from a recursive imputation of the missing values, which uses the imputed values

from the last period as an input for the imputation of the next period. Note that a recursive model requires to take a stand on the time-series dynamics of the factors and idiosyncratic component, which was not part of the estimation. Essentially, the $\hat{\beta}^{l,\text{B-XS}}$ could be different for different horizon forecasts. In this appendix, we present a more general model, which includes our baseline implementation as a special case. Our more general model estimates the dynamic multi-horizon structure from the data. We can show that our simple baseline implementation is actually close to an optimal model. Given its parsimonious structure, the main text only focuses on this transparent model.

Fundamentally, the key element of our model is to combine the information from the contemporaneous cross-section and the time-series dimension. A general model can be casted as a weighted average of separate forecasts that uses different information sets.

First, we start with the same pure cross-sectional factor model as before:

$$C_{i,l}^t = F_i^t \Lambda_l{}^\top + e_{i,l}^t \qquad \text{with } i = 1, ..., N_t \text{ and } l = 1, ..., L.$$

Given the estimated factors and loadings, we obtain our pure XS forecast:

$$\hat{C}_{i,l}^{t,\text{XS}} = \hat{F}_i^t \hat{\Lambda}_l^\top.$$

Importantly, this forecast is available for all entries. We have shown empirically, that a pure XS model can be improved when combined with time-series information. This is done in the weighted average step, where we will also distinguish between a model that only uses past information or also, in addition, future information:

B-XS weighted model:

$$\mathbb{E}\left[C_{i,t,l} | C_{i,t-s,l}, F_i^t\right] = w_s^{\text{XS},l} \hat{C}_{i,l}^{t,\text{XS}} + w_s^{\text{B},l} C_{i,t-s,l}.$$

BF-XS weighted model:

$$\mathbb{E}\left[C_{i,t,l} | C_{i,t-s,l}, C_{i,t+k,l}, F_i^t\right] = w_s^{\text{XS},l} \hat{C}_{i,l}^{t,\text{XS}} + w_s^{\text{B},l} C_{i,t-s,l} + w_k^{\text{F},l} C_{i,t+k,l}.$$

This model is closely related to our baseline benchmark model, but allows to deal with multi-horizon forecasts in a more systematic way. Consider $s = 1$, i.e. we use only a one-step ahead forecast. In

this case the weighted backward model can be expressed as

$$\mathbb{E}\left[C_{i,t,l}|C_{i,t-1,l},F_i^t\right] = \begin{pmatrix}(w_1^{\text{XS},l}\Lambda_1) & \cdots & (w_1^{\text{XS},l}\Lambda_K) & w_s^{\text{TS-B},l}\end{pmatrix}^\top \begin{pmatrix}\hat{F}_{i,1}^t & \cdots & \hat{F}_{i,K}^t & C_{i,t-1,l}\end{pmatrix}.$$

This means that the weighted model is a special case of the baseline benchmark model for a one-period prediction, but with a constraint on $\beta^{l,\text{B-XS}}$. The constraint is sensible as it imposes that $F_i^t\Lambda^\top$ captures the pure XS characteristic information, while the time-series information provides the right level for the forecast, without changing the relative cross-sectional weighting of the pure XS factors.

The weighted framework allows more flexibility for multi-period forecasts with a small number of parameters and without a priori imposing strong assumptions on the time-series structure. We could easily obtain a non-parametric model for the weights $w_s^{\text{XS},l}$, $w_s^{\text{B},l}$ and $w_k^{\text{F},l}$. One implementation, which could be viewed as a non-parametric estimation, is to simply estimate different models for each forecast horizon without further restricting them. For a specific characteristic l and a specific lack s, we could stack the characteristics $C_{i,t}^l$ over time and the cross-section and run a regression on the stacked values of $C_{i,t-s}^l$ and $\hat{C}_{i,l}^{t,\text{XS}}$. However, we suggest to impose some structure on the weights.

Guided by our empirical findings, the following parametric model provides a parsimonious and interpretable framework:

$$\begin{aligned} w_s^{\text{TS-B},l} &= a^{\text{B},l} + b^{\text{B},l}e^{-\gamma^{\text{B},l}s} \\ w_k^{\text{TS-F},l} &= a^{\text{F},l} + b^{\text{F},l}e^{-\gamma^{\text{F},l}k} \\ w_s^{\text{XS},l} &= a^{\text{XS},l} + b^{\text{XS},l}e^{-\gamma^{\text{XS},l}\min(s,k)}. \end{aligned}$$

This means that the B-XS and B-XS weighted model can be expressed as

BW weighted model:

$$\mathbb{E}\left[C_{i,t,l}|C_{i,t-s,l},F_i^t\right] = \left(a^{\text{XS},l} + b^{\text{XS},l}e^{-\gamma^{\text{XS},l}s}\right)\hat{C}_{i,l}^{t,\text{XS}} + \left(a^{\text{B},l} + b^{\text{B},l}e^{-\gamma^{\text{B},l}s}\right)C_{i,t-s,l}$$

BWFW weighted model:

$$\begin{aligned}\mathbb{E}\left[C_{i,t,l}|C_{i,t-s,l},C_{i,t+k,l},F_i^t\right] = &\left(a^{\text{XS},l} + b^{\text{XS},l}e^{-\gamma^{\text{XS},l}\min(s,k)}\right)\hat{C}_{i,l}^{t,\text{XS}} + \left(a^{\text{B},l} + b^{\text{B},l}e^{-\gamma^{\text{B},l}s}\right)C_{i,t-s,l} \\ &+ \left(a^{\text{F},l} + b^{\text{F},l}e^{-\gamma^{\text{F},l}k}\right)C_{i,t+k,l}\end{aligned}$$

This model has as two special cases: One special case keeps a weight of one on the last observed

value, the second special case interpolates linearly between the last and first observed values. The parameters of the weight functions can be easily estimated from minimizing the squared error of

$$\sum_{t=1}^{T}\sum_{i=1}^{N_t}\sum_{s=1}^{S} W_{i,l}^{t}\left(C_{i,t}^{l} - w_s^{\mathrm{XS},l}\hat{C}_{i,l}^{t,\mathrm{XS}} - w_s^{\mathrm{B},l}C_{i,t-s,l}\right)^2$$

on the observed data.

The parametric model formulation has the benefit of being easy to interpret. The value of $a+b$ measures the short-term effect, and γ measures the decay in information. A very persistent characteristic is expected to have a large value for a but a small value for b and γ.[13]

The more flexible model, that allows for horizon dependent weights, does not lead to substantial improvements relative to our baseline implementation. Table A.1 compares the global B-XS as implemented in the main text, and the more flexible global weighted B-XS. Note, that in-sample a more flexible model will by construction always result in smaller RMSE. However, the differences seem to be very small. For OOS missing-at-random we deal primarily with one-step ahead forecasts, and hence the additional flexibility of the weighed B-XS cannot help. The only case, where the more general structure can be relevant, is the OOS block-missing analysis. However, the improvements seem to be minor. We conclude that our simple model is almost as good as a more complex model. Hence, we favor the more parsimonious model as our baseline.

Table A.1: Imputation Error for Different Imputation Methods

	In-Sample			OOS MAR			OOS Block			OOS Logit		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
weighted B-XS	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.13	0.14	0.13	0.12	0.15
global B-XS	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.13	0.12	0.15

Note: This table shows imputation RMSE for the global B-XS and global weighted B-XS methods averaged over all characteristics and separately for monthly and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data either missing at random or missing in time-series blocks for 12 consecutive months.

[13] The weighted framework can be generalized to include a time-series forecasting model. This forecast could be based for example on an autoregressive model or a more complex non-parametric time-series model. Given the information set I_{t-s} and a forecasting model, we could include the forecast $\mathbb{E}\left[C_{i,t,l}|_{t-s}\right]$ in the weighted model with an additional weight $w_s^{\mathrm{TS},l}$.

Appendix B. Tables

Table B.1: Missing by Characteristic Quintiles

	All	ME Quintile				Characteristic Quintile			
		[1-2]	(2-3]	(3-4]	(4-5]	[1-2]	(2-3]	(3-4]	(4-5]
A2ME	12.43%	13.44%	10.51%	10.23%	9.93%	8.50%	9.56%	11.43%	15.25%
AC	43.20%	39.89%	34.04%	32.28%	26.67%	52.34%	26.01%	23.93%	51.18%
AT	12.43%	13.44%	10.51%	10.23%	9.93%	11.25%	10.20%	9.29%	9.01%
ATO	19.36%	22.33%	17.71%	16.24%	14.06%	19.27%	15.69%	14.11%	14.89%
B2M	10.69%	12.13%	8.67%	7.95%	6.63%	8.53%	7.75%	8.59%	12.31%
BETA_d	46.97%	56.44%	48.95%	44.73%	31.59%	39.19%	29.91%	28.54%	38.25%
BETA_m	35.85%	43.79%	37.57%	33.96%	23.76%	35.33%	22.39%	21.68%	32.85%
C2A	14.54%	15.49%	12.28%	12.10%	12.39%	15.45%	14.34%	12.39%	7.57%
CF2B	11.99%	14.17%	10.00%	8.86%	7.11%	9.73%	10.20%	10.09%	13.30%
CF2P	8.94%	10.81%	7.16%	5.38%	2.86%	8.62%	6.35%	6.36%	5.77%
CTO	19.35%	22.32%	17.70%	16.23%	14.06%	19.37%	15.25%	14.60%	15.24%
D2A	24.79%	25.89%	21.39%	20.77%	19.39%	22.07%	18.57%	18.61%	19.21%
D2P	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
DPI2A	55.95%	51.92%	52.41%	50.37%	44.98%	57.90%	37.42%	33.58%	38.17%
E2P	8.94%	10.81%	7.16%	5.38%	2.86%	8.70%	6.33%	5.94%	9.14%
FC2Y	28.24%	28.17%	24.02%	22.34%	23.87%	15.19%	17.68%	17.27%	20.42%
HIGH52	61.96%	70.83%	64.36%	60.54%	44.51%	83.61%	59.03%	49.68%	78.85%
INV	33.04%	38.42%	32.44%	30.16%	24.25%	43.89%	23.13%	21.88%	37.65%
IdioVol	0.04%	0.09%	0.03%	0.01%	0.00%	0.05%	0.03%	0.03%	0.05%
LEV	16.87%	16.14%	13.46%	14.17%	13.40%	12.68%	12.97%	13.45%	16.62%
ME	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
NI	32.01%	39.49%	32.54%	29.44%	22.76%	39.96%	23.09%	24.72%	32.03%
NOA	20.41%	23.11%	19.00%	17.98%	16.52%	17.71%	17.17%	17.08%	15.99%
OA	32.31%	24.86%	20.88%	20.51%	19.30%	40.22%	17.57%	15.58%	42.48%
OL	14.88%	16.34%	12.74%	12.30%	12.36%	15.26%	11.42%	11.68%	13.30%
OP	18.95%	14.32%	10.00%	8.81%	7.08%	10.94%	10.85%	9.61%	8.99%
PCM	17.12%	21.26%	16.81%	13.15%	10.61%	17.53%	14.05%	11.89%	10.13%
PM	13.91%	14.98%	11.53%	10.82%	9.91%	11.82%	11.73%	11.56%	14.21%
PROF	18.24%	21.22%	16.95%	15.13%	11.73%	18.78%	13.32%	12.78%	14.74%
Q	12.43%	13.44%	10.51%	10.23%	9.93%	14.38%	11.61%	9.76%	8.32%
R12_2	20.73%	26.04%	21.98%	19.41%	13.29%	36.47%	14.75%	11.49%	41.87%
R12_7	20.56%	25.75%	21.80%	19.32%	13.23%	39.37%	15.27%	12.00%	45.58%
R2_1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
R36_13	48.09%	58.13%	50.21%	45.42%	32.03%	57.91%	28.88%	22.84%	57.89%
R60_13	63.55%	74.31%	66.17%	60.78%	44.31%	63.36%	36.02%	29.05%	56.02%
RNA	21.66%	24.03%	19.65%	18.63%	17.24%	21.01%	16.50%	15.87%	18.25%
ROA	24.85%	28.86%	23.71%	21.98%	18.45%	25.90%	20.29%	17.08%	20.22%
ROE	23.15%	27.61%	21.93%	19.76%	15.17%	25.53%	17.74%	14.86%	20.98%
RVAR	0.04%	0.07%	0.03%	0.01%	0.03%	0.02%	0.02%	0.03%	0.04%
S2P	9.27%	11.08%	7.26%	5.42%	2.91%	7.87%	6.21%	6.50%	8.21%
SGA2S	28.27%	28.23%	24.03%	22.35%	23.87%	14.81%	17.56%	17.45%	20.65%
SPREAD	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
SUV	7.74%	10.50%	8.07%	6.30%	4.23%	28.97%	6.01%	7.66%	36.33%
TURN	5.55%	7.80%	5.57%	4.30%	2.82%	9.18%	4.80%	3.53%	3.96%
VAR	0.04%	0.07%	0.03%	0.01%	0.03%	0.02%	0.02%	0.03%	0.04%

Note: This table reports the percentage of missing observations for different size and characteristic quintiles. The means are pooled by stocks.

Table B.2: Lengths of Missing Blocks

	number of gaps	mean length	median length		number of gaps	mean length	median length
A2ME	11693	11.14	9	OA	3814	18.6	7
AC	11948	12	9	OL	11320	8.85	3
AT	11693	11.14	9	OP	6542	11.75	6
ATO	7550	11.95	9	PCM	10324	9.65	3
B2M	11617	11.16	9	PM	11535	9.54	3
BETA_d	1324	31.46	4	PROF	11595	11.3	9
BETA_m	1556	28.58	5	Q	11693	11.14	9
C2A	6599	12.18	6	R12_2	1406	42.02	23
CF2B	6447	11.93	6	R12_7	2165	26.92	7
CF2P	4770	13.93	6	R2_1	2040	25.54	6
CTO	7458	12.05	9	R36_13	1812	33.59	23
D2A	14002	14.67	9	R60_13	1169	44.34	48
D2P	2040	25.54	6	RNA	12979	9.61	6
DPI2A	5612	29.51	12	ROA	6968	12.42	9
E2P	4770	13.93	6	ROE	6818	12.57	9
FC2Y	7927	15.5	9	RVAR	2019	25.89	7
HIGH52	1137	23.45	4	S2P	5238	13.32	6
INV	13076	11.28	9	SGA2S	7919	15.52	9
IdioVol	2162	24.31	6	SPREAD	2085	25.01	6
LEV	13952	13.64	9	SUV	2129	22.96	4
ME	2040	25.54	6	TURN	2156	22.53	3
NI	8757	12.11	9	VAR	2019	25.89	7
NOA	4071	16.71	7				

Note: This table shows the number of missing blocks and their mean and median length for each characteristic.

Table B.3: OOS RMSE for Different Cross-Sectional Factor Models

Number of factors	all characteristics	quarterly characteristics	monthly characteristics
		local B-XS	
1	0.143	0.142	0.145
2	0.142	0.141	0.145
3	0.142	0.141	0.144
4	0.142	0.140	0.144
5	0.142	0.140	0.145
6	0.142	0.140	0.145
7	0.142	0.140	0.146
8	0.143	0.139	0.149
9	0.148	0.141	0.158
		global B-XS	
1	0.144	0.142	0.146
2	0.143	0.142	0.146
3	0.143	0.142	0.146
4	0.143	0.141	0.146
5	0.143	0.141	0.147
6	0.143	0.140	0.149
7	0.146	0.140	0.154
8	0.150	0.142	0.164
9	0.174	0.153	0.207
		local XS	
1	0.261	0.261	0.262
2	0.248	0.245	0.252
3	0.238	0.232	0.249
4	0.234	0.228	0.245
5	0.232	0.226	0.243
6	0.232	0.225	0.243
7	0.232	0.226	0.244
8	0.236	0.229	0.249
9	0.251	0.240	0.270

Note: This table shows the out-of-sample imputation RMSE for different number of factors for the local and global cross-sectional factor model with or without the past time-series information. For the out-of-sample analysis we mask 10% of the data randomly in time-series blocks of 12 consecutive months.

