Missing Financial Data

Svetlana Bryzgalova[1], Sven Lerner[2], Martin Lettau[3] and Markus Pelger[2]

[1]London Business School and CEPR [2]Stanford University [3]UC Berkeley, NBER, and CEPR

Missing financial data

Firm characteristics are crucial in asset pricing:

- Investment strategies (sorts, machine learning, panel models, etc.),
- Reduced-form, e.g. factors, and structural asset pricing models
- Test assets for models (e.g., double sorts),

Fundamental problem: Missing firm fundamentals

- Characteristic data is missing for a large number of firms,
- Massive in the presence of multiple characteristics

$\Rightarrow$ Unavoidable for Cochrane's (2011) "multivariate challenge"

Key questions: Does missing data matter and how should we deal with it?

Current standard of dealing with missing data strongly biased

- Only fully observed data $\Rightarrow$ sample selection
- Ad-hoc imputation (cross-sectional average, past observations)

This paper:

1. Key facts on missing characteristics
2. Novel method to impute missing values
3. Implications for asset pricing

Contribution of this paper I: Comprehensive empirical study

Stylized facts on missing fundamentals:

Fact #1: Missing data is prevalent:

- Almost all characteristics have missing observations
- Affects small and large, young and mature, profitable and distressed firms

Fact #2: Missingness particularly severe for multiple characteristics

- $> 70\%$ of the firms are missing some of the popular characteristics at any time
- 50% of market capitalization missing for fully observed panel

Fact #3: Data is not missing at random

- systematic patterns, clusters in time and characteristics
- more missingness for extreme realizations and smaller stocks

Fact #4: Returns depend on missingness

- higher expected returns for stocks with more observations
- missingness has a price impact even on simple anomaly-based strategies.

$\Rightarrow$ widespread implications for the "multivariate challenge" in asset pricing,

Contribution of this paper II: Novel Data Imputation Method

Challenges of data imputation:

- Characteristics are predictable ⇒ Avoid omitted variable bias
- Characteristics are not missing at random ⇒ Avoid selection bias
- Complex model estimated on observed data must be valid on missing data

⇒ Example: omitted variable and selection bias of median for small value stocks

Method: A cross-sectional and time-series factor model for characteristics

- Cross-sectional dependency (XS) explained by latent factors
- Persistence (TS) captured by a time-series model,
- Allows for general endogenous missing patterns: missingness can depend on time, stocks, characteristics and factor model

⇒ data-driven, transparent and simple-to-implement

Empirics:

- Comprehensive comparison of approaches to imputation,
- 40-50% reduction in the imputation error relative to existing benchmarks,
- TS and XS both matter, and depend on the characteristic and its missingness.

⇒ A reference dataset with imputed values for any follow-up work.

Literature (incomplete and partial list)

Missing financial data:

- GMM with missing data: Freyberger et al. (2021)
- Look-ahead-bias in imputation for out-of-sample investment: Blanchet et. al. (2022)
- Imputation for causal inference of publication effect: Xiong and Pelger (2022)

$\Rightarrow$ Different goal and complementary

Missing data in panel

- Latent factor models: Xiong and Pelger (2019), Bai and Ng (2021), Jin et al. (2021)
- Matrix completion: Athey et. al. (2018), Chen et al. (2019)
- Transfer learning with Target PCA: Duan, Pelger and Xiong (2022)

$\Rightarrow$ Only 2-D, challenge general missing patterns

Latent factor modeling in finance

- Unconditional: Connor et. al. (1988), Lettau and Pelger (2020a+b), Pelger (2019)
- Conditional: Kelly et. al. (2019), Pelger and Xiong (2021)

$\Rightarrow$ PCA type methods for fully observed panel of returns

Asset pricing with many characteristics

- Prediction: Freyberger et al. (2020), Gu et al. (2020), Kaniel et al. (2021)
- SDF modeling: Bryzgalova et al. (2019), Chen et al. (2019), Kozak et al. (2020)

$\Rightarrow$ Requires choices for missing data

Missing Data: Stylized Facts

Data

Dataset:

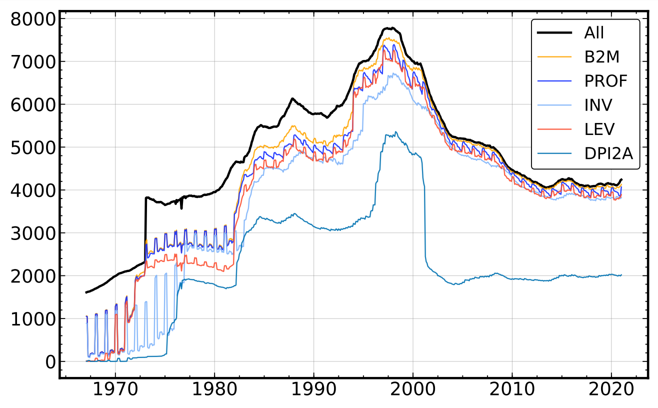
- Standard CRSP/Compustat universe + usual filters for outliers, exchanges, etc.
- Sample size: monthly returns 1967:07 – 2020:12
- 45 characteristics: value, investment, profitability, intangibles, past returns, trading frictions, etc.
- Characteristics raw values are converted into centered rank quantiles
- Characteristics are updated monthly or quarterly

Standard dataset for many modern asset pricing applications:

- the most popular characteristics, used individually and combined
- standard set of filters/transformations

Missing data: How big of a problem?

Even key firm characteristics are missing for many companies

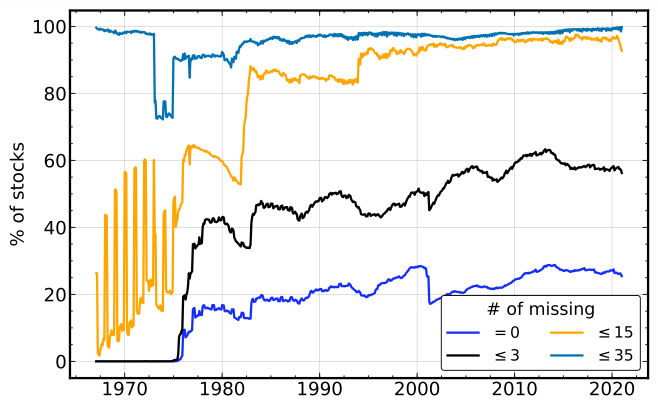


Firms with observed characteristic value

(value, gross profitability, investment, leverage, change in PPEI/assets)

- (Almost) any characteristic has missing observations
- The number of firms missing fundamentals is statistically and economically large
- Substantial cross-sectional and time variation

Elephant in the room: multiple characteristics

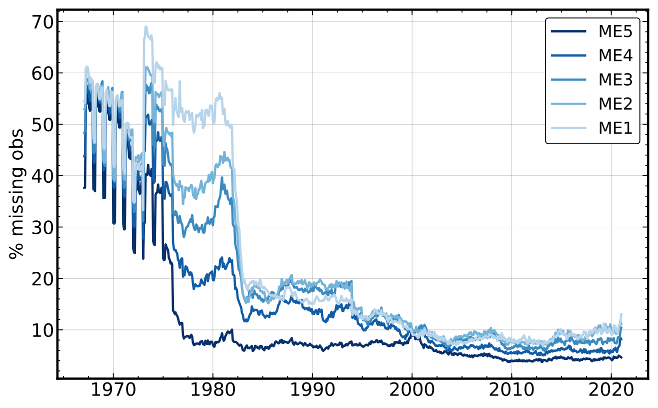


Percentage of firms missing some characteristics (from the list of 45)

- Missing data is a paramount problem whenever multiple characteristics are used
- $> 70\%$ of firms are missing at least some popular characteristics at any period
- Their total market cap is 48%

⇒ Using a fully observed panel of data may lead to massive sample selection: crucial for panel models, conditional factors, and machine learning.

