Missing Financial Data in Brazil

Ramiro Haase[†]

Abstract This study addresses the challenges of missing financial data in Brazil and its implications for asset pricing and corporate finance research. We propose combining information from multiple data sources, to generate a comprehensive dataset of firm characteristics. Our approach consists of using a two-step procedure, that leverages cross-sectional and time-series dependencies, to impute the missing data of three different data sources. After that, we compute the first principal component of a PCA of each firm characteristic to generate our combined dataset. Through an empirical analysis of the Brazilian market data, we demonstrate the effectiveness of our approach in mitigating the impact of missing data. Our findings highlight the importance of considering multiple data sources and implementing robust imputation methods to enhance the reliability and accuracy of financial research in Brazil.

Keywords: Missing data; firm characteristics; asset pricing; PCA.
JEL Code: C22, C38, G12.

1. Introduction

When studying asset pricing using Brazilian data, we face a considerable problem with the availability and quality of data on firm characteristics. Only recently in Brazil, in 2010, it became mandatory to report transactions and events following the International Financial Reporting Standards (IFRS). So, when researchers and practitioners want to use a considerable number of firm characteristics for a lot of stocks and/or a long period of time, missing data becomes a considerable problem that can induce bias in researches. If Asset Pricing and Corporate Finance researchers choose to use only fully observed data they are subject to selection bias as firm characteristics are not missing at random. On the other hand, using simple imputation methods (cross-sectional mean, last observed value) can induce omitted variable bias as firm characteristics depend both on the past and cross-section information (Bryzgalova et al., 2022; Freyberger et al., 2022).

Researchers can access firm characteristics data like balance sheet and income statement information for Brazilian companies from a number of different sources. However, each of these different sources has a number of problems with the supplied data, regarding the availability and quality of this data, as we show later in this paper, which can be a headache for researchers. If those problems are not properly addressed, they can induce bias in research results when using these datasets. Moreover, the data is not consistent across

This draft: April 7, 2024.
[†]São Paulo School of Economics, Fundação Getulio Vargas (FGV), Brazil.
email: ramirochaase@gmail.com

the different data sources. And is not the case that one data source is strictly preferred over the others. Each data source has it's own problems, different patterns of missing, for each firm characteristics. This makes comparing results from papers which use different data sources difficult. This is specially true for the Brazilian case where researchers do not have a consensus on what is the best data source for firm characteristics of public traded companies. In comparison, in the US most researchers use the Compustat data and so results are somewhat more comparable.

Bryzgalova et al. (2022) propose a imputation method designed with financial data in mind. Their method is a expansion on the work of Xiong and Pelger (2023) and consists of a two-step procedure that uses cross-sectional and time-series dependencies in firm characteristics to generate a fully observed firm characteristics data-set that does not impose look-ahead bias in future research using the imputed dataset.

Our novel approach consists of combining information from different sources. In our application of the procedure proposed by Bryzgalova et al. (2022) to Brazilian firm characteristics, we use three data sources, Economatica, Quantum and Compustat. These data sources offer the same firm characteristics but with different problems in the data of each. Each firm characteristic has a different missing pattern and amount of missingness in each of the three data sources. Moreover, none of the three data sources have the same number of companies available. To that end, we propose the combine use of these data sources in our estimations, using information from one to help estimate the other. That means that if we are interested in certain number of firm characteristics, we have three times as many firm characteristics in our data panel, with the same firm characteristics being collected across the three data sources. With that, we greatly increase the the cross-sectional information used in the estimation of our model. After the model is estimated using this merged dataset, we impute the missing data on each of the different data sources, Economatica, Quantum and Compustat. In future applications we can use these imputed datasets in two different ways. We can use one of those imputed datasets directly. Or we can generate a combined imputed data source. The later is our preference, and we propose generating this combined data source using the first principal component of a PCA of each firm characteristic and it self in all the three data sources. To illustrate, suppose we are interest in using Total Assets in some corporate finance or asset pricing research. We would first estimate the imputation model using the the Total Assets of all stocks from the three different data sources, that is, we would have Total Assets three times in our panel, one from each source. Than we should use the estimated model to impute the missing values of Total Assets

in each of the data sources. Finally, we compute the first principal component of the Total Assets with the three different imputed Total Assets. This resulting variable is our Total Asset from the combination of the three imputed data sources.

In this research, we estimate the imputation procedure proposed by Bryzgalova et al. (2022) and performed forecast comparisons of the model with benchmarks to a dataset of nine firm characteristics from Brazilian stocks collected from three different sources. We than use this model to generate a combined dataset that uses information from the three data sources. [1] In Section 1, we presented the Introduction of this paper. In Section 2, we present the two-step imputation method form Bryzgalova et al. (2022). In Section 3, we present our data for the Brazilian market and the empirical strategies applied during the execution of our forecasting exercise. In Section 4, we present and discuss the preliminary results found in this study. Finally, in Section 5, we present the conclusions of this study.