Table B.4: OOS RMSE in Characteristic Space for XS Factor Models

Number of factors	all characteristics	quarterly characteristics	monthly characteristics
Constant factor weights on ranks			
6	0.814	0.775	0.877
Factor model on normalized raw characteristics, global fit			
1	0.956	0.944	0.967
2	0.931	0.910	0.960
3	0.925	0.903	0.955
4	0.919	0.888	0.965
5	0.930	0.894	0.983
6	0.965	0.913	1.047
7	1.021	0.951	1.130
8	1.125	1.021	1.287
9	1.569	1.356	1.888
Factor model on normalized raw characteristics, local fit			
1	0.956	0.946	0.964
2	0.942	0.928	0.956
3	0.936	0.922	0.952
4	0.935	0.920	0.951
5	0.940	0.926	0.956
6	0.957	0.937	0.982
7	0.969	0.954	0.986
8	0.992	0.980	1.003
9	1.058	1.020	1.113
Factor model on kernel transformation of ranks global fit			
1	0.916	0.905	0.927
2	0.871	0.852	0.897
3	0.833	0.798	0.888
4	0.823	0.791	0.872
5	0.813	0.779	0.866
6	0.819	0.783	0.875
7	0.841	0.806	0.893
8	0.943	0.924	0.968
9	1.394	1.277	1.576
Factor model on kernel transformation of ranks local fit			
1	0.913	0.902	0.923
2	0.867	0.850	0.889
3	0.835	0.807	0.877
4	0.824	0.798	0.863
5	0.818	0.791	0.858
6	0.816	0.789	0.856
7	0.821	0.793	0.862
8	0.839	0.810	0.883
9	0.919	0.839	1.043

Note: This table shows the out-of-sample imputation RMSE in the original characteristic space without transforming characteristics into ranks. The characteristics are normalized by their cross-sectional mean and variance. The RMSE are further normalized by the RMSE of a medial that sets imputed values to zero, i.e. a simple median imputation. The first model is our baseline factor model estimated on ranks and transformed back into the characteristic space with the empirically estimated density function of each characteristic. We estimate the density function with the machine learning method, k-nearest neighbor. The second and third model estimates the factor model directly on the characteristics. In the fourth and fifth case, we estimate the factor model in the kernel transformed space with a Gaussian kernel and revert it back to the raw characteristics. For the out-of-sample analysis we mask 10% of the data randomly in time-series blocks of 12 consecutive months.

Table B.5: Imputation Error By Size Deciles

		In-Sample			OOS MAR			OOS Block		
size decile	method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
1	local B-XS	0.16	0.15	0.17	0.18	0.17	0.19	0.22	0.22	0.22
	local XS	0.23	0.24	0.22	0.26	0.27	0.26	0.25	0.26	0.24
	local B	0.17	0.16	0.18	0.19	0.17	0.20	0.24	0.24	0.23
2	local B-XS	0.15	0.14	0.15	0.16	0.16	0.17	0.20	0.21	0.19
	local XS	0.21	0.22	0.20	0.24	0.25	0.23	0.23	0.24	0.22
	local B	0.16	0.14	0.16	0.17	0.16	0.18	0.21	0.22	0.20
3	local B-XS	0.14	0.13	0.15	0.16	0.15	0.16	0.19	0.20	0.18
	local XS	0.20	0.21	0.20	0.23	0.24	0.23	0.22	0.23	0.22
	local B	0.15	0.14	0.16	0.16	0.16	0.17	0.20	0.22	0.20
4	local B-XS	0.14	0.13	0.15	0.16	0.15	0.16	0.19	0.19	0.18
	local XS	0.20	0.21	0.19	0.23	0.24	0.22	0.22	0.23	0.22
	local B	0.15	0.14	0.16	0.16	0.15	0.17	0.20	0.21	0.19
5	local B-XS	0.14	0.13	0.14	0.15	0.14	0.16	0.18	0.19	0.17
	local XS	0.20	0.21	0.19	0.22	0.23	0.22	0.22	0.22	0.21
	local B	0.15	0.13	0.15	0.16	0.15	0.16	0.19	0.21	0.18
6	local B-XS	0.13	0.12	0.14	0.15	0.14	0.15	0.18	0.18	0.17
	local XS	0.20	0.20	0.19	0.22	0.23	0.22	0.21	0.22	0.21
	local B	0.14	0.13	0.15	0.15	0.14	0.16	0.19	0.20	0.18
7	local B-XS	0.13	0.12	0.14	0.14	0.14	0.15	0.17	0.18	0.16
	local XS	0.19	0.20	0.19	0.22	0.22	0.22	0.21	0.22	0.21
	local B	0.14	0.13	0.14	0.15	0.14	0.15	0.18	0.20	0.17
8	local B-XS	0.13	0.12	0.14	0.14	0.13	0.15	0.17	0.18	0.16
	local XS	0.19	0.20	0.19	0.22	0.22	0.22	0.21	0.21	0.21
	local B	0.14	0.13	0.14	0.15	0.14	0.15	0.18	0.19	0.17
9	local bw	0.13	0.12	0.13	0.14	0.13	0.14	0.16	0.17	0.16
	local XS	0.19	0.20	0.19	0.22	0.22	0.22	0.21	0.21	0.21
	local B	0.14	0.12	0.14	0.14	0.13	0.15	0.18	0.18	0.17
10	local B-XS	0.12	0.11	0.13	0.13	0.13	0.14	0.16	0.16	0.15
	local XS	0.19	0.19	0.19	0.21	0.21	0.22	0.21	0.21	0.21
	local B	0.13	0.12	0.14	0.14	0.13	0.14	0.17	0.18	0.16

Note: This table shows out of sample imputation RMSE by imputation method for each size decile, overall and also for monthly updated and quarterly updated characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data either missing at random or missing in time-series blocks for 12 consecutive months.

Table B.6: Information used for Imputation for B-XS Model

Characteristic	F1	F2	F3	F4	F5	F6	Prev Val
A2ME	-0.008	0.077	0.007	-0.007	0.038	-0.008	0.816
AC	-0.001	-0.011	-0.009	-0.002	0.010	0.039	0.841
AT	-0.003	-0.001	0.001	-0.004	-0.000	0.001	0.990
ATO	-0.003	-0.009	-0.056	0.031	-0.014	-0.021	0.857
B2M	-0.002	0.032	0.002	0.004	0.023	0.004	0.916
BETA_d	0.006	-0.007	-0.007	-0.027	0.005	-0.018	0.943
BETA_m	0.003	-0.005	-0.004	-0.013	0.001	-0.009	0.965
C2A	0.002	-0.004	0.002	0.004	-0.001	-0.004	0.982
CF2B	-0.027	-0.008	0.002	0.022	0.041	-0.005	0.822
CF2P	-0.020	0.009	-0.009	-0.002	0.021	0.011	0.925
CTO	-0.001	-0.005	-0.042	0.013	-0.009	-0.002	0.909
D2A	0.003	-0.001	-0.008	-0.005	-0.004	-0.006	0.962
D2P	-0.004	0.002	0.001	0.002	0.000	-0.001	0.985
DPI2A	-0.018	-0.037	-0.013	-0.061	0.011	0.102	0.660
E2P	-0.030	0.002	-0.005	0.010	0.023	0.013	0.883
FC2Y	0.005	-0.005	0.005	0.006	0.007	-0.009	0.968
HIGH52	-0.051	-0.022	0.029	0.047	-0.106	0.019	0.681
INV	-0.019	-0.028	0.004	-0.029	0.018	0.062	0.850
IdioVol	0.132	-0.011	-0.059	-0.049	0.113	0.066	0.506
LEV	-0.001	0.008	-0.000	-0.011	0.003	0.000	0.969
ME	-0.013	-0.010	0.002	-0.013	-0.018	-0.010	0.948
NI	0.009	-0.012	0.007	-0.023	-0.001	0.030	0.910
NOA	-0.005	0.004	-0.014	-0.024	0.005	0.036	0.910
OA	-0.007	-0.007	-0.015	0.010	0.013	0.018	0.661
OL	0.008	-0.004	-0.038	0.017	-0.017	-0.010	0.908
OP	-0.026	-0.005	-0.019	-0.009	0.028	0.000	0.873
PCM	-0.002	-0.005	0.005	0.003	0.012	-0.003	0.971
PM	-0.014	-0.003	0.005	-0.001	0.011	0.005	0.947
PROF	-0.004	-0.002	-0.020	0.009	0.014	-0.016	0.934
Q	0.010	-0.082	-0.009	0.007	-0.037	0.005	0.802
R12_2	-0.036	-0.047	0.006	0.014	-0.055	0.101	0.721
R12_7	-0.043	-0.057	0.006	0.016	-0.061	0.136	0.652
R2_1	-0.042	-0.050	0.006	0.008	-0.122	0.003	-0.056
R36_13	-0.016	-0.018	-0.000	-0.003	0.016	0.002	0.906
R60_13	-0.012	-0.014	0.000	-0.003	0.011	0.000	0.941
RNA	-0.022	-0.022	-0.010	0.013	0.014	-0.003	0.886
ROA	-0.032	-0.024	-0.019	0.007	0.022	0.013	0.875
ROE	-0.042	-0.026	-0.017	0.005	0.039	0.009	0.835
RVAR	0.112	-0.008	-0.051	-0.038	0.100	0.060	0.584
S2P	-0.003	0.038	-0.040	0.005	0.012	-0.008	0.878
SGA2S	0.005	-0.004	0.005	0.006	0.006	-0.007	0.974
SPREAD	0.094	-0.008	-0.028	-0.023	0.070	0.045	0.626
SUV	0.001	-0.015	-0.003	-0.052	-0.019	-0.022	0.025
TURN	0.005	-0.067	-0.029	-0.143	-0.019	-0.069	0.637
VAR	0.126	-0.017	-0.059	-0.064	0.108	0.053	0.520

Note: This table shows the the regression coefficients on the cross-sectional factor model and the past time-series information of the global BW-XS model. We report the coefficients on each of the six factors and the past value.

Table B.7: Information used for Imputation for BF-XS Model

Characteristic	F1	F2	F3	F4	F5	F6	Prev Val	Next Val
A2ME	-0.005	0.045	0.004	-0.003	0.022	-0.005	0.466	0.432
AC	-0.001	-0.003	-0.002	-0.001	0.007	0.018	0.497	0.494
AT	-0.001	0.001	0.000	-0.002	0.000	0.000	0.507	0.490
ATO	-0.001	-0.002	-0.013	0.007	-0.003	-0.004	0.490	0.482
B2M	-0.000	0.018	0.000	0.003	0.013	0.000	0.487	0.477
BETA_d	0.002	-0.002	-0.002	-0.008	0.001	-0.005	0.498	0.489
BETA_m	0.001	-0.001	-0.001	-0.004	0.000	-0.003	0.493	0.500
C2A	0.001	-0.002	0.001	0.002	-0.001	-0.003	0.502	0.493
CF2B	-0.005	-0.002	-0.000	0.006	0.016	-0.001	0.491	0.491
CF2P	-0.005	0.005	-0.003	-0.000	0.009	0.002	0.500	0.481
CTO	-0.001	-0.002	-0.014	0.003	-0.002	0.001	0.490	0.482
D2A	0.001	-0.000	-0.002	-0.000	-0.002	-0.004	0.500	0.497
D2P	-0.001	0.001	0.000	0.000	0.001	-0.000	0.499	0.499
DPI2A	-0.004	-0.009	-0.005	-0.026	0.008	0.046	0.477	0.477
E2P	-0.011	0.002	-0.003	0.004	0.015	0.005	0.480	0.484
FC2Y	0.002	-0.002	0.002	0.002	0.001	-0.003	0.500	0.491
HIGH52	-0.023	-0.020	0.015	0.023	-0.067	0.019	0.426	0.440
INV	-0.006	-0.006	0.002	-0.014	0.011	0.024	0.490	0.478
IdioVol	0.058	-0.007	-0.027	-0.026	0.052	0.030	0.390	0.408
LEV	-0.000	0.002	-0.000	-0.004	0.001	0.000	0.502	0.492
ME	-0.006	-0.005	0.001	-0.007	-0.009	-0.005	0.509	0.466
NI	0.002	-0.003	0.003	-0.008	0.001	0.011	0.497	0.491
NOA	-0.002	0.002	-0.004	-0.012	0.004	0.019	0.488	0.488
OA	-0.001	-0.002	-0.004	0.004	0.007	0.015	0.498	0.497
OL	0.003	-0.003	-0.017	0.008	-0.008	-0.005	0.483	0.478
OP	-0.007	-0.002	-0.006	-0.002	0.013	0.000	0.487	0.488
PCM	-0.001	-0.002	0.002	0.001	0.007	-0.001	0.498	0.493
PM	-0.007	-0.001	0.002	-0.001	0.009	0.002	0.488	0.490
PROF	-0.001	-0.001	-0.006	0.003	0.005	-0.006	0.498	0.489
Q	0.005	-0.047	-0.005	0.003	-0.021	0.004	0.461	0.432
R12_2	-0.016	-0.024	0.002	0.008	-0.024	0.059	0.442	0.455
R12_7	-0.018	-0.027	0.002	0.008	-0.025	0.071	0.447	0.432
R2_1	-0.043	-0.050	0.006	0.009	-0.122	0.003	-0.057	-0.034
R36_13	-0.004	-0.005	-0.000	-0.001	0.006	-0.001	0.493	0.492
R60_13	-0.002	-0.003	0.000	-0.001	0.003	-0.001	0.497	0.494
RNA	-0.005	-0.005	-0.003	0.002	0.008	0.002	0.493	0.489
ROA	-0.007	-0.005	-0.005	0.001	0.007	0.002	0.499	0.477
ROE	-0.009	-0.006	-0.004	0.001	0.012	0.002	0.494	0.476
RVAR	0.022	-0.001	-0.010	-0.006	0.022	0.014	0.484	0.440
S2P	-0.002	0.021	-0.021	0.004	0.007	-0.005	0.478	0.460
SGA2S	0.002	-0.002	0.002	0.002	0.001	-0.003	0.502	0.490
SPREAD	0.058	-0.007	-0.018	-0.020	0.045	0.027	0.368	0.412
SUV	0.001	-0.016	-0.004	-0.053	-0.018	-0.022	0.034	0.037
TURN	0.005	-0.046	-0.021	-0.100	-0.012	-0.047	0.391	0.387
VAR	0.054	-0.009	-0.026	-0.032	0.049	0.024	0.397	0.411

Note: This table shows the the regression coefficients on the cross-sectional factor model and the past time-series information of the global BF-XS model. We report the coefficients on each of the six factors and the past and future values.

Table B.8: Imputation Error For Different Size Filters

estimation	evaluation	aggregate	quarterly	monthly
$< \$ 1$ firms	$< \$ 1$ firms	0.09	0.10	0.05
	$\geq \$ 1$ firms	0.16	0.15	0.17
	all	0.16	0.15	0.17
$\geq \$ 1$ firms	$< \$ 1$ firms	0.26	0.30	0.24
	$\geq \$ 1$ firms	0.14	0.14	0.14
	all	0.14	0.14	0.14
all	$< \$ 1$ firms	0.26	0.30	0.24
	$\geq \$ 1$ firms	0.14	0.14	0.14
	all	0.14	0.14	0.14

Note: This figure shows the imputation RMSE for the global B-XS model across fits and evaluations on firms with filters based on share prices.

Table B.9: Imputation Results with and without Financial Firms

estimation	evaluation	aggregate	quarterly	monthly
financial firms	financial firms	0.14	0.13	0.14
	non financial firms	0.14	0.13	0.14
non financial firms	financial firms	0.14	0.14	0.14
	non financial firms	0.14	0.14	0.14

This figure shows the imputation RMSE for the global B-XS model across fits and evaluations on financial and non-financial firms.