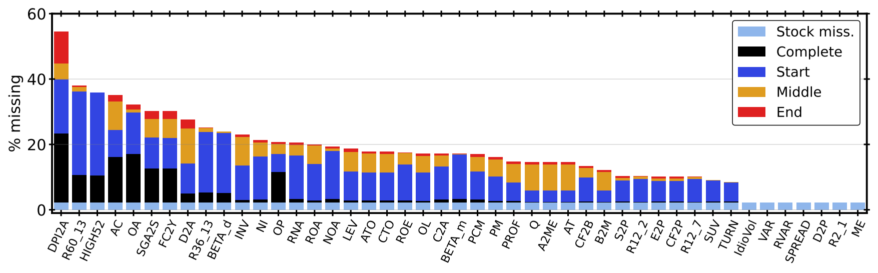
Missingness affects both small and large caps



Percentage of missing firm-month observations within quintiles

- Historically smaller companies used to have worse data coverage
- Last 20 years: similar patterns

When are characteristics missing?



Start = no previous observations

End = no further observations

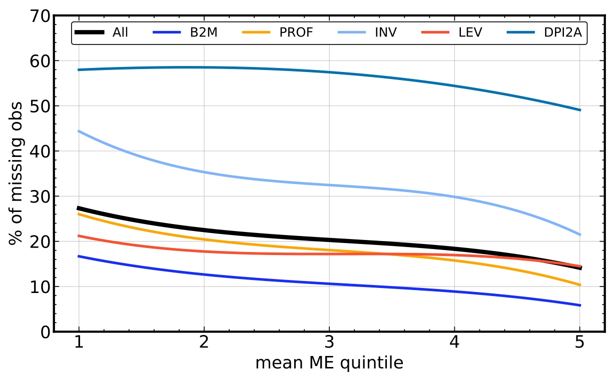
Middle = some previous and future observations

Complete = completely missing

- Some characteristics are mechanically missing for younger firms (e.g., LTrev)
- Many characteristics are missing after having been previously observed
- Some characteristics are missing at the end of the company's life
- Some are never observed

⇒ Imputation needs to allow for different information sets

Which stocks have missing observations?

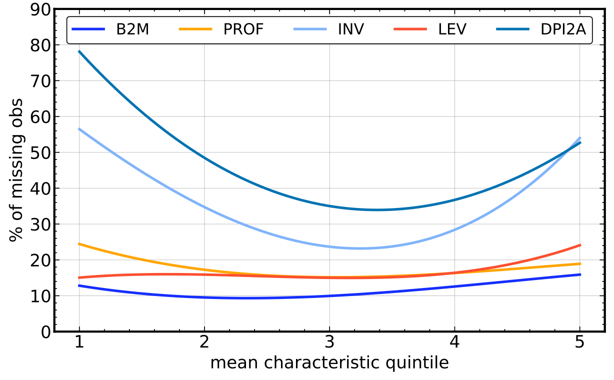


Percentage of missing characteristics by size quintiles

- Smaller companies have more missing observations
- Complex interactions of size and heterogeneous missingness

⇒ Firms with observed data are different ⇒ selection bias

Which characteristic realizations are missing?

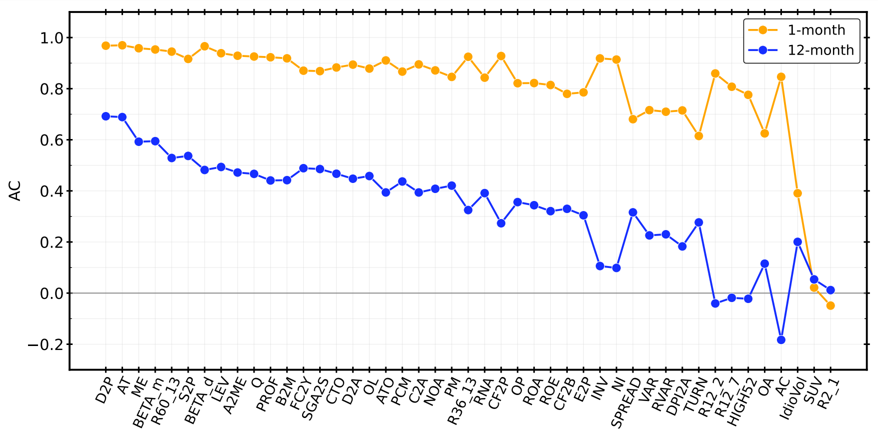


Percentage of missing characteristics by average characteristic quintiles

- More extreme realizations of characteristics are more likely to be unobserved
- U-shaped pattern generalizes to most characteristics
- Missingness depends on characteristic realization
- Endogenous missingness $\Rightarrow$ challenging statistical problem

Characteristics Dependency

Characteristics are persistent

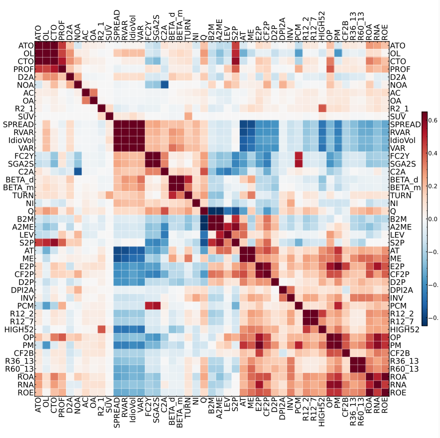


Average sample autocorrelation for each characteristic

- Many characteristics are very persistent
- Past (and future) values have information for missing values

⇒ Disregarding time dependency when imputing values might lead to a bias

Characteristics are cross-sectionally correlated



Pairwise correlations in characteristics, averaged over time and stocks

- Strong cross-sectional dependence
- Contemporaneous correlated characteristics have information for missing values
- Challenge: A model for the complex dependencies (avoid omitted variable bias)

We need a way of imputing characteristics using both cross-section and time-series.

Model

Model formulation

Characteristics form a 3-dimensional vector space:

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

- Cross-sectional stock dimension $i = 1, ..., N_t$
- Time-series dimension $t = 1, ..., T$
- Different characteristics $l = 1, ..., L$

$\Rightarrow$ Goal: A low-dimensional model for cross-sectional and time-series dependency

Our baseline model uses centered rank quantiles:

- Stationarity in the cross-section and over time and deals with outliers
- Simple mapping between rank quantiles and raw values through empirical density
- Similar results in raw characteristic space after appropriate kernel transformation

Cross-sectional factor model

Approximate factor structure for $N_t \times L$ characteristic matrix C^t at time t:

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

- Allows for a separate factor model for each time t,
- K latent factors: $F^t \in \mathbb{R}^{N_t \times K}$ and $\Lambda^t \in \mathbb{R}^{L \times K}$,
- without missing values, estimate model with PCA applied to $C^t C^{t\top}$.

A general approach to estimation:

1. Estimate F^t_i as the eigenvectors of the K largest eigenvalues of

$$\tilde{\Sigma}^{\mathrm{XS},t}_{i,j} = \frac{1}{|Q^t_{i,j}|} \sum_{l \in Q^t_{i,j}} C^t_{i,l} C^t_{j,l},$$

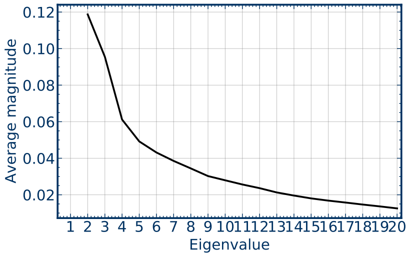
   with $Q^t_{i,j}$ set of characteristics observed for stocks i and j at time t
2. Estimate loadings Λ^t_l from the characteristic regression:

$$\hat{\Lambda}^t_l = \left(\sum_{i=1}^{N_t} W^t_{i,l} \hat{F}^t_i \hat{F}^{t\top}_i \right)^{-1} \left(\sum_{i=1}^{N_t} W^t_{i,l} \hat{F}^t_i C^t_{i,l} \right),$$

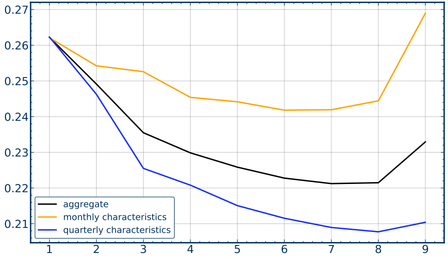
   where $W^t_{i,l} = 1$ if char. l is observed for stock i at time t and $W^t_{i,l} = 0$ o/w.