2. Method

Our dataset of quarterly observed, firm characteristics can be represented in the three-dimensional vector:

$$C_{i,t,l}$$

where we have that:

- Cross-section of stocks $i = 1,\ldots,N_t$;[2]
- Time-series $t = 1,\ldots,T$;
- Different characteristics $l = 1,\ldots,L$;

We will use a upper index notation to signal that we are fixing any one of these three dimensions and selecting a two dimensional matrix of the data. As an example, we can select the matrix $C^t_{i,l}$, the $N_t \times L$ matrix of characteristics at time t. Bryzgalova et al. (2022) exploits both the Cross-sectional (XS) dependency and the Time-series (TS) persistence. This in turn allows for general endogenous missing patterns. In the first step of our model, we

[1] We use this resulting imputed data panel in other Asset Pricing research to estimate a modified version of the Stochastic Discount Factor proposed by Andrews and Gonçalves (2020) for the Brazilian market.

[2] The subscript t in the number of stocks, N_t, denotes that we can have a different number of stocks at each time t, as is indeed the case in our sample.

add cross-sectional information in our model by estimating a K-factor model for each t, with $F^t \in \mathbb{R}^{N_t \times K}$ and $\Lambda^t \in \mathbb{R}^{L \times K}$ as follows

$$C^t_{i,l} = F^t_i (\Lambda^t_l)^\top + e^t_{i,l} \tag{1}$$

Without any missing value we would be able to estimate F^t and Λ^t as the eigenvectors of the K largest eigenvalues of the $L \times L$ matrix

$$\frac{1}{N^t} \sum_{i=1}^{N^t} C^t_i (C^t_i)^\top$$

that is, a simple PCA. However, due to the presence of missing data, we follow Xiong and Pelger (2023) and compute $L \times L$ "characteristic covariance matrix" as

$$\hat{\Sigma}^{\text{XS},t}_{l,p} = \frac{1}{Q^t_{l,p}} \sum_{i \in Q^t_{l,p}} C^t_{i,l} C^t_{i,p} \tag{2}$$

Where $Q^t_{l,p}$ is the set of all stocks that are observed for the two characteristics l and p at time t.[3]

The characteristics loadings Λ^t are estimated as the scaled eigenvectors ($\hat{V}^t$) of the K largest eigenvalues ($\hat{D}^t$) of $\hat{\Sigma}^{\text{XS},t}_{l,p}$

$$\hat{\Lambda}^t = \hat{V}^t \left(\hat{D}^t\right)^{1/2} \tag{3}$$

The characteristic factors F^t are estimated from a regularized ridge regression on the loadings

$$\hat{F}^{t,\gamma}_i = \left(\frac{1}{L} \sum_{l=1}^{L} W^t_{i,l} \hat{\Lambda}^t_l \left(\hat{\Lambda}^t_l\right)^\top + \gamma I_K \right)^{-1} \left(\frac{1}{L} \sum_{l=1}^{L} W^t_{i,l} \hat{\Lambda}^t_l C^t_{i,l} \right) \tag{4}$$

Where $W^t_{i,l}$ is 0 if characteristic l is missing for stock i at time t and 1 if it is observed. $\gamma \geqslant 0$ is the regularization parameter. Missing values are imputed with the estimated common component

$$\hat{F}^{t,\gamma}_i \left(\hat{\Lambda}^t_l\right)^\top$$

In the second step of our estimation, we proceed to add the time-series information (TS) of our data-set and combine it with the cross-sectional (XS) estimations of the previous step. In this second step, Bryzgalova et al. (2022)

[3] Note that we do not exclude $l = p$ and that $|Q^t_{l,p}| \leqslant N^t$ by construction.

propose two possible time-series models, a backward cross-sectional model (B-XS), which relies only on past observed information, and a backward forward cross-sectional model (BF-XS), which uses past and future information.[4]

Given our estimates for the cross-sectional characteristics common component, $\hat{F}_i^{t,\gamma}\left(\hat{\Lambda}_l^t\right)^\top$, obtained in the first step, we use a time-series regression to estimate $\beta^{l,t}$ for each combination of firm characteristic l and stock i as:

Backward Cross-Sectional Model (B-XS)

$$\hat{C}_{i,t}^{l,\mathrm{B-XS}} = \left(\hat{\beta}^{l,t,\mathrm{B-XS}}\right)^\top \left(\hat{F}_i^t\left(\hat{\Lambda}_l^t\right)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1}\right) \tag{5}$$