Table B.10: Univariate Long-Short Decile Factors with and without Imputation

	(1) Fully Observed					(2) Obs ≥ 10					(3) Specific Char Observed					(2) + (3)					(2) - (3)				
	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%	mean	stdev	Sharpe	vw %	%
A2ME	3.52	11.2	0.31	0.22	0.41	2.44	9.06	0.27	0.4	0.63	1.66	6.27	0.26	0.92	0.99	1.66	6.27	0.26	0.93	1.0					
AC	1.79	5.22	0.34	0.22	0.41	1.87	5.0	0.37	0.4	0.63	1.97	5.07	0.39	0.7	0.79	1.98	5.05	0.39	0.72	0.82	2.15	13.98	0.15	0.02	0.03
AT	1.41	4.51	0.31	0.22	0.41	1.35	4.34	0.31	0.4	0.63	1.38	4.24	0.32	0.92	0.99	1.38	4.24	0.32	0.93	1.0					
ATO	2.14	5.29	0.4	0.22	0.41	2.13	5.21	0.41	0.4	0.63	2.25	5.31	0.42	0.89	0.99	2.25	5.31	0.42	0.89	0.99	1.88	18.31	0.1	0.01	0.0
B2M	2.38	7.74	0.31	0.22	0.41	2.53	7.28	0.35	0.4	0.63	2.3	5.71	0.4	0.93	0.99	2.3	5.71	0.4	0.93	1.0	1.69	25.72	0.07	0.0	0.0
BETA_d	2.11	8.7	0.24	0.22	0.41	2.25	8.5	0.27	0.4	0.63	2.18	8.03	0.27	0.78	0.95	2.18	8.03	0.27	0.79	0.95					
BETA_m	2.38	9.21	0.26	0.22	0.41	2.61	8.8	0.3	0.4	0.63	2.65	8.53	0.31	0.85	0.96	2.65	8.53	0.31	0.86	0.96					
C2A	2.09	6.19	0.34	0.22	0.41	2.04	5.98	0.34	0.4	0.63	2.21	6.09	0.36	0.91	0.99	2.22	6.1	0.36	0.91	0.99	4.87	15.3	0.32	0.0	0.0
CF2B	1.79	5.25	0.34	0.22	0.41	1.89	5.02	0.38	0.4	0.63	1.87	4.71	0.4	0.92	0.99	1.88	4.71	0.4	0.92	0.99	1.04	20.62	0.05	0.0	0.0
CF2P	1.89	6.53	0.29	0.22	0.41	2.09	6.65	0.31	0.4	0.63	1.52	3.98	0.38	0.93	0.99	1.52	3.98	0.38	0.93	1.0					
CTO	1.87	4.97	0.38	0.22	0.41	1.95	4.93	0.4	0.4	0.63	2.07	4.96	0.42	0.89	0.99	2.07	4.96	0.42	0.9	0.99	1.2	18.15	0.07	0.01	0.0
D2A	1.67	5.81	0.29	0.22	0.41	1.79	5.25	0.34	0.4	0.63	1.86	5.34	0.35	0.78	0.91	1.83	5.21	0.35	0.83	0.94	1.95	7.2	0.27	0.04	0.03
D2P	2.3	5.1	0.45	0.22	0.41	2.32	4.88	0.48	0.4	0.63	2.12	4.17	0.51	1.0	1.0	2.12	4.17	0.51	1.0	1.0					
DPI2A	1.76	6.17	0.29	0.22	0.41	1.74	5.9	0.3	0.4	0.63	1.85	5.92	0.31	0.5	0.61	1.84	5.91	0.31	0.61	0.77	1.58	9.68	0.16	0.11	0.16
E2P	2.27	6.39	0.36	0.22	0.41	2.23	5.97	0.37	0.4	0.63	2.08	5.23	0.4	0.93	0.99	2.08	5.23	0.4	0.93	1.0	3.25	25.49	0.13	0.0	0.0
FC2Y	1.73	4.65	0.37	0.22	0.41	1.76	4.58	0.38	0.4	0.63	1.93	4.99	0.39	0.75	0.8	1.95	4.99	0.39	0.78	0.83	2.73	7.24	0.38	0.03	0.03
HIGH52	1.19	4.37	0.27	0.22	0.41	1.23	4.23	0.29	0.4	0.63	1.27	3.94	0.32	0.66	0.9	1.27	3.94	0.32	0.66	0.9					
INV	1.78	6.27	0.28	0.22	0.41	1.91	6.07	0.32	0.4	0.63	1.96	6.01	0.33	0.83	0.97	1.96	5.98	0.33	0.84	0.98	0.85	15.38	0.06	0.01	0.0
IdioVol	2.78	8.45	0.33	0.22	0.41	2.91	8.47	0.34	0.4	0.63	3.37	8.98	0.37	1.0	1.0	3.37	8.98	0.38	1.0	1.0					
LEV	2.07	6.06	0.34	0.22	0.41	1.91	6.0	0.32	0.4	0.63	1.75	5.74	0.31	0.86	0.95	1.75	5.72	0.31	0.88	0.97	2.15	12.19	0.18	0.02	0.02
ME	1.48	4.59	0.32	0.22	0.41	1.48	4.46	0.33	0.4	0.63	1.43	4.3	0.33	1.0	1.0	1.43	4.3	0.33	1.0	1.0					
NI	1.91	6.26	0.31	0.22	0.41	1.87	5.99	0.31	0.4	0.63	1.84	5.57	0.33	0.83	0.97	1.84	5.58	0.33	0.84	0.97	2.07	10.69	0.19	0.01	0.0
NOA	1.64	5.97	0.27	0.22	0.41	1.65	5.75	0.29	0.4	0.63	1.64	5.04	0.33	0.89	0.99	1.64	5.04	0.33	0.89	0.99	1.25	30.89	0.04	0.01	0.0
OA	1.8	5.29	0.34	0.22	0.41	1.78	5.09	0.35	0.4	0.63	1.94	5.12	0.38	0.75	0.8	1.93	5.13	0.38	0.76	0.83	1.11	23.06	0.05	0.01	0.03
OL	1.81	4.98	0.36	0.22	0.41	1.94	4.99	0.39	0.4	0.63	2.03	4.92	0.41	0.9	0.99	2.04	4.92	0.42	0.91	0.99	2.63	15.15	0.17	0.01	0.0
OP	1.88	5.28	0.36	0.22	0.41	1.87	5.12	0.36	0.4	0.63	1.97	5.04	0.39	0.83	0.93	1.97	5.04	0.39	0.84	0.93	2.19	23.16	0.09	0.0	0.0
PCM	1.63	4.98	0.33	0.22	0.41	1.61	4.76	0.34	0.4	0.63	1.7	4.81	0.35	0.89	0.99	1.7	4.81	0.35	0.91	0.99	3.96	13.19	0.3	0.01	0.0
PM	1.79	5.71	0.31	0.22	0.41	1.69	5.27	0.32	0.4	0.63	1.55	4.48	0.35	0.9	0.98	1.55	4.48	0.35	0.91	0.99	2.74	19.2	0.14	0.01	0.0
PROF	1.87	4.83	0.39	0.22	0.41	1.92	4.73	0.41	0.4	0.63	1.99	4.79	0.41	0.9	0.99	1.99	4.79	0.42	0.9	0.99	2.06	18.77	0.11	0.0	0.0
Q	1.6	5.13	0.31	0.22	0.41	1.65	4.96	0.33	0.4	0.63	1.78	5.23	0.34	0.92	0.99	1.79	5.23	0.34	0.93	1.0					
R12_2	2.34	6.46	0.36	0.22	0.41	2.41	6.24	0.39	0.4	0.63	2.59	6.34	0.41	0.92	0.98	2.59	6.34	0.41	0.92	0.98					
R12_7	2.26	6.61	0.34	0.22	0.41	2.39	6.48	0.37	0.4	0.63	2.65	6.53	0.41	0.92	0.98	2.66	6.53	0.41	0.92	0.98					
R2_1	1.85	6.17	0.3	0.22	0.41	1.92	5.93	0.32	0.4	0.63	2.07	5.94	0.35	1.0	1.0	2.07	5.94	0.35	1.0	1.0					
R36_13	1.94	6.5	0.3	0.22	0.41	2.06	6.36	0.32	0.4	0.63	2.02	6.05	0.33	0.77	0.95	2.01	6.05	0.33	0.78	0.95	0.49	14.98	0.03	0.01	0.0
R60_13	1.76	6.08	0.29	0.22	0.41	1.87	6.12	0.31	0.4	0.63	1.89	5.87	0.32	0.64	0.91	1.89	5.86	0.32	0.65	0.91	1.77	14.03	0.13	0.01	0.0
RNA	1.85	5.13	0.36	0.22	0.41	1.79	4.88	0.37	0.4	0.63	1.89	4.92	0.39	0.87	0.98	1.89	4.91	0.38	0.88	0.98	2.99	16.75	0.18	0.01	0.0
ROA	1.66	5.07	0.33	0.22	0.41	1.64	4.79	0.34	0.4	0.63	1.75	4.91	0.36	0.87	0.98	1.76	4.9	0.36	0.87	0.98					
ROE	1.7	5.15	0.33	0.22	0.41	1.63	4.76	0.34	0.4	0.63	1.75	4.84	0.36	0.87	0.98	1.76	4.84	0.36	0.88	0.98					
RVAR	2.83	9.11	0.31	0.22	0.41	2.98	8.95	0.33	0.4	0.63	3.5	9.25	0.38	1.0	1.0	3.5	9.25	0.38	1.0	1.0					
S2P	2.81	6.95	0.4	0.22	0.41	2.51	6.41	0.39	0.4	0.63	2.69	6.39	0.42	0.92	1.0	2.69	6.38	0.42	0.93	1.0	-0.06	26.34	-0.0	0.0	0.0
SGA2S	1.78	5.1	0.35	0.22	0.41	1.78	4.85	0.37	0.4	0.63	1.97	5.37	0.37	0.75	0.8	1.99	5.31	0.37	0.78	0.83	2.65	7.38	0.36	0.03	0.03
SPREAD	3.17	10.44	0.3	0.22	0.41	3.32	10.19	0.33	0.4	0.63	3.68	9.59	0.38	1.0	1.0	3.68	9.59	0.38	1.0	1.0					
SUV	1.78	5.24	0.34	0.22	0.41	1.8	5.13	0.35	0.4	0.63	1.96	5.0	0.39	0.93	1.0	1.96	5.0	0.39	0.94	1.0					
TURN	2.49	7.96	0.31	0.22	0.41	2.61	7.75	0.34	0.4	0.63	2.74	7.63	0.36	0.94	1.0	2.74	7.63	0.36	0.94	1.0					
VAR	2.77	9.32	0.3	0.22	0.41	2.92	9.2	0.32	0.4	0.63	3.35	9.39	0.36	1.0	1.0	3.35	9.39	0.36	1.0	1.0					

This table shows the mean, volatility, Sharpe ratio, value percentage and percentage of total stocks used to construct long-short decile factors. The quantile cutoffs are based on deciles of fully present NYSE data. We consider (1) fully observed data, (2) only ≥ 10 observed characteristics, where the rest is imputed, (3) specific characteristic present, other characteristics missing, (4) union of 3 and 2 (in 3 or 2), (5) not in (3) but in (2). Mean and standard deviations are reported as percentages. We use the global B-XS model to impute missing values that have prior observations available and the global XS model for the case without prior observations.

Table B.11: Firm Characteristics

Acronym	Name	Definition	Reference	Freq
A2ME	Assets to market cap	Total assets (AT) over market capitalization (PRC*SHROUT) as of current month	Bhandari (1988)	Q
AC	Accrual	Change in operating working capital per split-adjusted share from the fiscal year end t-2 to t-1 divided by book equity (defined in B2M) per share in t-1. Operating working capital per split-adjusted share is defined as current assets (ACTQ) minus cash and short-term investments (CHEQ) minus current liabilities (LCTQ) minus debt in current liabilities (DLCQ) minus income taxes payable (TXPQ).	Sloan (1996)	Q
AT	Total Assets	Total Assets (ATQ)	Gandhi and Lustig (2015)	Q
ATO	Net sales over lagged net operating assets	Net sales (SALEQ) over lagged net operating assets. Net operating assets are the difference between operating assets and operating liabilities (defined in NOA)	Soliman (2008)	Q
B2M	Book to Market Ratio	Book equity is shareholder equity (SH) plus deferred taxes and investment tax credit (TXDITCQ), minus preferred stock (PSTKQ). SH is shareholdersâ□□ equity (SEQQ). If missing, SH is the sum of common equity (CEQQ) and preferred stock (PSQ). If missing, SH is the difference between total assets (ATQ) and total liabilities (LTQ). The market value of equity (PRC*SHROUT) is as of the current month.	Fama and French (1992)	Q
Beta_d	CAPM Beta	Product of correlations between the excess return of stock i and the market excess return and the ratio of volatilities. We calculate volatilities from the standard deviations of daily log excess returns over a one-year horizon requiring at least 120 observations. We estimate correlations using overlapping three-day log excess returns over a five-year period requiring at least 750 non-missing observations.	Frazzini and Pedersen (2014)	M
Beta_m	Market Beta	Coefficient of the market excess return from the regression on excess returns in the past 60 months (24 months minimum)	Fama & MacBeth (1973)	M
C2A	Ratio of cash and short-term investments to total assets	Ratio of cash and short-term investments (CHEQ) to total assets (ATQ)	Palazzo (2012)	Q
CF2B	Free Cash Flow to Book Value	Cash flow to book value of equity is the ratio of net income (NIQ), depreciation and amortization (DPQ), less change in working capital (WCAPCH), and capital expenditure (CAPX) over the book-value of equity (defined in B2M)	Hou et al. (2011)	Q
CF2P	Cashflow to price	Cashflow over market capitalization (PRC*SHROUT) as of currrent month. Cashflow is defined as income before extraordinary items (IBQ) plus depreciation and amortization (DPQ) plus deferred taxes (TXDBQ).	Desai, Rajgopal & Venkatachalam (2004)	Q
CTO	Capital turnover	Ratio of net sales (SALEQ) to lagged total assets (ATQ)	Haugen and Baker (1996)	Q
D2A	Capital intensity	Ratio of depreciation and amortization (DPQ) to total assets (ATQ)	Gorodnichenko and Weber (2016)	Q
D2P	Dividend Yield	Total dividends (DIVAMT) paid from July of t-1 to June of t per dollar of equity (ME) in June of t	Litzenberger and Ramaswamy (1979)	M
DPI2A	Change in property, plants, and equipment	Changes in property, plants, and equipment (PPEGTQ) and inventory (INVTQ) over lagged total assets (ATQ)	Lyandres, Sun, and Zhang (2008)	Q
E2P	Earnings to price	The earnings used in months (t, t+1, t+2) are the earning from the quarter reported at time t (IBQ). P (actually ME) is price times shares outstanding at the end of current month.	Basu (1983)	Q
FC2Y	Fixed costs to sales	Ratio of selling, general, and administrative expenses (XSGAQ), research and development expenses (XRDQ), and advertising expenses (XADQ) to net sales (SALEQ)	D'□□Acunto, Liu, Pflueger, and Weber (2016)	Y
HIGH52	Closeness to past year high	The ratio of stock price at the end of the current calendar month and the highest daily price in the past year	George and Hwang (2004)	M
IdioVol	Idiosyncratic volatility	"Standard deviation of the residuals from a regression of excess returns on the Fama and French three-factor model"	Ang, Hodrick, Xing, and Zhang (2006)	M
INV	Investment	Change in total assets (ATQ) from the fiscal quarter ending in month t-12 to the fiscal quarter ending in t, divided by t-12 total assets	Cooper, Gulen, and Schill (2008)	Q
LEV	Leverage	Ratio of long-term debt (DLTTQ) and debt in current liabilities (DLCQ) to the sum of long-term debt, debt in current liabilities, and stockholders'□□ equity (SEQQ)	Lewellen (2015)	Q
ME	Size	Total market capitalization at the end of the current month defined as price times shares outstanding	Fama and French (1992)	M
LT_Rev	Long-term reversal	Cumulative return from 60 months before the return prediction to 13 months before	Jegadeesh and Titman (2001)	M
TURN	Turnover	Turnover is last month'□□s volume (VOL) over shares outstanding (SHROUT)	Datar, Naik, and Radcliffe (1998)	M
NI	Net Share Issues	The change in the natural log of split-adjusted shares outstanding (CSHO*AJEX) from the fiscal yearend in t-2 to the fiscal yearend in t-1	Pontiff and Woodgate (2008)	M
NOA	Net operating assets	Difference between operating assets minus operating liabilities scaled by lagged total assets (ATQ). Operating assets are total assets (ATQ) minus cash and short-term investments (CHEQ), minus investment and other advances (IVAOQ). Operating liabilities are total assets (ATQ), minus debt in current liabilities (DLCQ), minus long-term debt (DLTTQ), minus minority interest (MIBQ), minus preferred stock (PSTKQ), minus common equity (CEQQ).	Hirshleifer, Hou, Teoh, and Zhang (2004)	Q
OA	Operating accruals	Changes in non-cash working capital minus depreciation (DPQ) scaled by lagged total assets (ATQ). Non-cash working capital is defined in Accrual (AC)	Sloan (1996)	Q
OL	Operating leverage	Sum of cost of goods sold (COGSQ) and selling, general, and administrative expenses (XSGAQ) over total assets (ATQ)	Novy-Marx (2011)	Q
OP	Operating profitability	Annual revenues (REVTQ) minus cost of goods sold (COGSQ), interest expense (TIEQ), and selling, general, and administrative expenses (XSGAQ) divided by book equity (defined in B2M)	Fama and French (2015)	
PCM	Price to cost margin	Difference between net sales (SALEQ) and costs of goods sold (COGSQ) divided by net sales (SALEQ)	Bustamante and Donangelo (2016)	Q
PM	Profit margin	Operating income after depreciation (OIADPQ) over net sales (SALEQ)	Soliman (2008)	Q

Note: Continued on next page.