Asymptotic theory and proofs: Xiong and Pelger (2019).

A factor model for characteristics



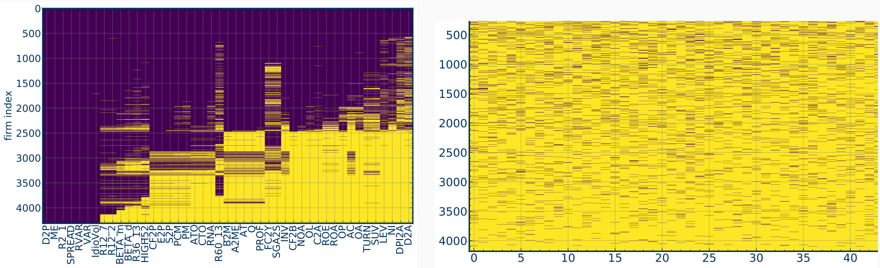
(a) Eigenvalues of $\tilde{\Sigma}^{XS,t}_{l,p}$ (time-averaged)



(b) Out-of-sample RMSE for different number of factors

- Strong factor structure in characteristics:
- $K = 6$ factors capture most of the cross-sectional variation (our baseline model).
- Extensive robustness results for different number of factors.
- Characteristic factors have economic interpretation.

Distribution of missingness



(a) Sample month: April 1981

(b) Simulated missing-at-random

Joint distribution of missing patterns (yellow missing)

- Missingness clusters in time and cross-section, and is heterogenous.
- Characteristics are not missing at random.
- Selection bias for model estimated on observed data assuming missing-at-random
- Our approach allows for general missing patterns (different from most literature)

Assumptions on missingness

Assumptions on probability of missingness $\mathbb{P}\left(W_{i,l}^{t}=0\right)=: p_{i,l}^{t}$:

Dependence on time t:

- No assumption on temporal structure as different factor model for each t
- Examples: block-missing, mixed-frequency, dependence on prior missingness ...

Dependence on characteristic l:

- Characteristic specific heterogeneity
- Examples: Investment more likely to miss than B2M

Dependence on stock i:

- Extremely general dependence on features of stocks
- General time-varying and characteristic-specific function $p_{i,l}^{t}=f_{l,t}(F_i^t, S_i^t)$ of unknown stock-specific features $S_i^t \in \mathbb{R}^r$ and the stock-specific factors F_i^t
- Examples: small stocks or more extreme realizations more likely to miss

Identification restrictions:

- Missingness $W_{i,l}^{t}$ independent of loadings Λ_l^t and error $e_{i,l}^{t}$
- Same characteristic covariance matrix $\tilde{\Sigma}_{i,j}^{\text{XS},t}$ on partially and fully observed data

$\Rightarrow$ Intuition: Identify "similar" stocks from observed data

Adding time-series information

Combine XS (cross-sectional) with TS (time-series) information:

- **B-XS-Model:** (backward-cross-sectional)

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

- **BF-XS-Model:** (backward-forward-cross-sectional)

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

The framework includes several important special cases:

1. Time-series AR(1) model (B): $\beta^{l,\text{B-XS}} = \begin{pmatrix} \beta^{\text{B}} & 0 & \cdots & 0 \end{pmatrix}$.
2. Last observed value (PV): $\beta^{l,\text{B-XS}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$.
3. Cross-sectional median: $\beta^{l,\text{B-XS}} = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}$ (ranks centered at 0).

For estimation, stack cross-sectional and time-series information in $X^{l,t}_{i}$ and run the following regression (averaged over observed stocks):

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

Imputation models

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

Different imputation methods sorted by the size of the information set

- Current standard in the literature: Cross-sectional median or previous value
- The need for past/future information restricts available options for imputation

$\Rightarrow$ Different types of missing values might benefit from different methods

Global and local factor models

Global model assumes that factor composition Λ and β stays constant over time:

- Global model estimated with global (pooled) regression

$$\hat{\beta}^{l} = \left(\sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(W_{i,l}^{t} X_i^{l,t} X_i^{l,t\top} \right) \right)^{-1} \left(\sum_{t=1}^{T} \sum_{i=1}^{N_t} \left(W_{i,l}^{t} X_i^{l,t} C_{i,t}^{l} \right) \right)$$

- Estimate rotation of global Λ from average characteristic covariance matrix

$$\tilde{\Sigma}_{l,p}^{\mathrm{XS}} = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{|O_{l,p}^{t}|} \sum_{i \in O_{l,p}^{t}} C_{i,l}^{t} C_{i,p}^{t} \right)$$

- Local model estimates factor models and $\hat{\beta}^{l,t}$ for each t independently

Local vs. global tradeoff:

- Global estimation is more efficient (uses more information)
- Local estimation allows for time-variation (less bias) and avoids look-ahead bias

Imputing Characteristics

Evaluation metrics

We evaluate the models based on their RMSE (root mean squared errors).

- Aggregated RMSE for the model prediction $\hat{C}_{i,t,l}$ averaged over all stocks, time-periods and characteristics:
  $$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}$$
- RMSE for each characteristic separately:
  $$\text{RMSE}_l = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}$$
- RMSE for each time period:
  $$\text{RMSE}_t = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2}$$

Out-of-sample evaluation:

- OOS Block-missing: Masking 10% of characteristics in blocks of 1 year
- OOS Missing-at-random: Masking 10% of characteristics randomly
- OOS Propensity: Masking with empirical distribution of missing data
- In-sample results on observed characteristics

Aggregate results

Method	all characteristics	quarterly characteristics	monthly characteristics
global BF-XS	**0.18**	**0.17**	**0.18**
global F-XS	0.19	0.18	0.20
global B-XS	0.19	0.19	0.20
global XS	0.23	0.22	0.24
global B	0.21	0.21	0.22
local B-XS	**0.20**	**0.20**	**0.20**
local B	0.21	0.21	0.22
local XS	0.23	0.23	0.24
prev val	0.23	0.22	0.25
XS median	0.29	0.28	0.29
industry median	0.29	0.28	0.29

Out-of-sample RMSE for block-missing characteristics

- Baseline models:
  - local B-XS (no look-ahead-bias)
  - global BF-XS (full possible information)
- Current standard (cross-sectional median and last observed value) is the worst
- Similar results for missing-at-random and in-sample

Imputation error for different types of missingness

	In-Sample			OOS MAR			OOS Block		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
				Start of the sample					
global BF-XS	–	–	–	–	–	–	–	–	–
global F-XS	**0.11**	**0.06**	**0.15**	**0.19**	**0.20**	**0.19**	**0.20**	**0.19**	**0.22**
global B-XS	–	–	–	–	–	–	–	–	–
global XS	0.21	0.19	0.23	0.32	0.33	0.29	0.26	0.25	0.27
global B	–	–	–	–	–	–	–	–	–
local B-XS	–	–	–	–	–	–	–	–	–
local XS	**0.23**	**0.22**	**0.25**	**0.28**	**0.27**	**0.29**	**0.26**	**0.25**	**0.27**
prev	–	–	–	–	–	–	–	–	–
local B	–	–	–	–	–	–	–	–	–
XS-median	0.32	0.32	0.32	0.32	0.32	0.31	0.31	0.31	0.30
ind-median	0.32	0.32	0.32	0.32	0.32	0.31	0.31	0.31	0.30