Backward-Forward Cross-Sectional Model (BF-XS)

$$\hat{C}_{i,t}^{l,\mathrm{BF-XS}} = \left(\hat{\beta}^{l,t,\mathrm{BF-XS}}\right)^\top \left(\hat{F}_i^t\left(\hat{\Lambda}_l^t\right)^\top \quad C_{i,l}^{t-1} \quad \hat{e}_{i,l}^{t-1} \quad C_{i,l}^{t+1} \quad \hat{e}_{i,l}^{t+1}\right) \tag{6}$$

Where $\hat{e}_{i,l}^{t-1} = C_{i,l}^{t-1} - \hat{F}_i^{t-1,\gamma}\left(\hat{\Lambda}_l^{t-1}\right)^\top$. Other models are used as benchmark:

Cross-sectional (XS)

$$\hat{C}_{i,t}^{l,XS} = \left(\hat{\beta}^{l,t,\mathrm{XS}}\right)^\top \left(\hat{F}_i^t\left(\hat{\Lambda}_l^t\right)^\top\right)$$

Time Series (B)

$$\hat{C}_{i,t}^{l,B} = \left(\hat{\beta}^{l,t,\mathrm{B}}\right)^\top \left(C_{i,l}^{t+1}\right)$$

Previous Value (PV)

$$\hat{C}_{i,t}^{l,PV} = C_{i,t-1}^{l}$$

Cross-sectional Median (XS-M)

$$\hat{C}_{i,t}^{l,XS-M} = 0^{[5]}$$

We estimated all our models for our sub-sample and we proceed to perform a forecasting exercise to evaluate the performance of these models compared to each other.

[4] The results of the empirical investigations of Bryzgalova et al. (2022) suggests that the BF-XS produces the best forecasting results. This is indeed the case in our results also. Nevertheless, given that we do not want to be at risk of inducing look-ahead bias in our imputation of the missing values, we opted to only use the B-XS model in our data imputation.

[5] We use rank normalized data to the interval $[-0.5, 0.5]$, and so 0 is the cross-sectional median, see Section 3.

3. Data and Empirical Strategies

In our empirical investigation, we will use the procedure proposed in Bryzgalova et al. (2022) to investigate the forecasting capability of different models using firm information from Brazilian stocks. We start this section with a introduction of our dataset. We use as our firm characteristics the following 9 firms fundamentals, which are constructed following Gonçalves (2021):

- $(A_{i,t})$ **Total Assets**.
- $(B_{i,t})$ **Total Book Debt**: Current + Non-current Liabilities.
- $(BE_{i,t})$ **Book Equity** - Davis et al. (2000): Net Equity.
- $(C_{i,t})$ **Cash and Short-term Investment**: Cash and Cash Equivalents + Investment.
- $(E_{i,t})$ **Income Before Extraordinary Items**: Net Income.
- $(GP_{i,t})$ **Gross Profit** - Novy-Marx (2013).
- $(ME_{i,t})$ **Market Equity**.
- $(PO_{i,t})$ **Net Payout** - Boudoukh et al. (2007): ($\sum$ Earnings) $\times$ Outstanding Shares.
- $(Y_{i,t})$ **Total Revenue**: Net Revenue.

Our dataset is collected form three sources, Economatica, Quantum and Compustat. It consists of information of Brazilian stocks firm characteristics with 104 time observations of quarterly data, starting in the first quarter of 1997 up to the final quarter of 2022. Our data was collected giving preference for consolidated information, but when this was not available, we use non-consolidated information. For the appropriate characteristics, information was collected in Brazilian Real, in units with fourteen decimals. We have information on stocks that are/were traded in the B3, and formerly in the BM&FBOVESPA, stock exchanges, but we exclude information on firms of the Financial and Utilities Bovespa Business Sector, following Gonçalves (2021). Data was deflated to prices of the fourth quarter of 2022 using the IPCA (BCB series 433).

The three data sources offer the same 9 firms characteristics, but they are very different in terms of problems with data. To that end, we use all data sources in our estimations, using information from one to help estimate

our imputation method for the others. With that in mind, we end up 27 firm characteristics, the same nine as before but for each of our three datasets. In what follows we will describe our data sources and the problems with each one.[6]

From Quantum, we have 567 individual stocks, this small number of firms, even for the Brazilian market is one of the first problems. We were not able to obtain data from Quantum for much of the companies in Brazil who stop being publicly traded during our sample period. From Economatca we have 1,399 individual stocks, including much of the companies that stopped being publicly traded during our sample for one reason or another. We combine these datasets considering only stocks i that have at least one observation of one characteristic l in at least on time t for both data sources. This amount to 555 individual stocks. Some stocks in our sample are traded for all the spam of our analyzed period, however, most stocks are not. Some stocks start being traded at some time t in the middle of our sample, some stocks stop being traded at some time t in the middle of our sample and finally some stocks start and stop being traded inside our sample period. To make sure that we are only looking at "true" missing values, that is, missing values that occur inside the trading period of a stock, for each stock i, we removed all dates before first non-missing observation and all dates after last non-missing observation. That is, we consider that a stock i started being traded at the first time t for which we have a characteristic l observation and that a stock i stopped being traded at the last time t for which we have a characteristic l observation for it. In Table 1 below, we have the number of stocks and total lines of information we would have if we chose to work with only fully observed data instead of applying the chosen imputation method. We would lose more than 20% of the number of different stocks and more than 50% of the total information.