Acronym	Name	Definition	Reference	Freq
PROF	Profitability	Gross profit (GP) divided by the book value of equity (defined in B2M)	Ball, Gerakos, Linnainmaa, and Nikolaev (2015)	Y
Q	Tobin's Q	"Tobinâ□□s Q is total assets (ATQ), the market value of equity (SHROUT times PRC) minus cash and short-term investments (CEQQ), minus deferred taxes (TXDBQ) scaled by total assets (ATQ)"	Kaldor (1966)	Q
R12_2	Momentum	To be included in a portfolio for month t (formed at the end of month t-1), a stock must have a price for the end of month t-13 and a good return for t-2. In addition, any missing returns from t-12 to t-3 must be -99.0, CRSP's code for a missing price. Each included stock also must have ME for the end of month t-1.	Fama and French (1996)	M
R12_7	Intermediate momentum	Cumulative return from 12 months before the return prediction to seven months before	Novy-Marx (2012)	M
R36_13	Long-term reversal	Cumulative return from 36 months before the return prediction to 13 months before	De Bondt and Thaler (1985)	M
R2_1	Short-term reversal	Lagged one-month return	Jegadeesh and Titman (1993)	M
RNA	Return on net operating assets	Ratio of operating income after depreciation (OIADPQ) to lagged net operating assets. Net operating assets are the difference between operating assets minus operating liabilities. (defined in NOA)	Soliman (2008)	Q
ROA	Return on assets	Income before extraordinary items (IBQ) to lagged total assets (ATQ)	Balakrishnan, Bartov, and Faurel (2010)	Q
ROE	Return on equity	Income before extraordinary items (IBQ) to lagged book-value of equity (defined in B2M)	Haugen and Baker (1996)	Q
RVAR	Residual Variance	Variance of the residuals from a regression of excess returns in the past two months on the CAPM model	Ang, Hodrick, Xing, and Zhang (2006)	M
S2P	Sales to price	Ratio of net sales (SALEQ) to the market capitalization (ME)	Lewellen (2015)	Q
SGA2S	Selling, general and administrative expenses to sales	Ratio of selling, general and administrative expenses (XSGAQ) to net sales (SALEQ)	Freyberger, Neuhierl, Weber (2017)	Q
SPREAD	Bid-ask spread	The average daily bid-ask spread in the current month	Chung and Zhang (2014)	M
SUV	Standard unexplained volume	Difference between actual volume and predicted volume in the current month. Predicted volume comes from a regression of daily volume on a constant and the absolute values of positive and negative returns. Unexplained volume is standardized by the standard deviation of the residuals from the regression	Garfinkel (2009)	M
VAR	Variance	Variance of daily returns in the past 60 days	Ang, Hodrick, Xing, and Zhang (2006)	M

Note: This table summarizes the information about the 45 characteristics. We report the abbreviation, name, definition, reference and updating frequency.

Table B.12: CRSP and Compustat dependencies in the construction of characteristics

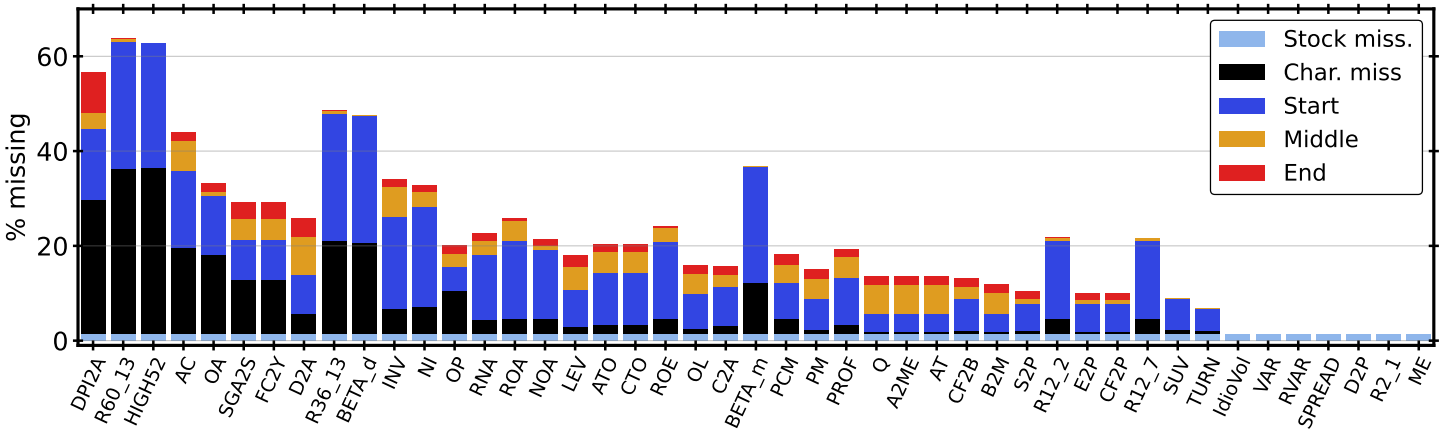
Characteristic	CRSP Dependencies		Compustat Dependencies	
	Monthly	Daily	Quarterly	Yearly
A2ME	prc, shrout		atq	
AC			actq, atq, ceqq, cheq, dlcq, lctq, ltq, pstkq, pstkq, seqq, txditcq, txpq	
AT			atq	
ATO			atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, pstkq, saleq	
B2M	prc, shrout		atq, ceqq, ltq, pstkq, pstkq, seqq, txditcq	
BETA_d	ret	ret		
BETA_m	ret			
C2A			atq, cheq	
CF2B			atq, capxy, ceqq, dpq, ltq, niq, pstkq, pstkq, seqq, txditcq, wcapchy	
CF2P	prc, shrout		dpq, ibq, txdbq	
CTO			atq, saleq	
D2A			atq, dpq	
D2P	divamt, prc, shrout			
DPI2A			atq, invtq, ppegtq	
E2P	prc, shrout		ibq	
FC2Y			saleq, xrdq, xsgaq	xad
HIGH52	prc	prc		
INV			atq	
IdioVol	ret	ret		
LEV			dlcq, dlttq, seqq	
ME	prc, shrout			
NI			ajexq, cshoq	
NOA			atq, atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, pstkq	
OA			actq, atq, cheq, dlcq, dpq, lctq, txpq	
OL			atq, cogsq, xsgaq	
OP			atq, ceqq, cogsq, ltq, pstkq, pstkq, revtq, seqq, tieq, txditcq, xsgaq	
PCM			cogsq, saleq	
PM			oiadpq, saleq	
PROF			atq, ceqq, ltq, pstkq, pstkq, seqq, txditcq	gp
Q	prc, shrout		atq, ceqq, txdbq	
R12_2	prc, prc, ret, shrout			
R12_7	ret			
R2_1	ret			
R36_13	ret			
R60_13	ret			
RNA			atq, atq, ceqq, cheq, dlcq, dlttq, ivaoq, mibq, oiadpq, pstkq	
ROA			atq, ibq	
ROE			atq, ceqq, ibq, ltq, pstkq, pstkq, seqq, txditcq	
RVAR	ret	ret		
S2P	prc, shrout		saleq	
SGA2S			saleq, xsgaq	
SPREAD	ret	askhi, bidlo		
SUV	ret	ret, vol		
TURN	shrout, vol			
VAR	ret	ret		

This table shows the CRSP and Compustat dependencies in the construction of characteristics. We report for each characteristic, which CRSP and Compustat variables are used in the construction and the corresponding updating frequency.

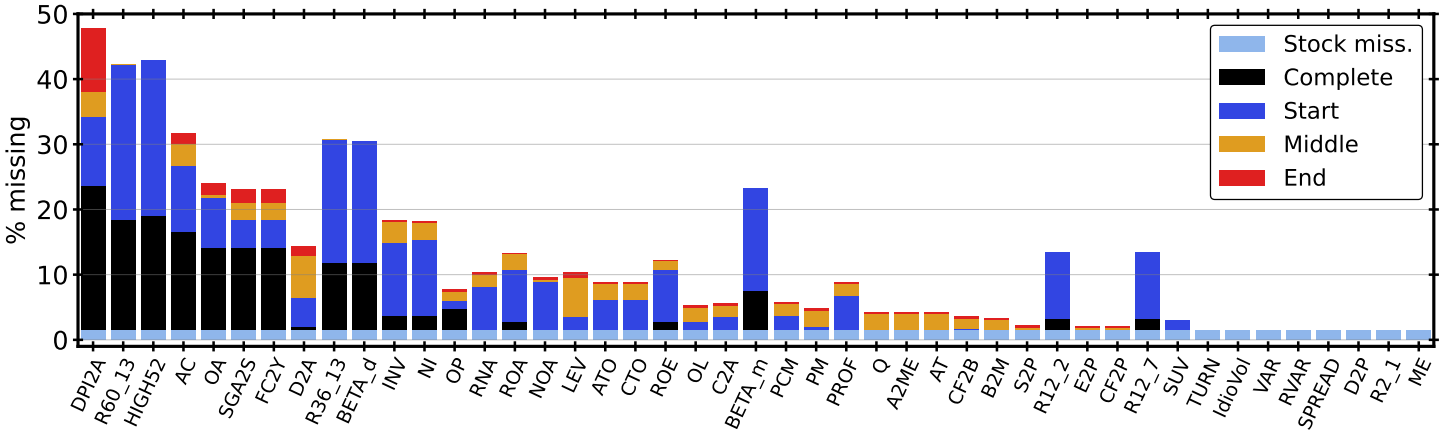
Appendix C. Figures

Figure C.1: Missing Observations by Characteristic Pooled by Stocks

(a) Pooled Mean across Stocks (Equally-weighted)

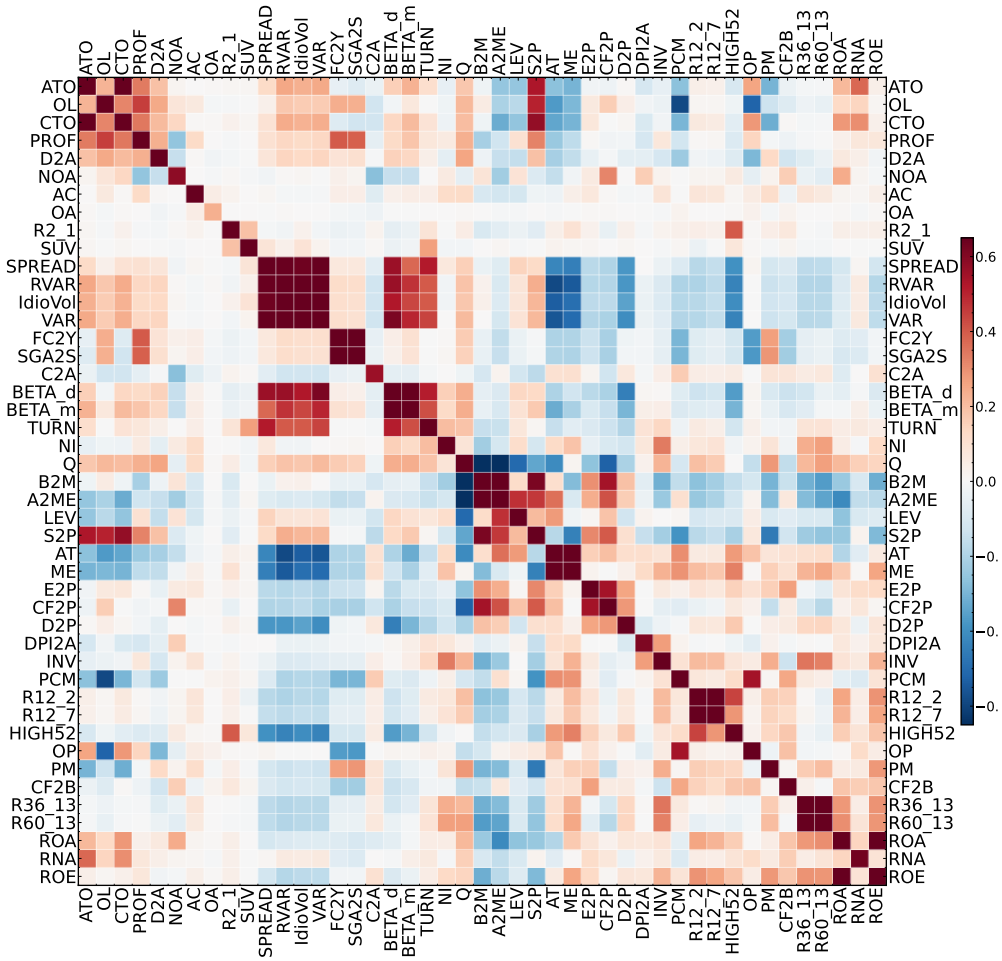


(b) Pooled Mean across Stocks (Value-weighted)



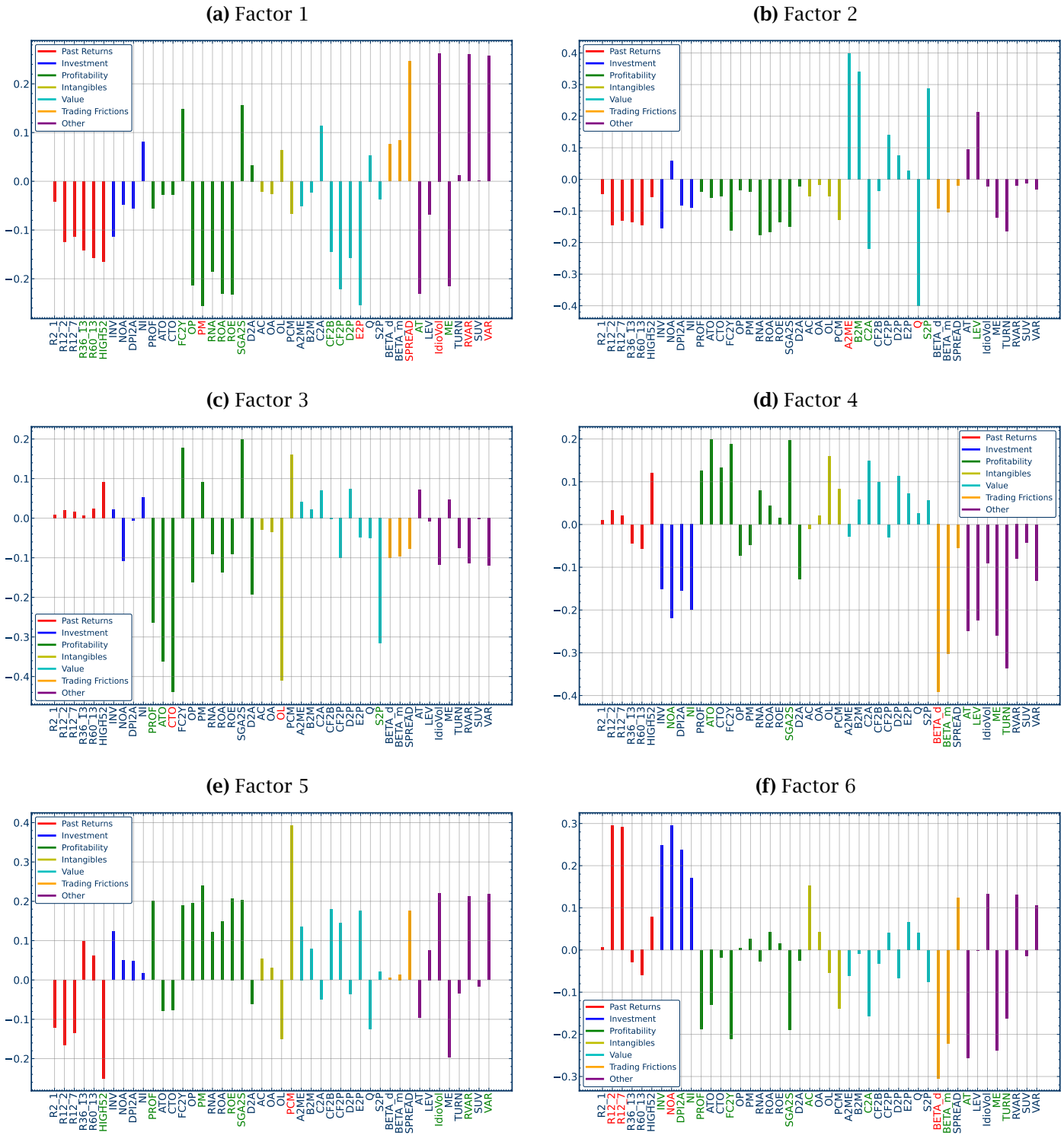
Note: This figure shows the average percentage of missing observations for each characteristic. The means are pooled by stocks, which are equally weighted in the top panel and value-weighted in the bottom panel. We decompose the missing values in those missing at the start (no previous observations), the middle (some previous and future observations), the end (no further observations) and completely missing.

Figure C.2: Heatmap of Pairwise Correlation from 1967–1976



Note: This figure shows the pairwise correlations across time and stocks for each characteristic. The time period is the early sample from 1967-1976.