In- and out-of-sample RMSE for different types of missing observations

- Bold indicates best local (lock-ahead-bias free) and global model
- Our baseline models dominate across all the missing patterns
- Availability of models depends on type of missingness

Imputation error for different types of missingness

	In-Sample			OOS MAR			OOS Block		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
				Middle of the sample					
global BF-XS	**0.11**	**0.10**	**0.13**	**0.15**	**0.16**	**0.15**	**0.17**	**0.17**	**0.19**
global F-XS	0.10	0.07	0.14	0.17	0.17	0.16	0.19	0.18	0.20
global B-XS	0.15	0.16	0.14	0.17	0.17	0.16	0.19	0.19	0.20
global XS	0.19	0.18	0.21	0.24	0.24	0.25	0.23	0.22	0.24
global B	0.16	0.17	0.15	0.18	0.18	0.16	0.21	0.21	0.22
local B-XS	**0.15**	**0.16**	**0.14**	**0.17**	**0.17**	**0.16**	**0.20**	**0.20**	**0.20**
local XS	0.21	0.20	0.22	0.24	0.24	0.25	0.23	0.23	0.24
prev	0.18	0.18	0.18	0.20	0.20	0.20	0.23	0.22	0.25
local B	0.16	0.17	0.15	0.18	0.18	0.16	0.21	0.21	0.22
XS-median	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.29
ind-median	0.29	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.29

In- and out-of-sample RMSE for different types of missing observations

- Bold indicates best local (lock-ahead-bias free) and global model
- Our baseline models dominate across all the missing patterns
- Availability of models depends on type of missingness

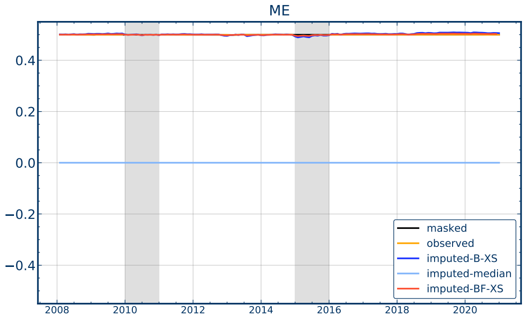
Imputation error for different types of missingness

	In-Sample			OOS MAR			OOS Block		
Method	all	quarterly	monthly	all	quarterly	monthly	all	quarterly	monthly
End of the sample									
global BF-XS	–	–	–	–	–	–	–	–	–
global F-XS	–	–	–	–	–	–	–	–	–
global B-XS	**0.19**	**0.21**	**0.16**	**0.20**	**0.21**	**0.19**	**0.22**	**0.22**	**0.22**
global XS	0.24	0.25	0.23	0.30	0.29	0.30	0.26	0.25	0.26
global B	0.21	0.23	0.18	0.21	0.22	0.19	0.25	0.25	0.24
local B-XS	**0.20**	**0.22**	**0.16**	**0.21**	**0.21**	**0.19**	**0.23**	**0.23**	**0.22**
local XS	0.27	0.27	0.26	0.29	0.28	0.31	0.26	0.26	0.27
prev	0.23	0.24	0.21	0.23	0.24	0.22	0.27	0.26	0.28
local B	0.21	0.23	0.18	0.21	0.22	0.19	0.25	0.25	0.24
XS-median	0.35	0.36	0.34	0.33	0.33	0.33	0.32	0.32	0.32
ind-median	0.35	0.36	0.34	0.33	0.33	0.33	0.32	0.32	0.32

In- and out-of-sample RMSE for different types of missing observations

- Bold indicates best local (lock-ahead-bias free) and global model
- Our baseline models dominate across all the missing patterns
- Availability of models depends on type of missingness

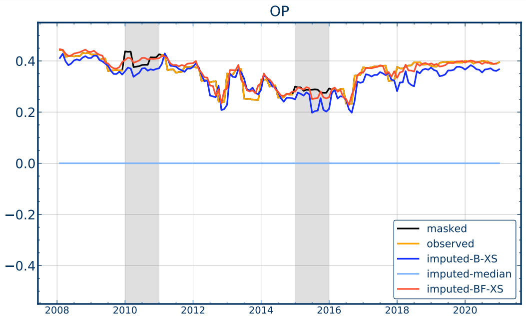
Illustration: Persistent characteristics



Size of Microsoft: Model-implied and observed time-series

- Size as representative persistent characteristic; other examples: AT, D2P, LEV
- Gray blocks: 1-year out-of-sample imputation
- B-XS and BF-XS extremely precise
- Time-series observation provides close to perfect prediction
- Median wrong level and dynamics

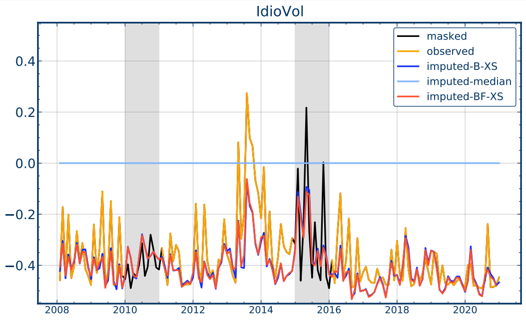
Illustration: Persistent and volatile characteristics



Operating Profitability of Microsoft: Model-implied and observed time-series

- OP as representative persistent and volatile characteristic; other examples: B2M, E2P, INV
- Gray blocks: 1-year out-of-sample imputation
- B-XS "anchors" at last observed value, dynamics from cross-section
- BF-XS connects endpoints, dynamics from cross-section
- Median wrong level and dynamics

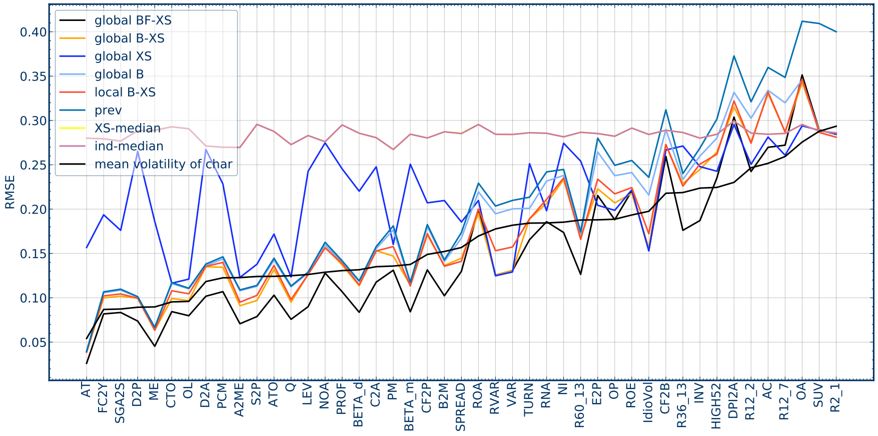
Illustration: Volatile characteristics



Idiosyncratic Volatility of Microsoft: Model-implied and observed time-series

- IVol as representative volatile characteristic; other examples: R2_1, R12_2, SUV
- Gray blocks: 1-year out-of-sample imputation
- Dynamics driven by cross-sectional contemporaneous factors
- Median wrong level and dynamics

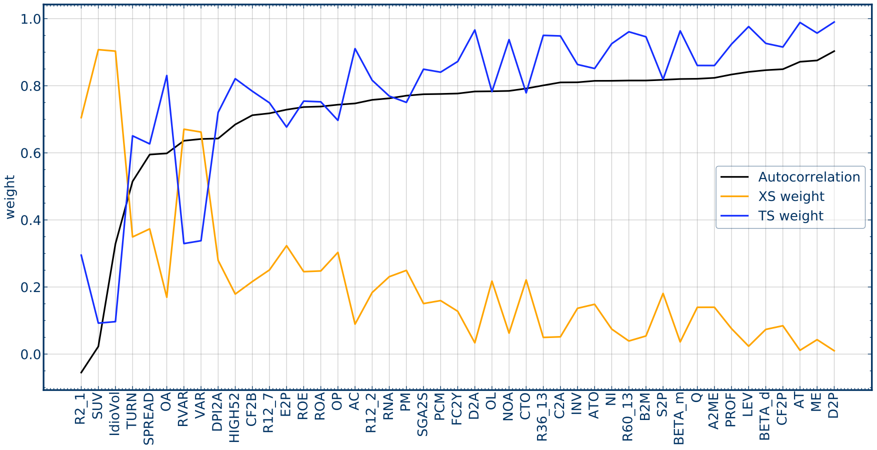
Comparison of imputation methods



Out-of-sample RMSE by imputation method across individual block-missing characteristics

- Characteristics are sorted in ascending order based on their volatility (black line)
- Imputation for persistent characteristics benefits from the TS data
- Imputation for more volatile characteristics relies more on XS information

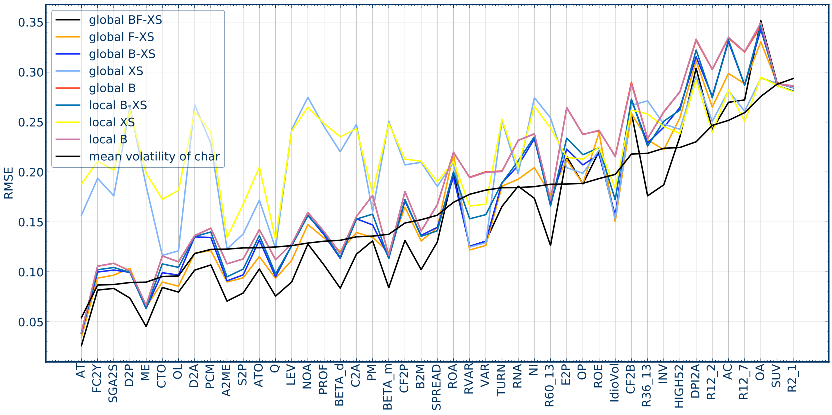
Information used for imputation



Relative importance of the TS and XS components of B-XS (L1 norm)

- Characteristics are sorted in ascending order based on their persistence
- Persistent characteristics put more weight on TS information
- Volatile characteristics put more on XS information

Local vs global imputation

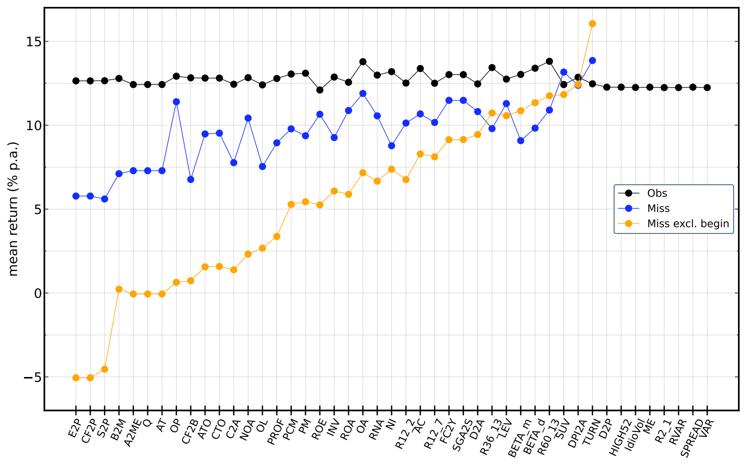


Out-of-sample RMSE by imputation method across individual block-missing characteristics

- Global models are slightly better
- Highly volatile characteristics benefit more from local models

Asset Pricing Results

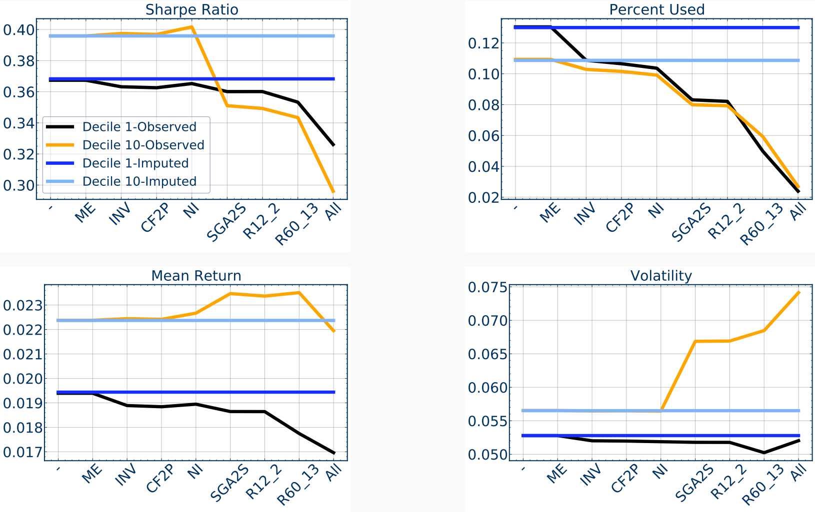
Missingness matters for simple portfolio strategies



Returns on long-only portfolios that include/exclude particular characteristics

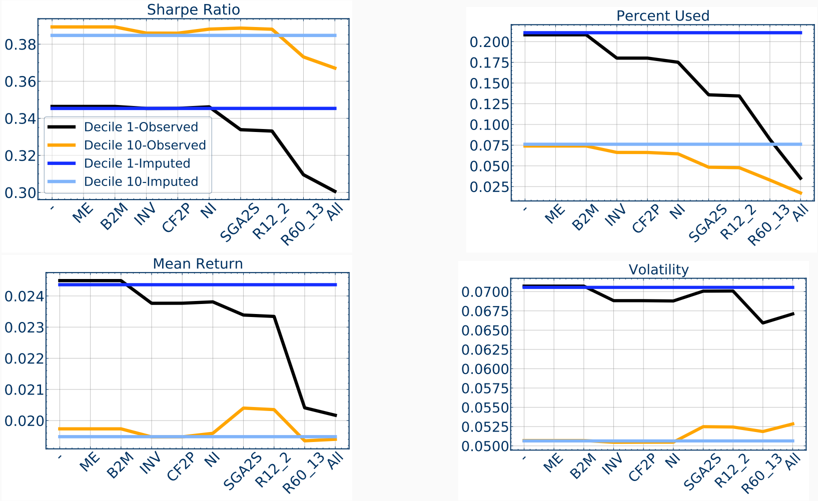
- Portfolios are formed by buying stocks with observed/missing characteristic value
- Significant difference in returns for many characteristics

Book-to-Market, conditional on other observables



- Return on the value sorts, requiring additionally observed characteristics
- Missing information has a direct impact on the return of the simplest strategies
- Effect is larger when multiple signals are used

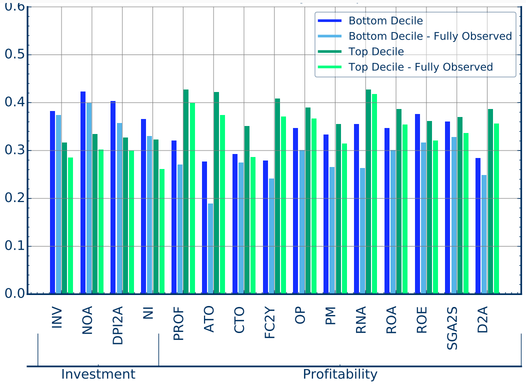
Operating Profitability, conditional on other observables



⇒ Missingness has a stronger impact when multiple characteristics are used

⇒ Implications for multiple sorts, machine learning, and the whole "multidimensional challenge" in asset pricing.