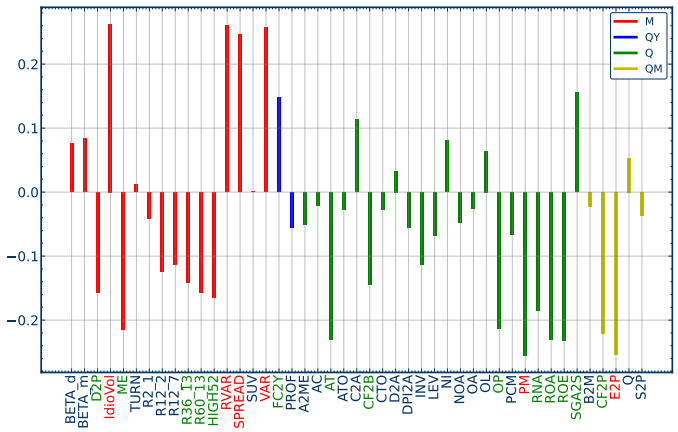
Figure C.3: Composition of Latent Factors by Characteristic Categories



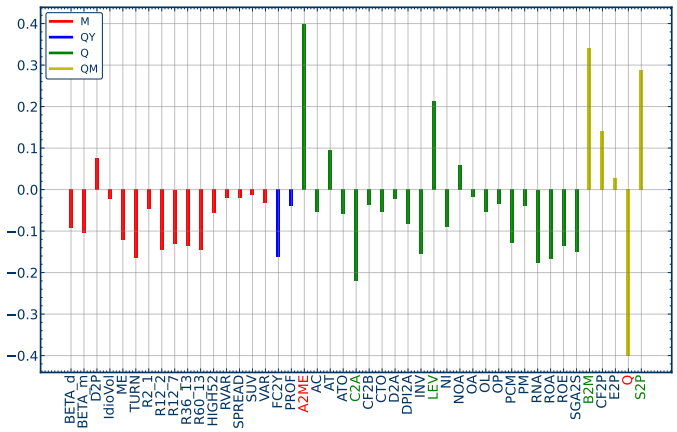
Note: This figure shows the global factor loadings on the characteristics for the first 6 factors. The loadings are colored by the category to which the characteristic belongs.

Figure C.4: Composition of Latent Factors Grouped by Frequencies

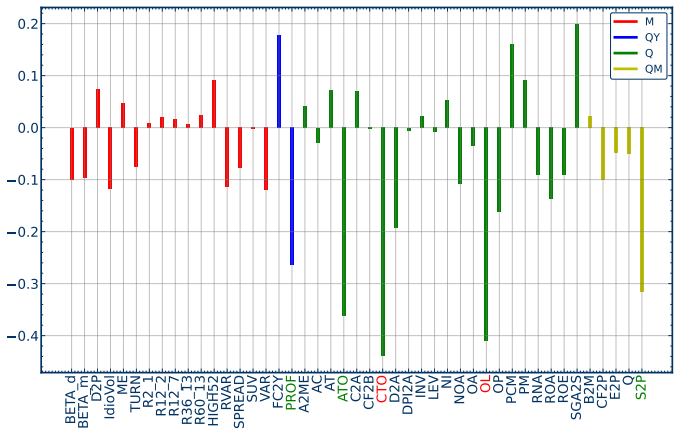
(a) Factor 1



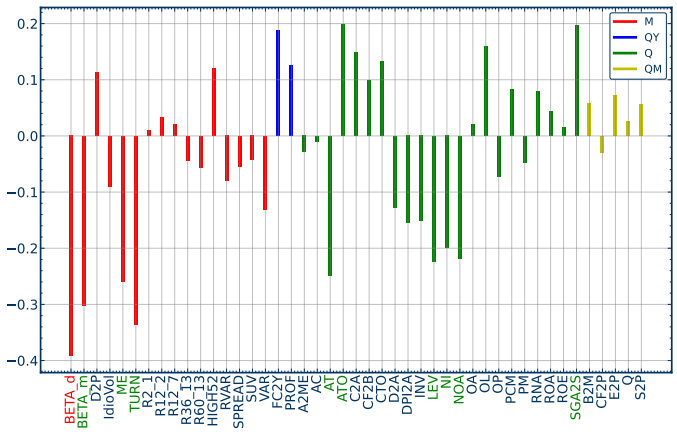
(b) Factor 2



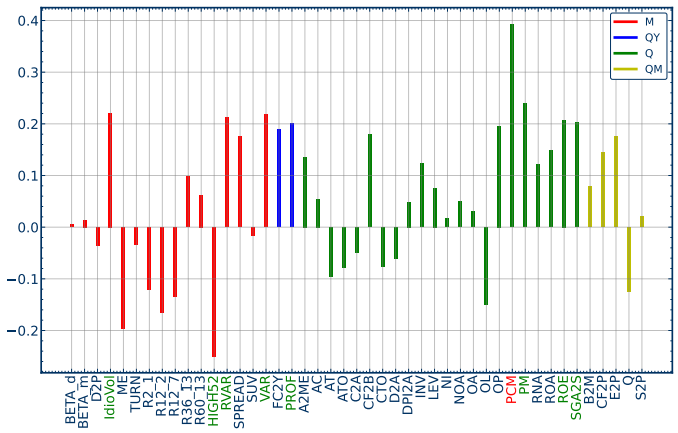
(c) Factor 3



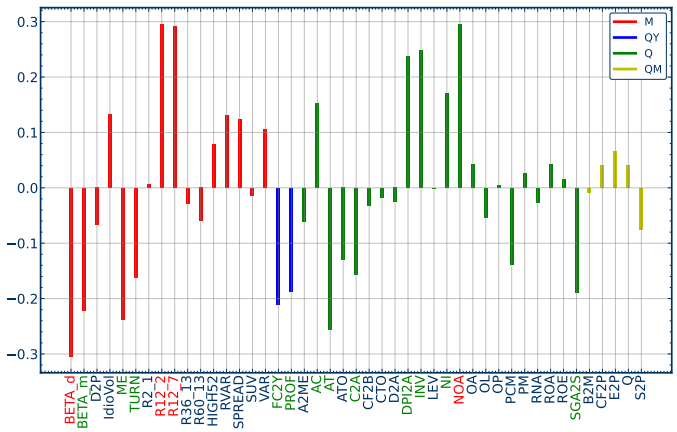
(d) Factor 4



(e) Factor 5

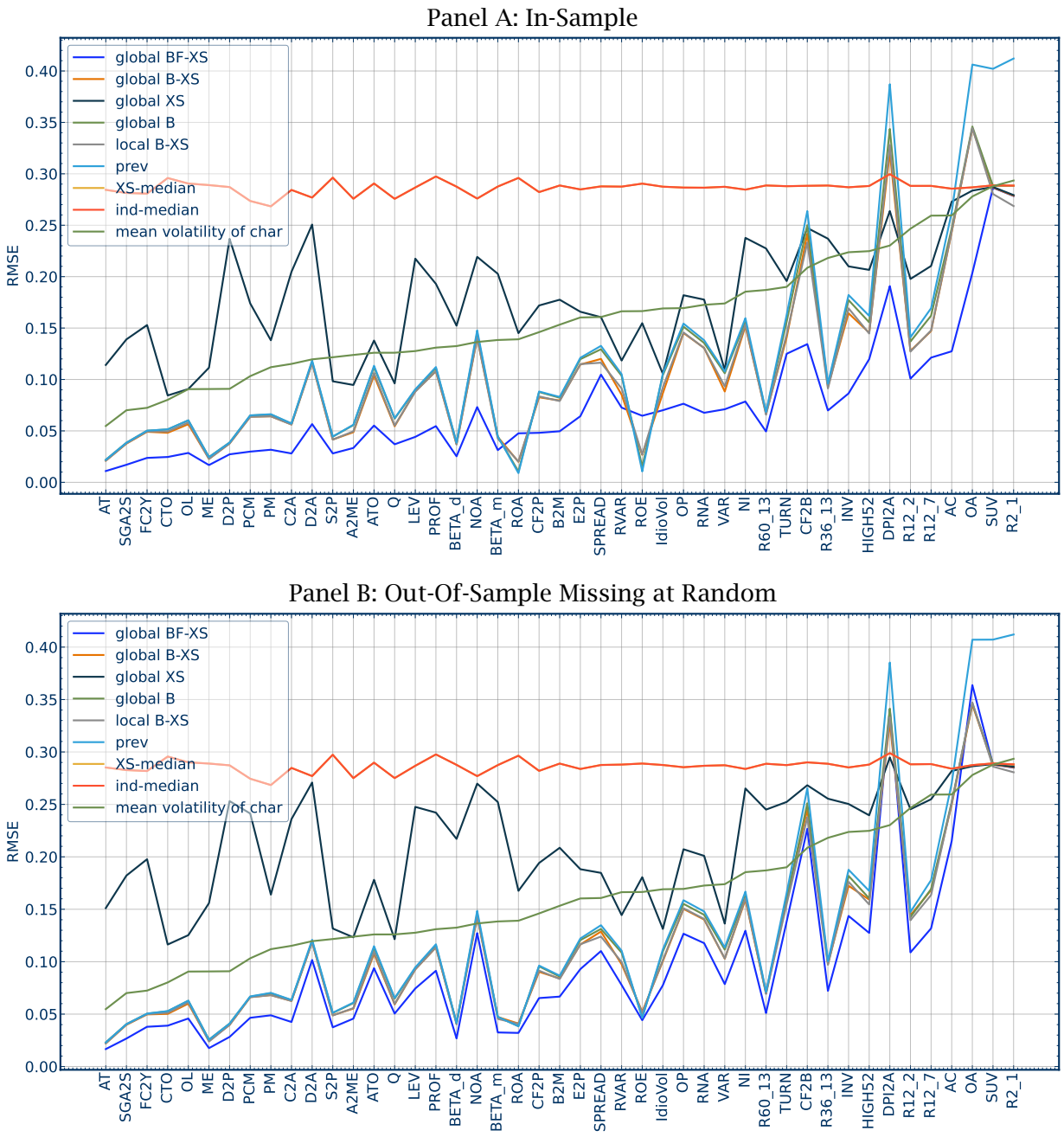


(f) Factor 6



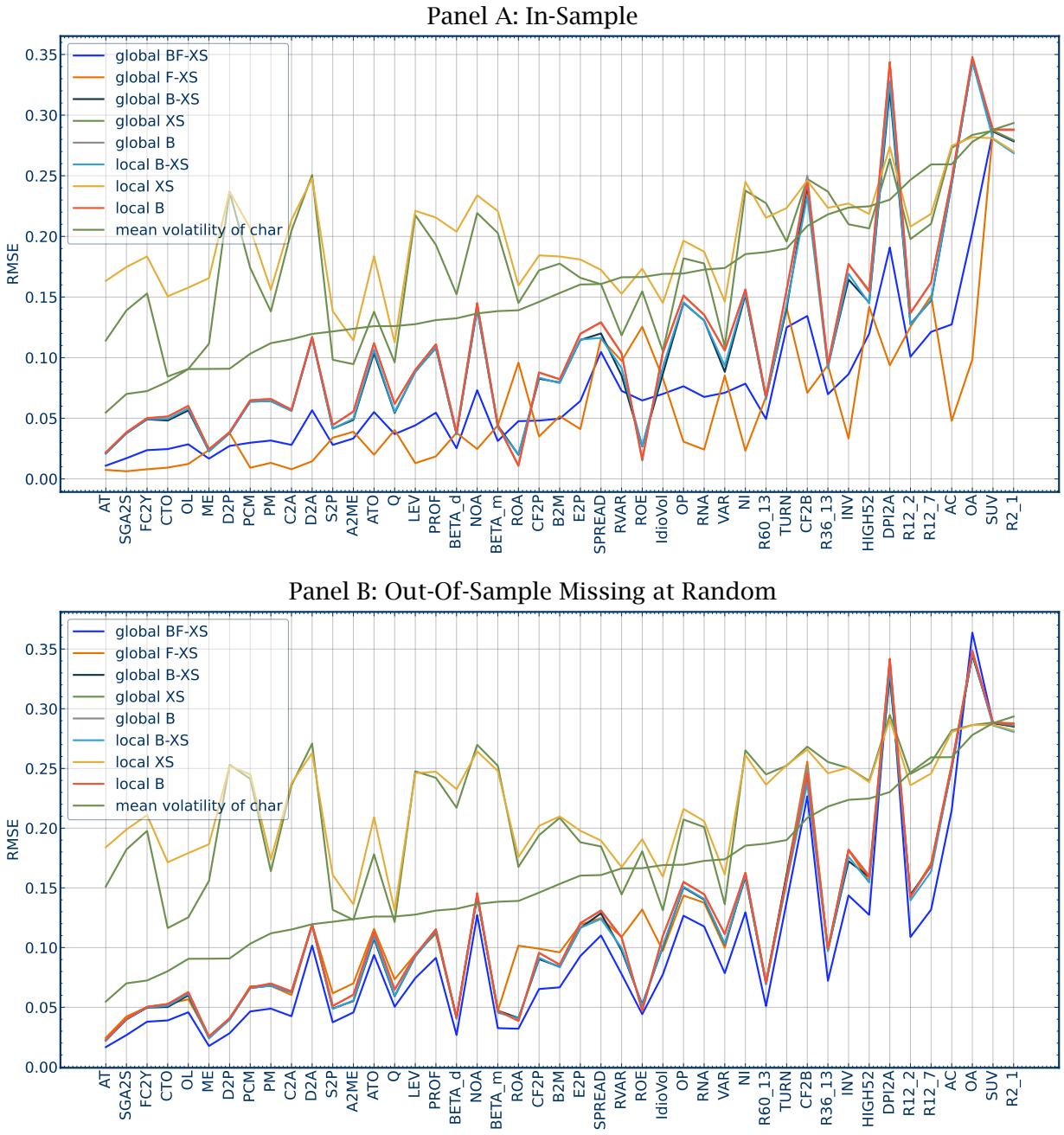
Note: This figure shows the global factor loadings on the characteristics for the first 6 factors. The loadings are colored by the frequency at which the characteristic is updated.

Figure C.5: Imputation Error For Individual Characteristics



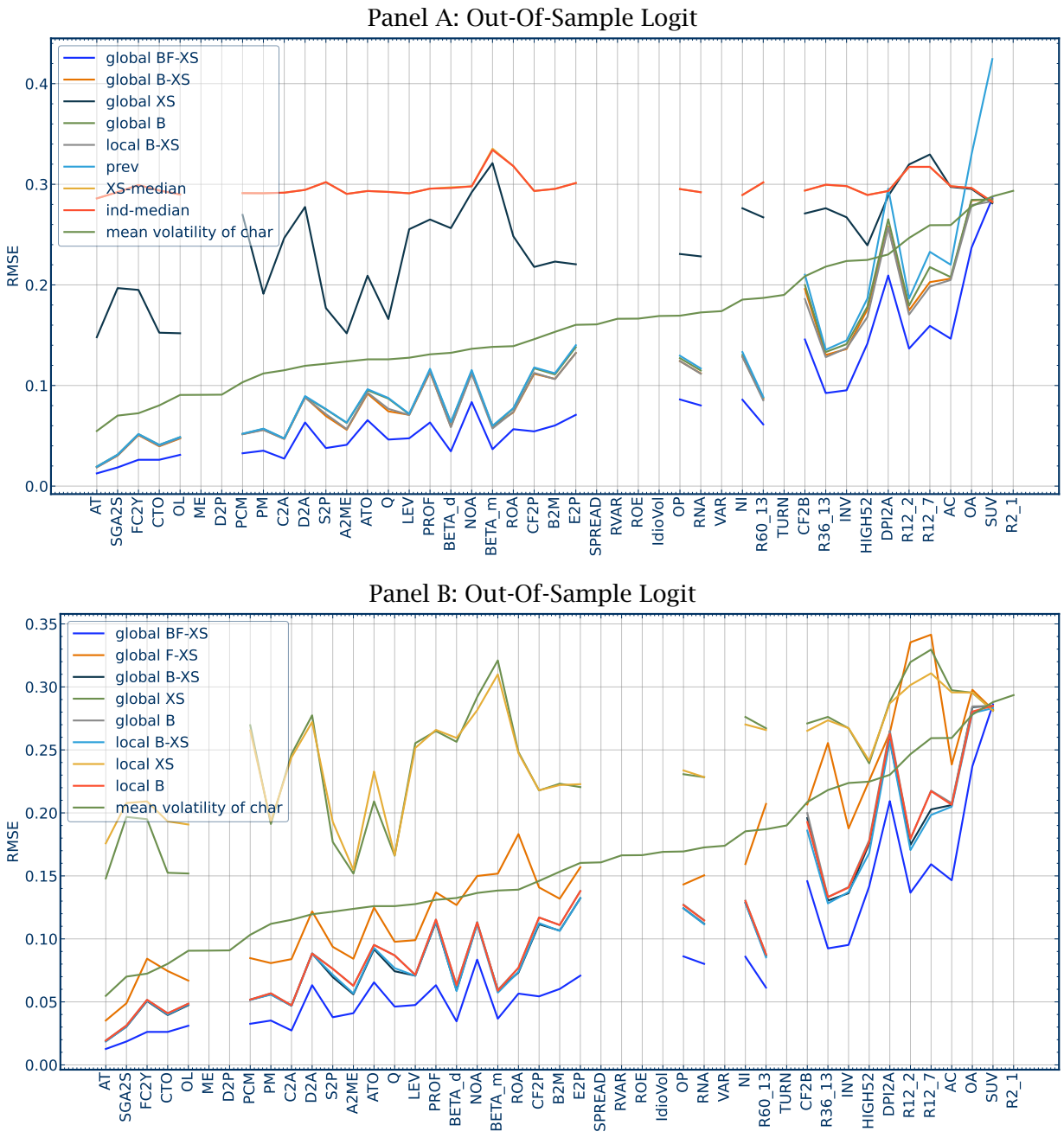
Note: This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data missing at random.