Univariate Portfolio Sorts with and without Missing Values

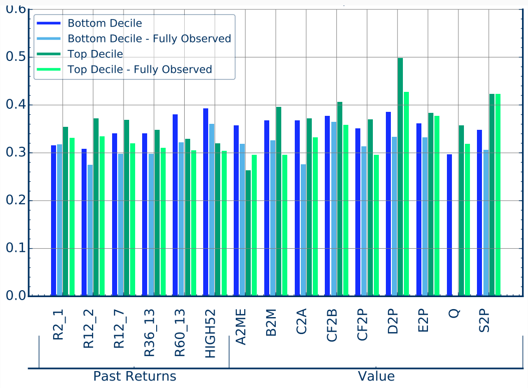


Sharpe ratios of top and bottom decile of sorted portfolios

- Sorts of stocks with observed single characteristic or all 45 characteristics
- Lower Sharpe ratios for fully observed subset
- Mean returns: complex interaction between characteristic and missingness
- Higher volatility: restricted sample has less diversification

⇒ Selection bias applies to all characteristic sorts

Univariate Portfolio Sorts with and without Missing Values

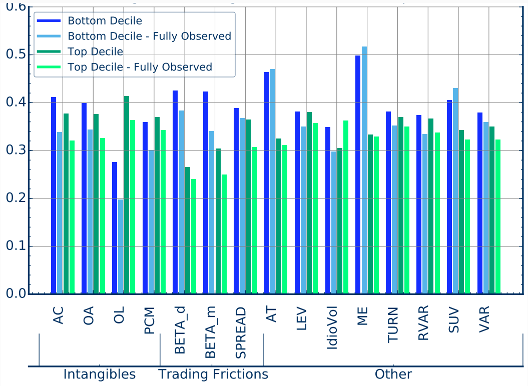


Sharpe ratios of top and bottom decile of sorted portfolios

- Sorts of stocks with observed single characteristic or all 45 characteristics
- Lower Sharpe ratios for fully observed subset
- Mean returns: complex interaction between characteristic and missingness
- Higher volatility: restricted sample has less diversification

⇒ Selection bias applies to all characteristic sorts

Univariate Portfolio Sorts with and without Missing Values



Sharpe ratios of top and bottom decile of sorted portfolios

- Sorts of stocks with observed single characteristic or all 45 characteristics
- Lower Sharpe ratios for fully observed subset
- Mean returns: complex interaction between characteristic and missingness
- Higher volatility: restricted sample has less diversification

⇒ Selection bias applies to all characteristic sorts

Conclusion

Conclusion

A systematic study of missing data in characteristics:

- the problem is pervasive and affects even simple investment strategies,
- complex and endogenous patterns of missingness,
- simple solutions do not work.

A novel method to impute characteristic values:

- a parsimonious model for characteristic structure,
- time-series AND cross-sectional dependence,
- automatically captures a wide range of dependencies and missing patterns.

Outlook:

- a growing challenge in the presence of big data and machine learning,
- numerous implications for asset pricing and corporate finance.

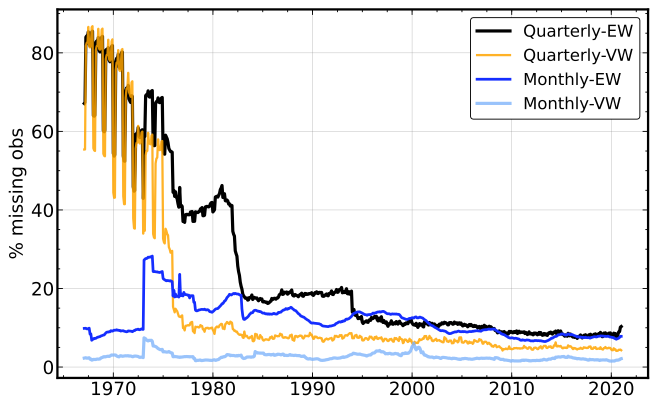
We will provide a publicly available dataset for researchers.

Appendix

Firm characteristics

Past Returns	Investment	Profitability	Intangibles	Value	Trading Frictions
Momentum	Investment	Operating profitability	Accrual	Book to Market Ratio	Size
Short-term Reversal	Net operating assets	Profitability	Operating accruals	Assets to market cap	Turnover
Long-term Reversal	Change in prop. to assets	Sales over assets	Operating leverage	Cash to assets	Idiosyncratic Volatility
Return 2-1	Net Share Issues	Capital turnover	Price to cost margin	Cash flow to book value	CAPM Beta
Return 12-2		Fixed costs to sales		Cashflow to price	Residual Variance
Return 36-13		Profit margin		Dividend to price	Total assets
		Return on net assets		Earnings to price	Market Beta
		Return on assets		Tobin's Q	Close to High
		Return on equity		Sales to price	Spread
		Expenses to sales		Leverage	Unexplained Volume Variance
		Capital intensity			

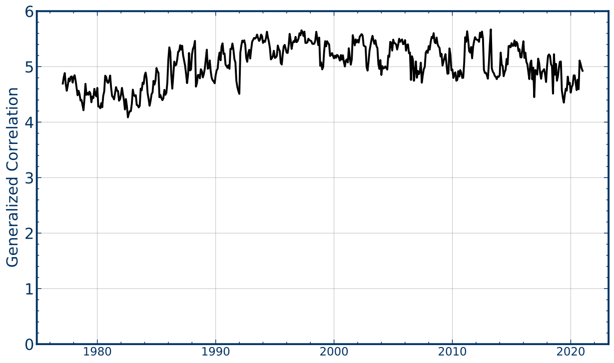
Missingness affects both quarterly and monthly characteristics



Percentage of missing quarterly and monthly updated characteristics

- Historically quarterly updated (usually accounting based) characteristics have more missing values than monthly updated (usually price based) characteristics
- Last 20 years: similar patterns

Generalized Correlation of Global and Local Factor Weights



Generalized Correlation of Global and Local Factor Weights

- Generalized correlation of constant global Λ with time-varying local Λ^t
- Six-factor model $\Rightarrow$ generalized correlation of 6 means the same span

$\Rightarrow$ Global and local loadings very close

Imputation Error For Different Size Filters

fit	eval	aggregate	quarterly	monthly
$<$ \$ 1 firms	$<$ \$ 1 firms	0.13	0.16	0.08
	$\geq$ \$ 1 firms	0.17	0.17	0.16
	all	0.17	0.17	0.16
$\geq$ \$ 1 firms	$<$ \$ 1 firms	0.38	0.25	0.44
	$\geq$ \$ 1 firms	0.15	0.14	0.15
	all	0.15	0.14	0.15
all	$<$ \$ 1 firms	0.38	0.25	0.44
	$\geq$ \$ 1 firms	0.15	0.14	0.15
	all	0.15	0.14	0.15

Imputation RMSE For Different Size Filters

$\Rightarrow$ Results are robust to size filters

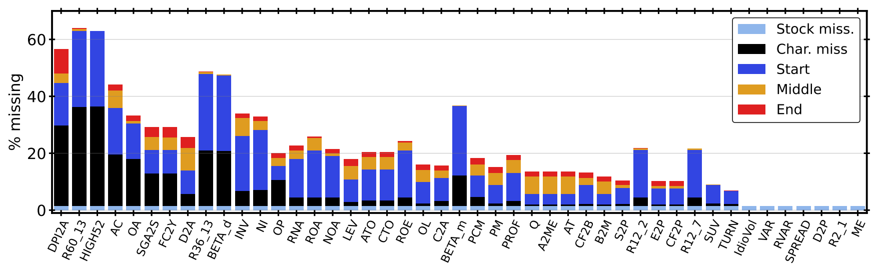
Imputation Results with and without Financial Firms

fit	eval	aggregate	quarterly	monthly
financial firms	financial firms	0.15	0.13	0.16
	non financial firms	0.15	0.13	0.16
non financial firms	financial firms	0.15	0.14	0.16
	non financial firms	0.15	0.14	0.15

Imputation RMSE with and without financial firms

$\Rightarrow$ Results are robust to excluding financial firms

Pooled Mean across Stocks (Equally-weighted)