Figure C.6: Global and Local Imputation For Individual Characteristics



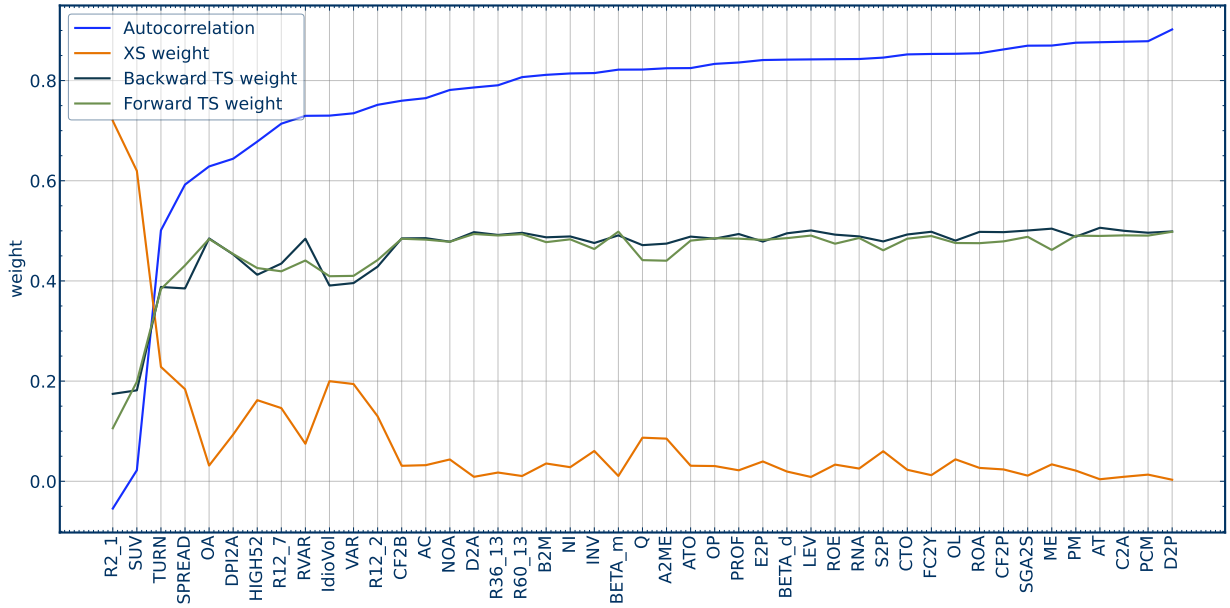
Note: This figure shows the imputation RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. We report the imputation error in-sample evaluated over all observed data, and out-of-sample for masked characteristics from the fully present subset of the data. For the out-of-sample analysis we mask 10% of the data missing at random.

Figure C.7: Imputation Error For Individual Characteristics



Note: This figure shows the imputation out-of-sample RMSE by imputation method across individual characteristics. The characteristics are sorted in ascending order based on the time-series standard deviation of characteristics. The logit masking is based on the logistic regression model with all covariates and fixed effects as estimated in Table 2. The lines have empty entries for characteristics that are always observed.

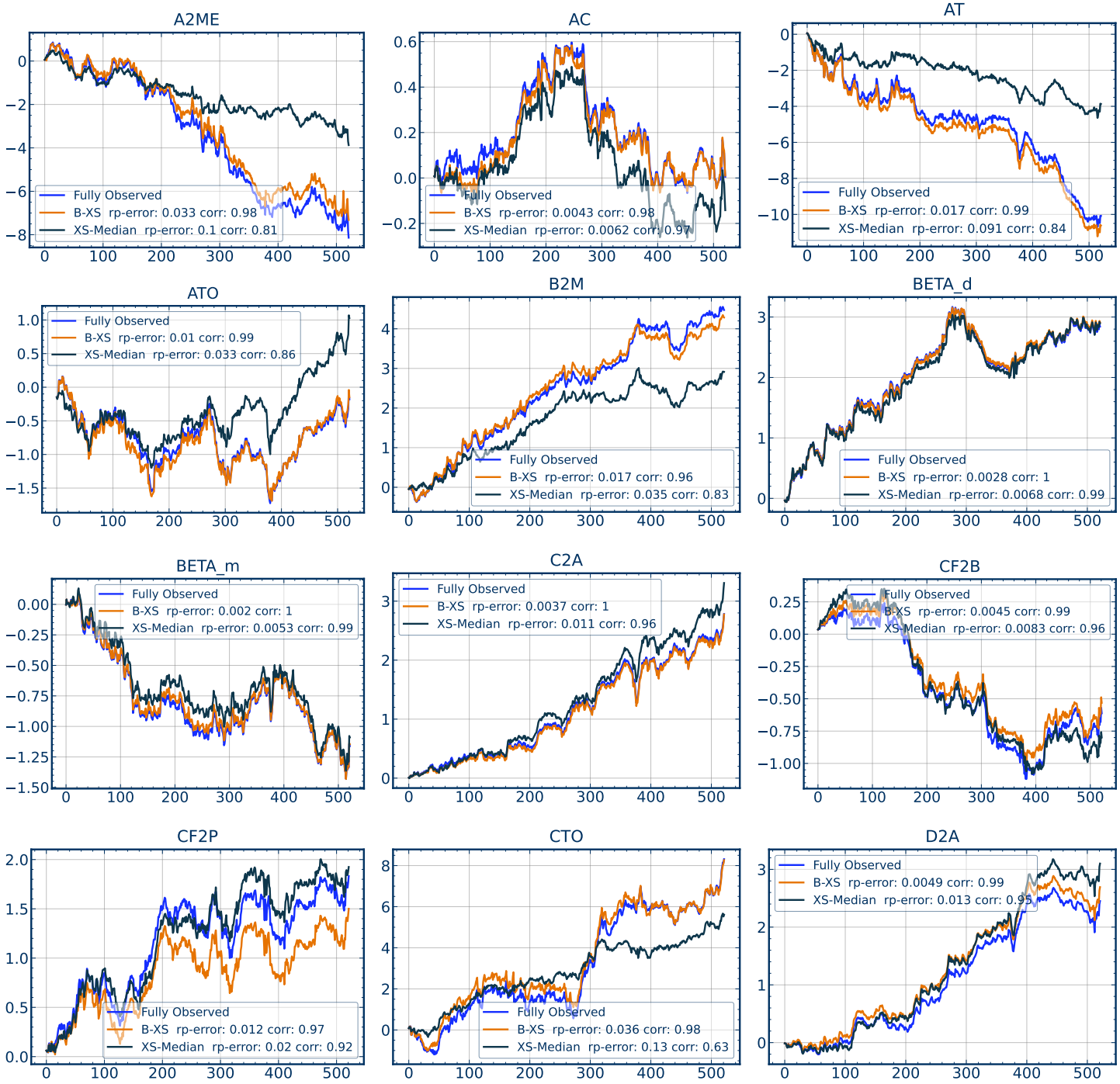
Figure C.8: Information used for Imputation for BF-XS model



Note: This figure shows the regression coefficients on the cross-sectional factor model and the time-series information. The XS weight denotes the sum of absolute values of the coefficients on the cross-sectional factor model. The characteristics are sorted in ascending order based on their autocorrelation.

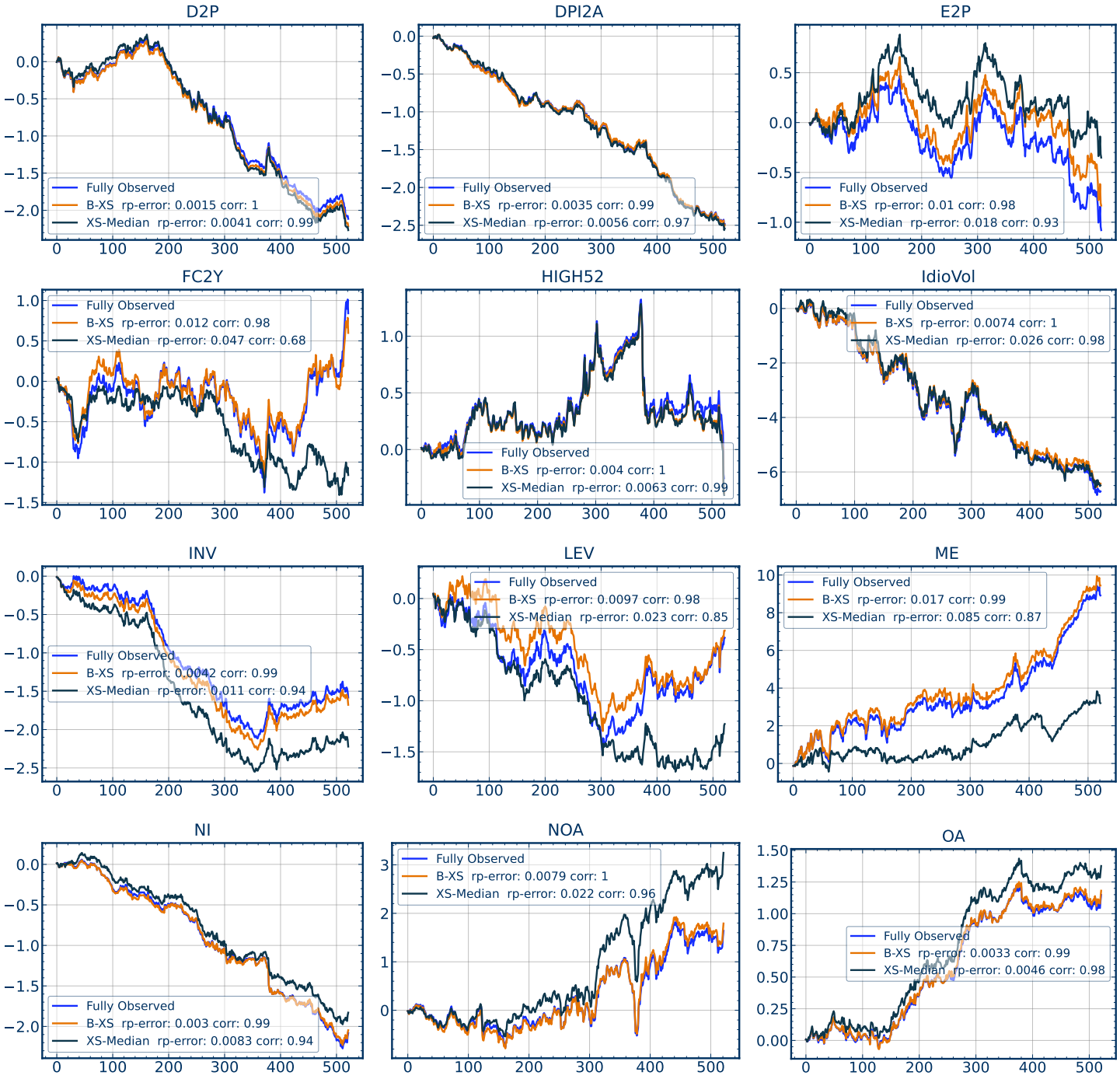
-

Figure C.9: Characteristic mimicking factor portfolios



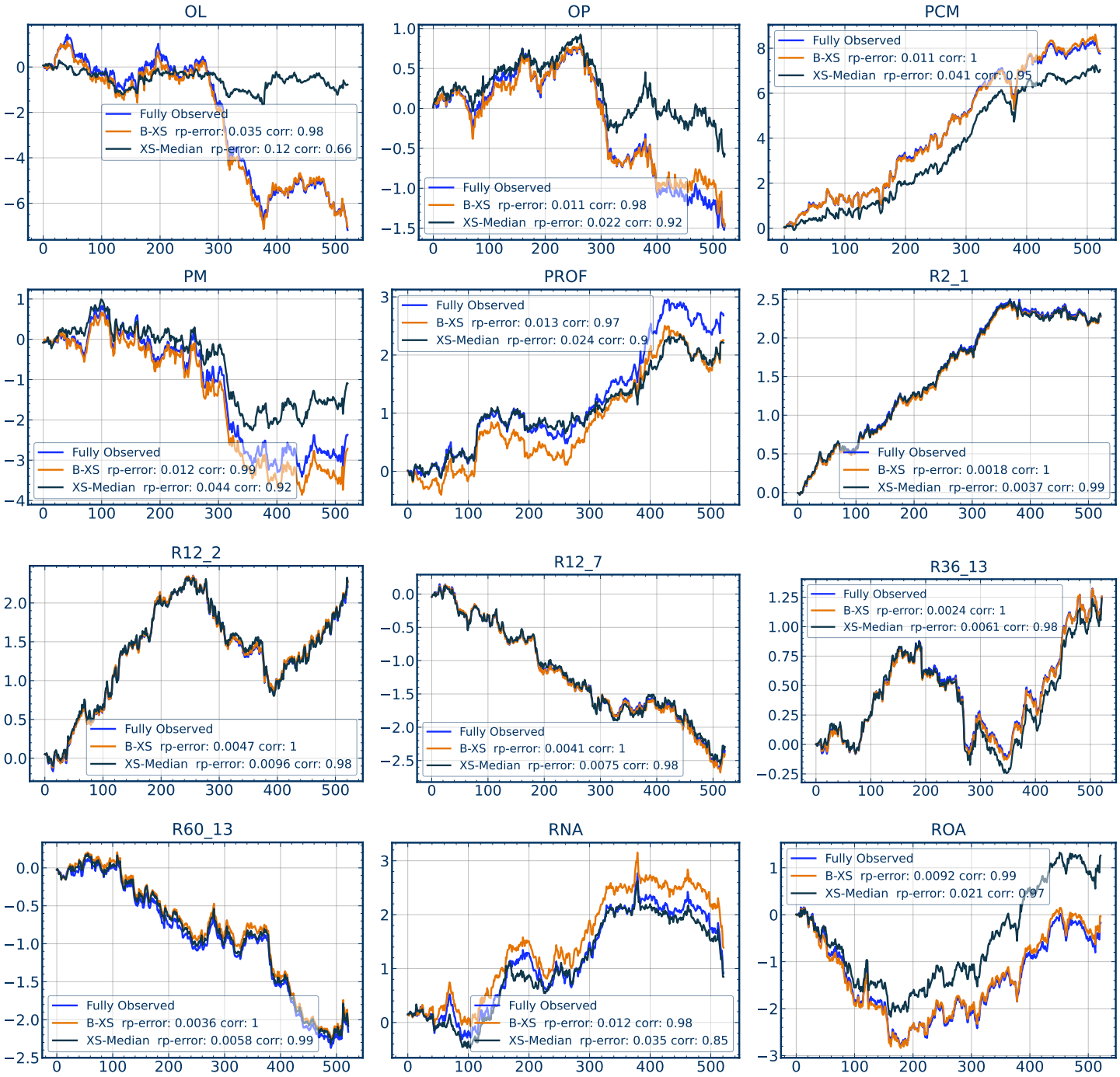
Note: These figures show the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference. We report the correlation and absolute error in characteristic risk premia.