Start = no previous observations

End = no further observations

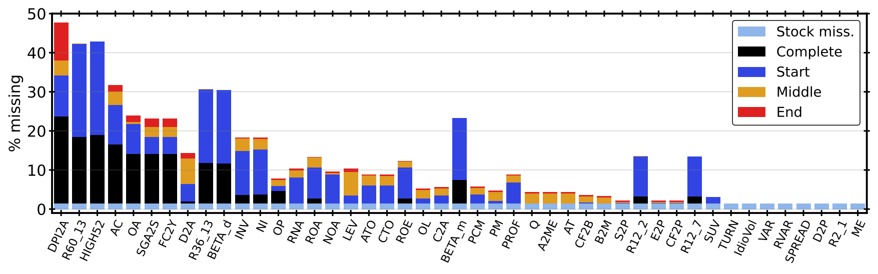
Middle = some previous and future observations

Complete = completely missing

- Some characteristics are mechanically missing for younger firms (e.g., LTrev)
- Many characteristics are missing after having been previously observed
- Some characteristics are missing at the end of the company's life
- Some are never observed

$\Rightarrow$ Imputation needs to allow for different information sets

Pooled Mean across Stocks (Value-weighted)



Start = no previous observations

End = no further observations

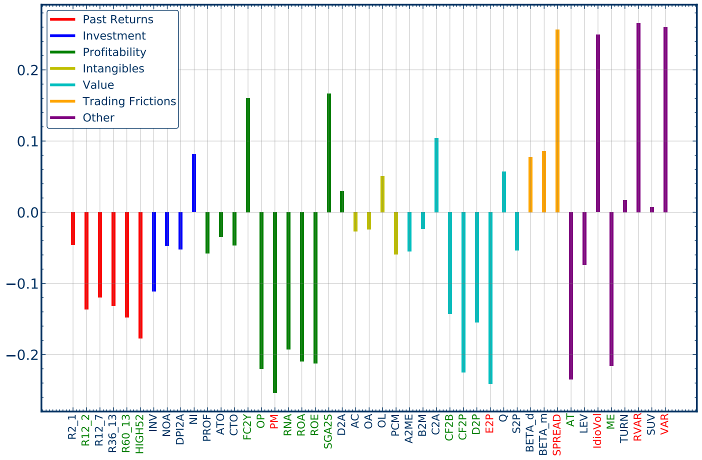
Middle = some previous and future observations

Complete = completely missing

- Some characteristics are mechanically missing for younger firms (e.g., LTrev)
- Many characteristics are missing after having been previously observed
- Some characteristics are missing at the end of the company's life
- Some are never observed

⇒ Imputation needs to allow for different information sets

Composition of Latent Factor 1

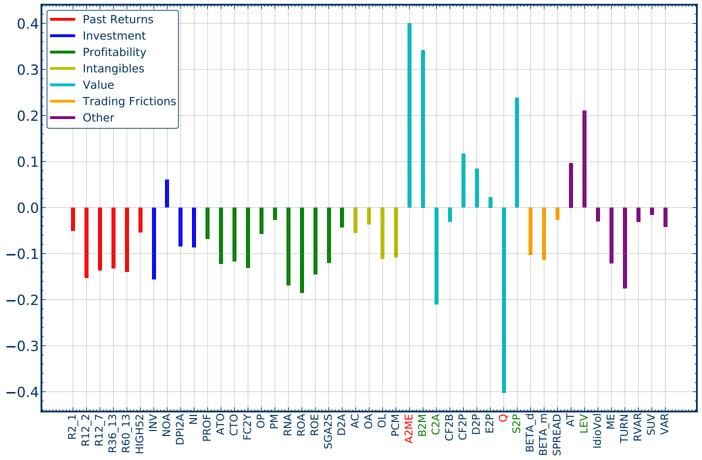


Composition of first latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 1st factor = high volatility characteristics factor

Composition of Latent Factor 2

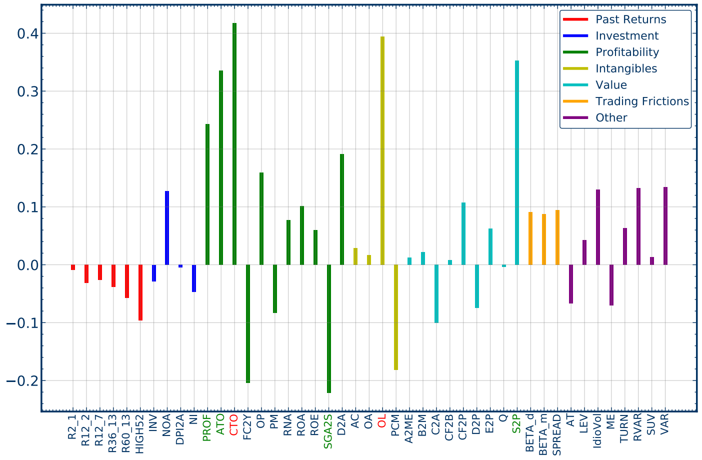


Composition of second latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 2nd factor = value factor

Composition of Latent Factor 3

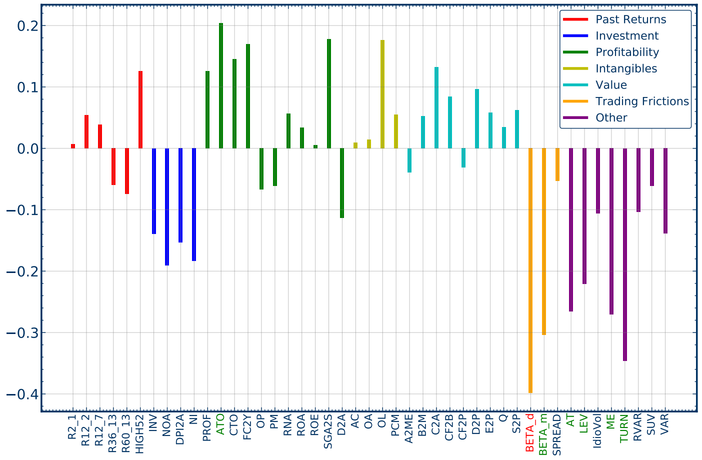


Composition of third latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 3rd factor = profitability factor

Composition of Latent Factor 4

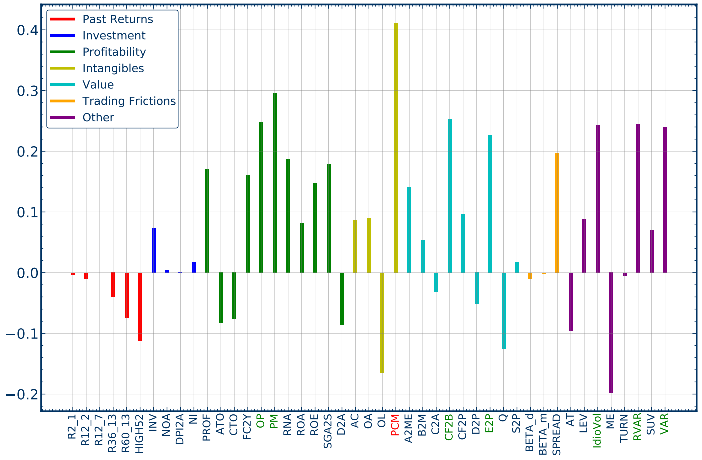


Composition of fourth latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 4th factor = trading friction factor

Composition of Latent Factor 5

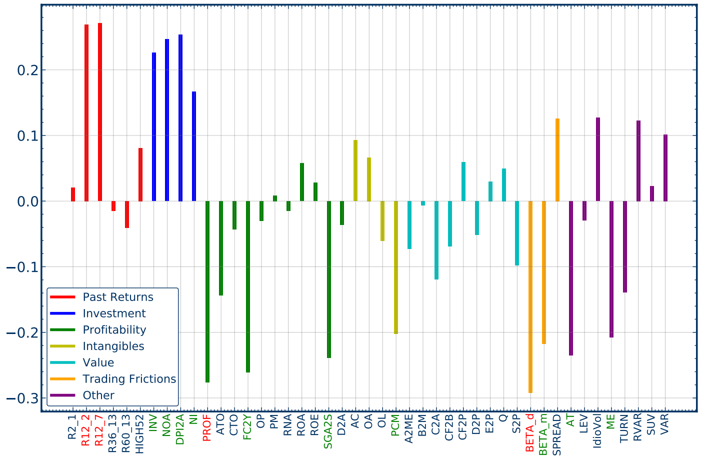


Composition of fifth latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 2nd factor = persistent characteristics factor