Figure C.10: Characteristic mimicking factor portfolios



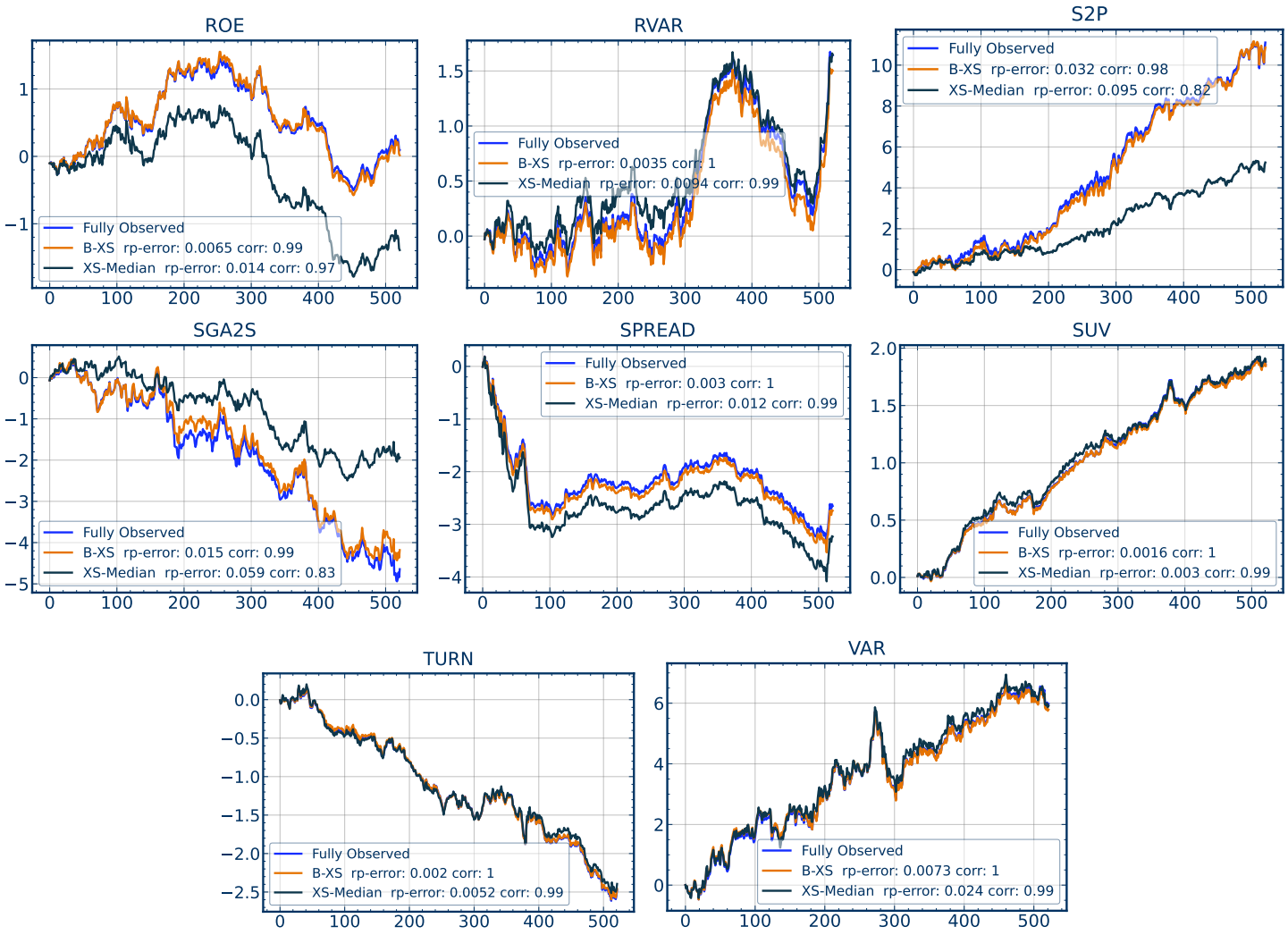
Note: These figures show the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference. We report the correlation and absolute error in characteristic risk premia.

Figure C.11: Characteristic mimicking factor portfolios



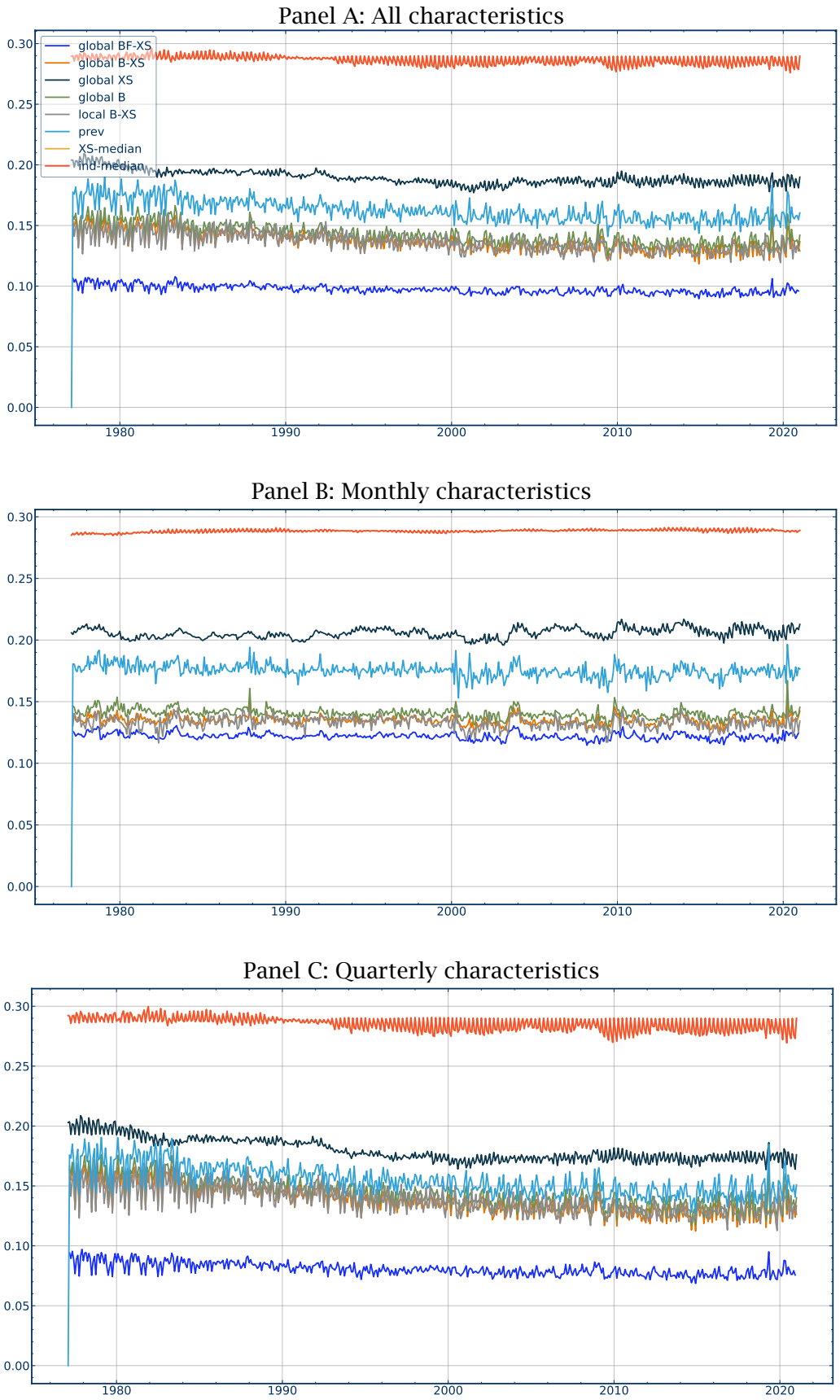
Note: These figures show the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference. We report the correlation and absolute error in characteristic risk premia.

Figure C.12: Characteristic mimicking factor portfolios



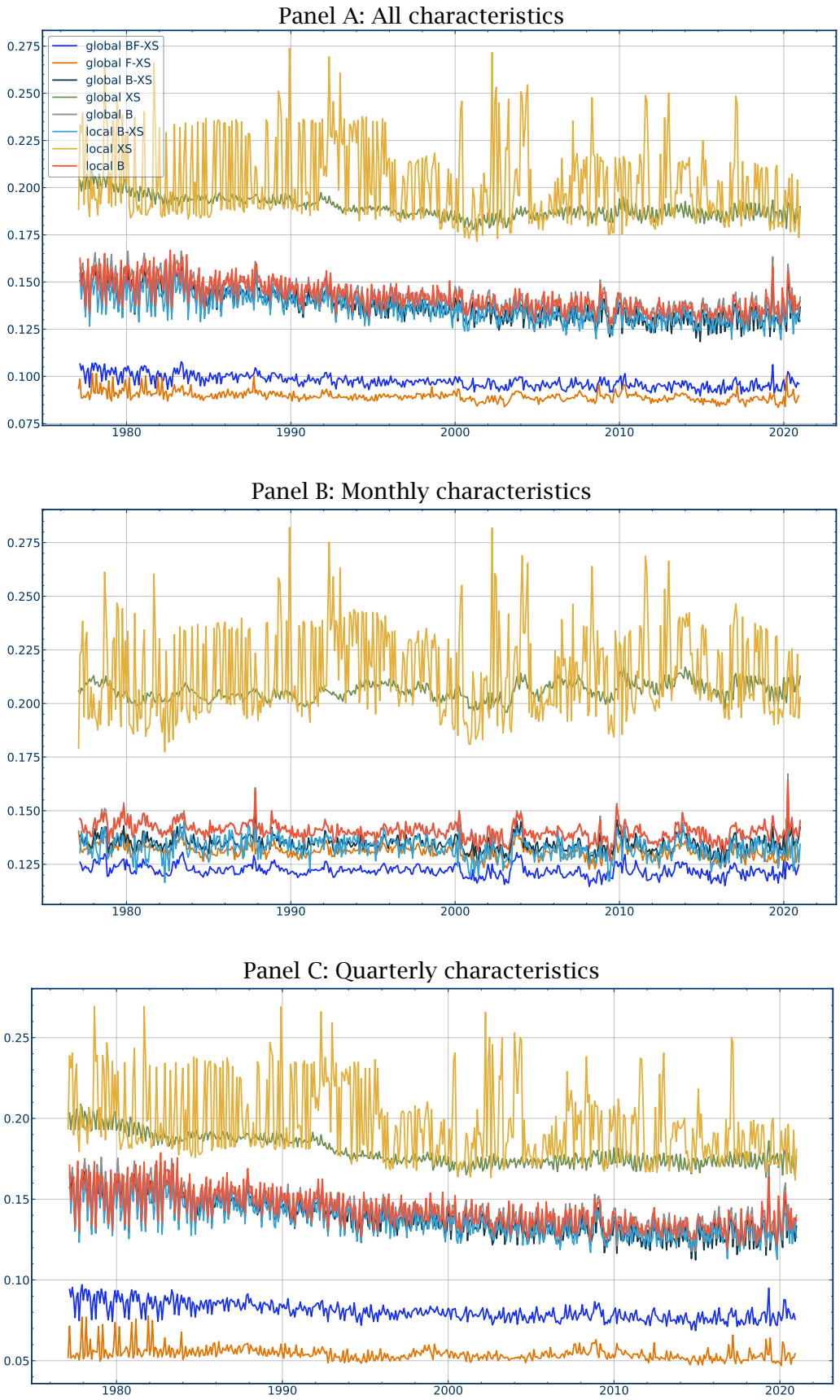
Note: These figures show the time-series of cumulative excess returns of characteristic mimicking factor portfolios with and without imputation. We estimate characteristic mimicking factor portfolios with cross-sectional regressions of stock excess returns on characteristics. We mask the characteristic values based on the empirical pattern with the logistic regression propensity. The masked values are imputed either with the local B-XS or median value. The mimicking portfolio without masking is the reference. We report the correlation and absolute error in characteristic risk premia.

Figure C.13: In-Sample Imputation Error Over Time



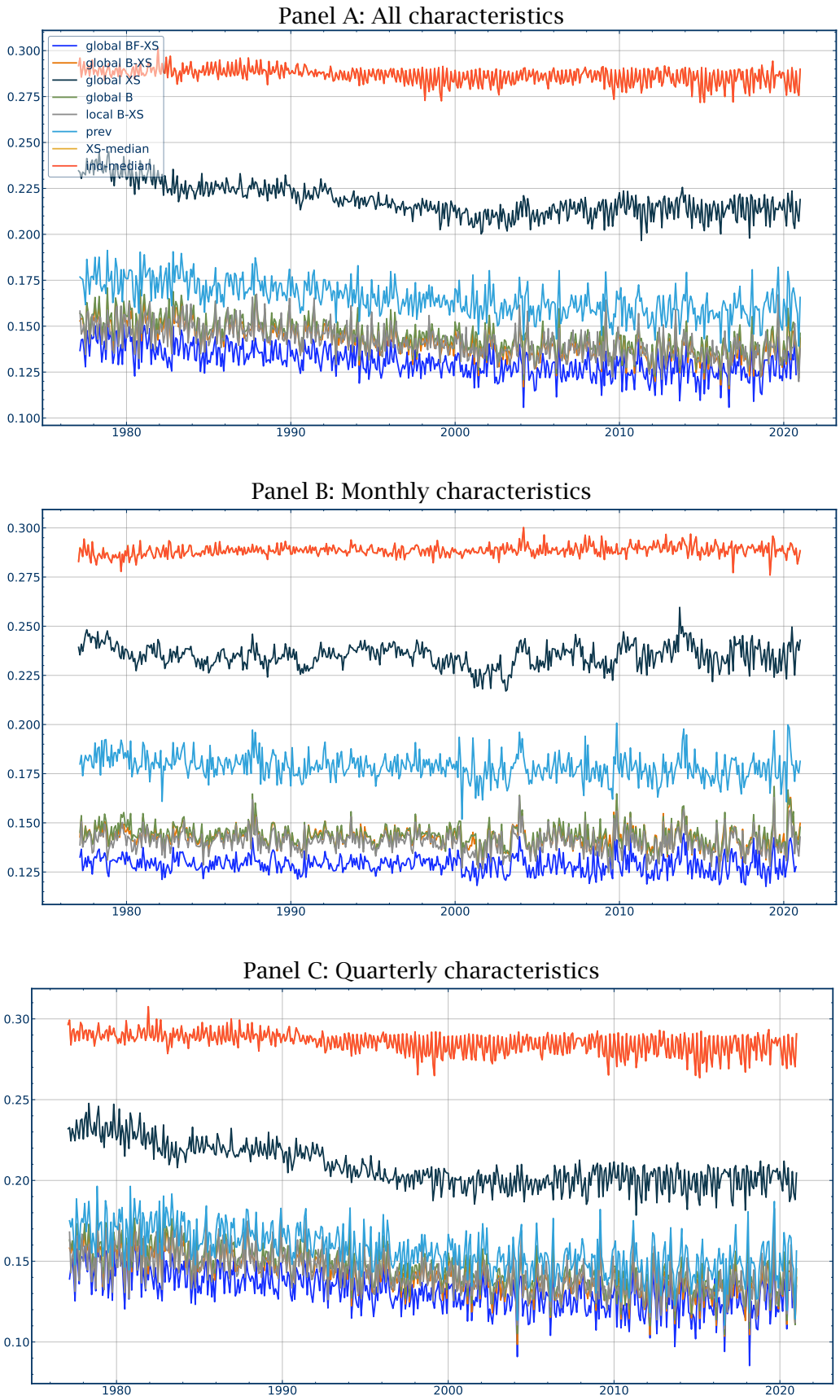
Note: This figure shows in-sample time series RMSE_t for different imputation methods. This is evaluated over all observed data in the sample.

Figure C.14: In-Sample Imputation Error Over Time



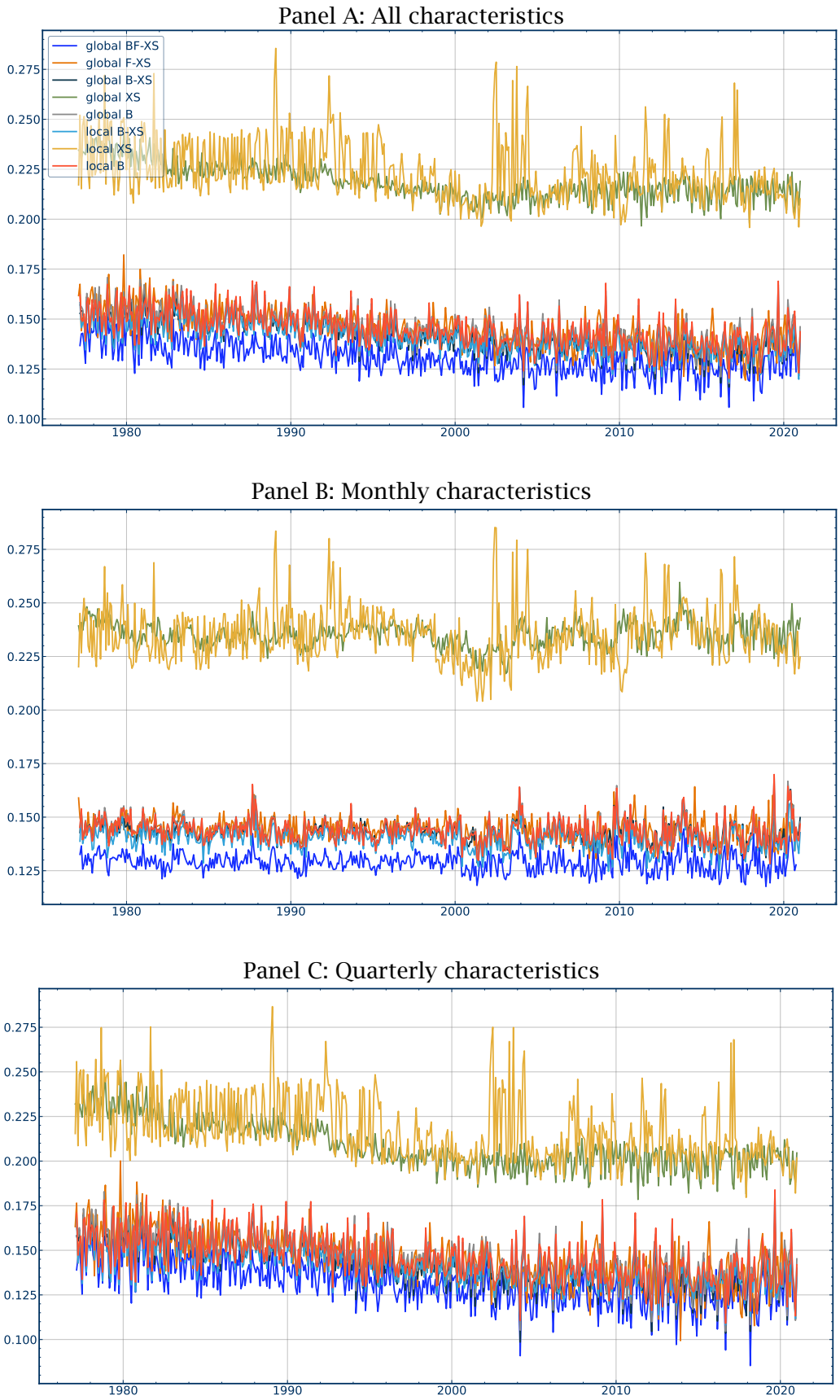
Note: This figure shows in-sample time series RMSE_t for different imputation methods. This is evaluated over all observed data in the sample.

Figure C.15: Out-Of-Sample Missing at Random Imputation Error Over Time



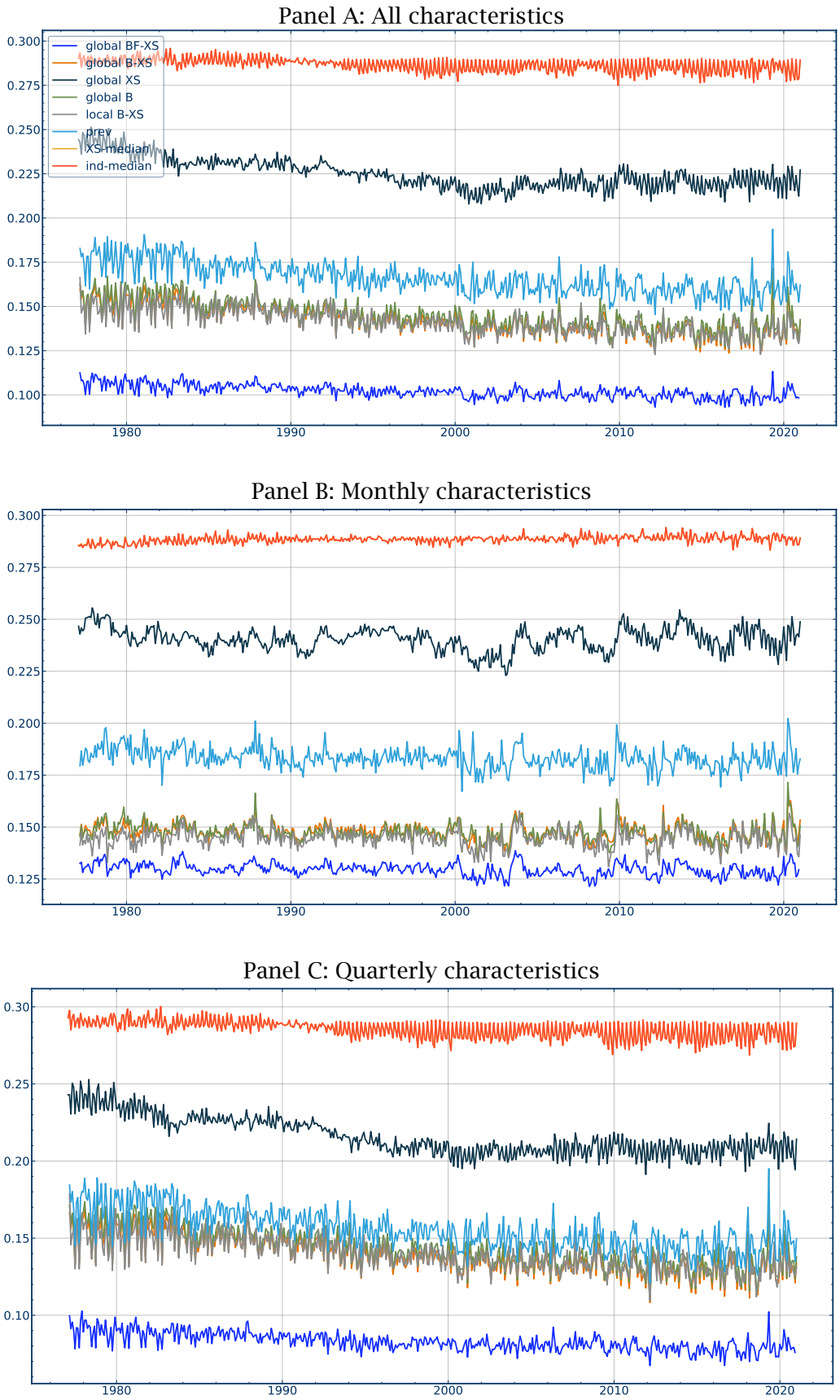
Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure C.16: Out-Of-Sample Missing at Random Imputation Error Over Time



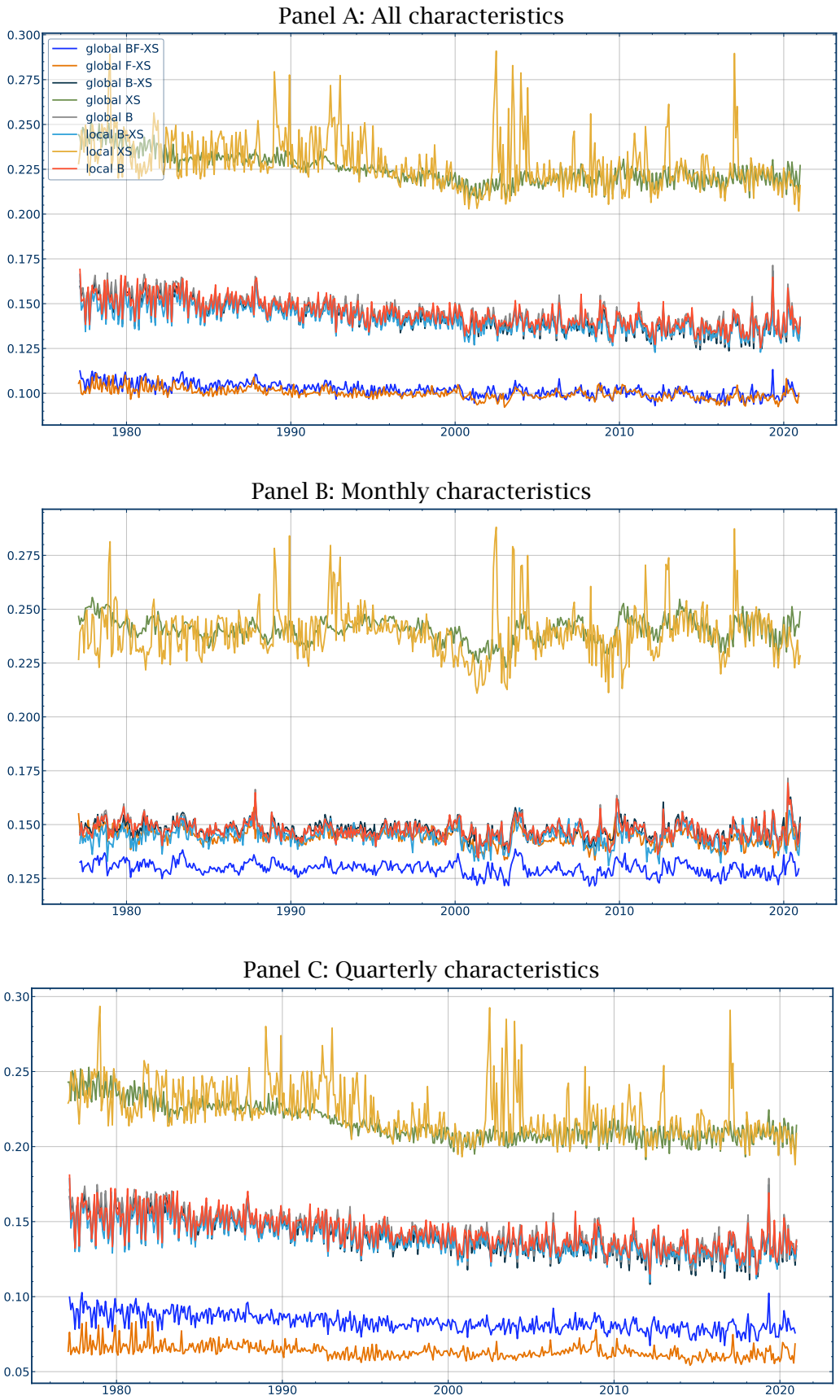
Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure C.17: Out-Of-Sample Block Missing Imputation Error Over Time



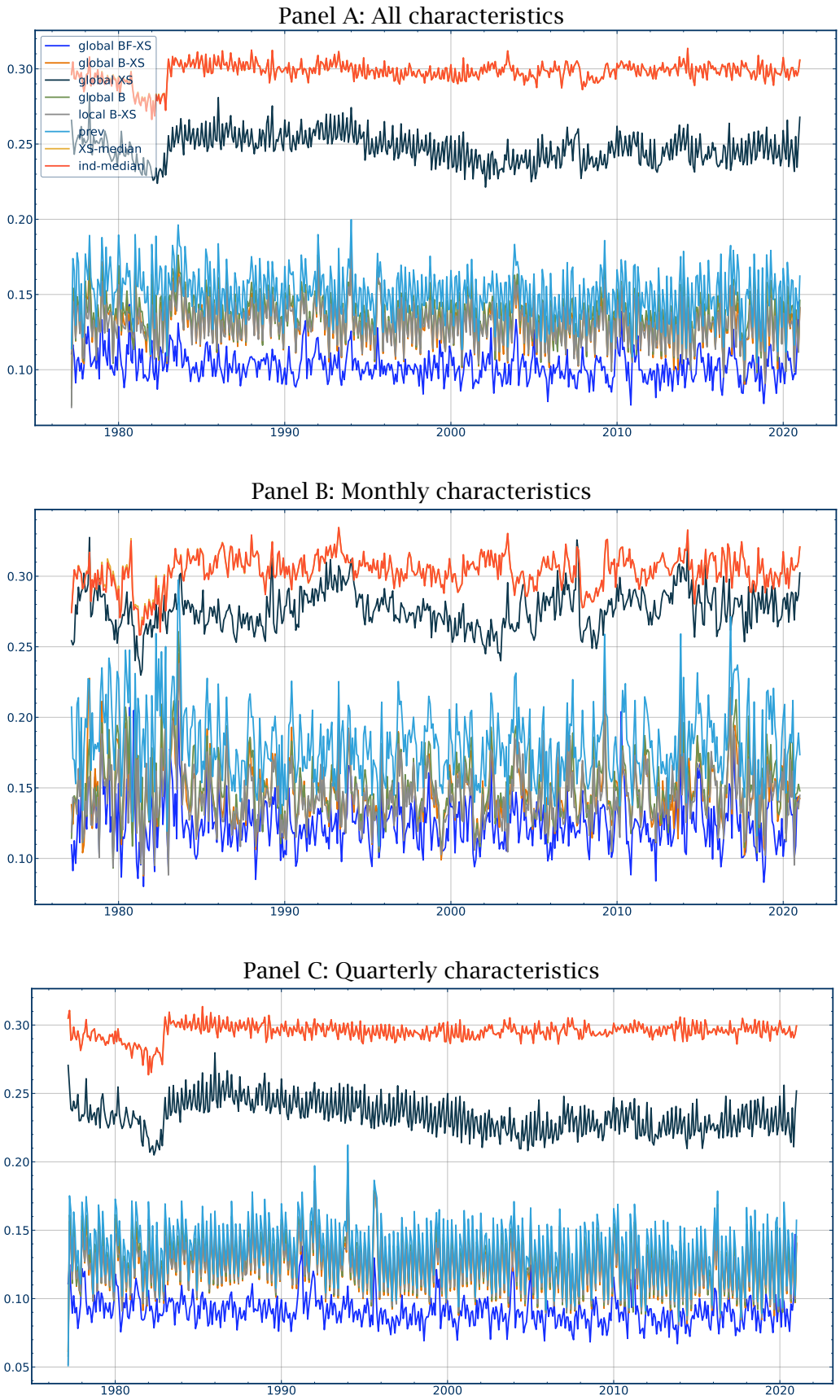
Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure C.18: Out-Of-Sample Block Missing Imputation Error Over Time



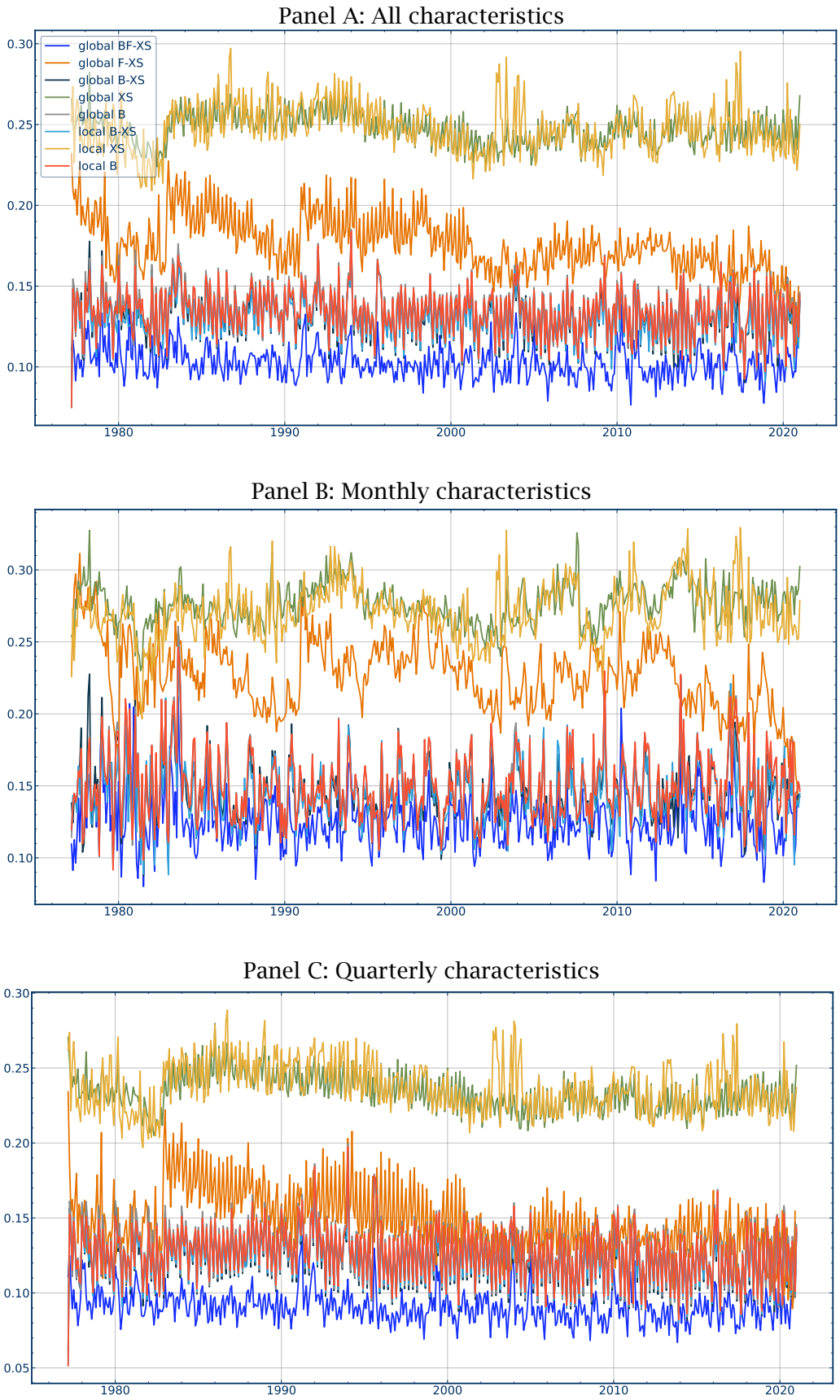
Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure C.19: Out-Of-Sample Logit Missing Imputation Error Over Time



Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.

Figure C.20: Out-Of-Sample Logit Missing Imputation Error Over Time



Note: This figure shows out-of-sample time series RMSE_t for different imputation methods. This is evaluated over the masked out-of-sample characteristics.