Composition of Latent Factor 6

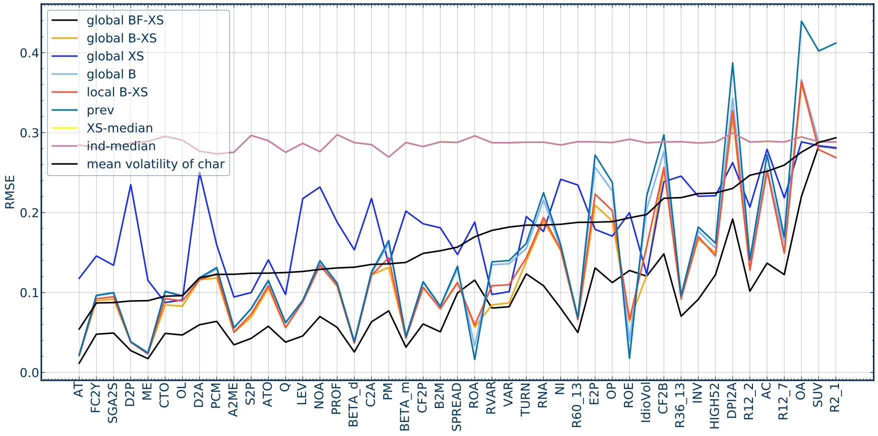


Composition of sixth latent factors by characteristic categories

- The loadings are colored by characteristic category

⇒ 6th factor = long past returns and investment

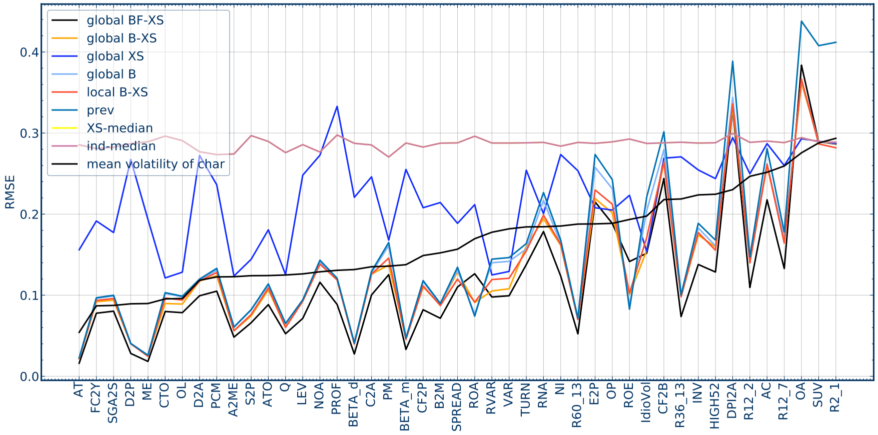
Comparison of imputation methods



In-sample RMSE by imputation method across individual characteristics

- Characteristics are sorted in ascending order based on their volatility (black line)
- Imputation for persistent characteristics benefits from the TS data
- Imputation for more volatile characteristics relies more on XS information

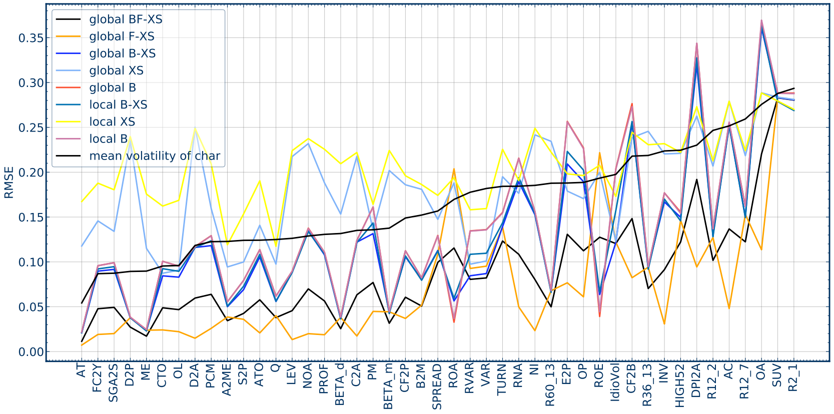
Comparison of imputation methods



Out-of-sample RMSE by imputation method across individual MAR characteristics

- Characteristics are sorted in ascending order based on their volatility (black line)
- Imputation for persistent characteristics benefits from the TS data
- Imputation for more volatile characteristics relies more on XS information

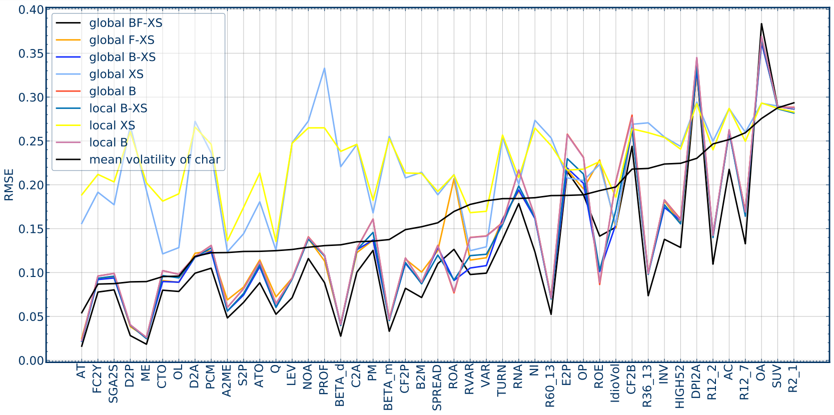
Local vs global imputation



In-sample RMSE by imputation method across individual characteristics

- Global models are slightly better
- Highly volatile characteristics benefit more from local models

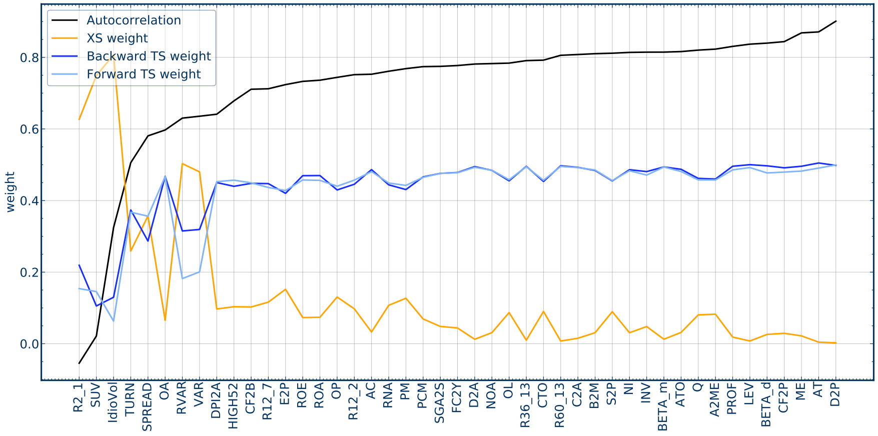
Local vs global imputation



Out-of-sample RMSE by imputation method across individual MAR characteristics

- Global models are slightly better
- Highly volatile characteristics benefit more from local models

Information used for imputation



Relative importance of the TS and XS components (L1 norm)

- Characteristics are sorted in ascending order based on their persistence
- Imputation for persistent characteristics benefits from the TS data
- Imputation for more volatile characteristics relies more on XS information