Table 1
Amount of Stocks

Source	Stocks	Stocks*	%	Total Lines	Total Lines*	%
Quantum	555	437	0.79	44,232	21,075	0.48
Economatica	555	436	0.79	41,925	15,472	0.37

Number of stocks and total lines of stock i time t information when using the proposed imputation method versus when using only fully observed data (marked as *).

In Figure 1 below we have the evolution in time of the amount of stocks in each of our sources and in our combined sample. We can see a clear upwards trend in the number of stocks of each sample, it stars with a small number of

[6]Data from Compustat was just recently added and will be commented on the next draft.

stocks, and it peaks close to the end of our sample. This is expected given the development experienced by the Brazilian capital market after 1994. However, we also see that the number of stocks fluctuates considerably. This is because for some firms in our sample, especially before 2011, we only observe some characteristics at their balance sheets at one quarter per year. We can see that even towards the end of our sample, there are still a lot of stocks for which firms we do not observe at all quarters.

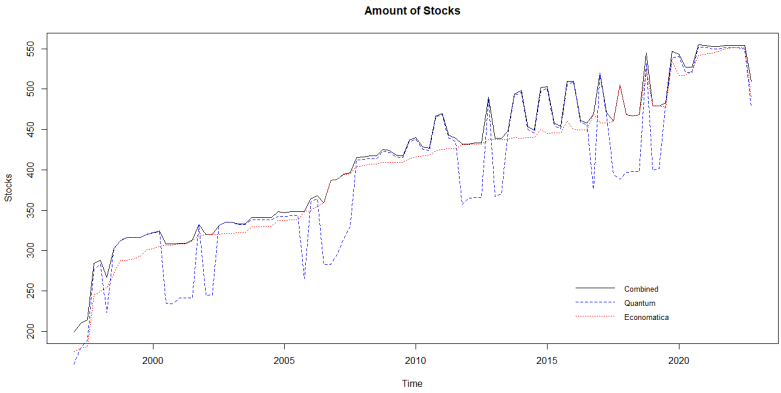

Figure 1
Amount of Stocks

In Figure 2 below, we show that the two sources of data have different patterns of missingness in our analyzed period. The data from Economatica shows a downward trend in the amount of missing for most characteristics, except for $C_{i,t}$ (Cash and Short-term Investment), which is missing for all firms in our sample until 2010. We have a considerable amount of firms for which we do not observe some of the characteristics at all quarters of the year. This is evident by the frequent dips in the missing percentage of some characteristics up until 2010. The fluctuation in the amount of missing is greatly reduced after 2011 for mos characteristics and we finally have the first observations of $C_{i,t}$. These unwanted problems are reduced in the final stretch of sample because in 2010 we have the mandatory full adoption in Brazil of the International Financial Reporting Standards (IFRS), which improved the quality of the financial reports provided by firms in Brazil. Data from Quantum fluctuates much more, with some characteristics having no observed values for any stocks at some quarters, and the missing patterns do not appear to improve over time.

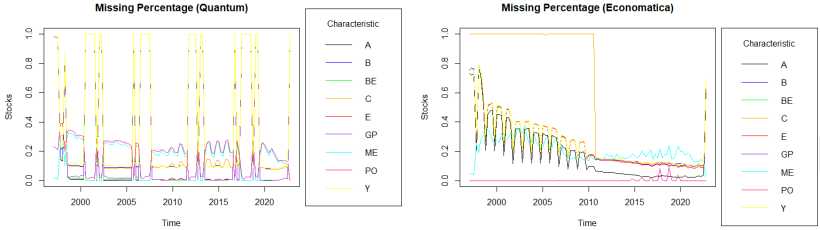

Figure 2
How much data is missing in each source

In Table 2 below we have, by characteristic and data source the amount of missing observations in our dataset. We observe that in general, the amount of missing data per characteristic is lower in the Economatica data than in the Quantum data.

Table 2
Missing Data

Characteristic	Quantum	Economatica
A	0.31	0.16
B	0.38	0.24
BE	0.31	0.23
C	0.38	0.52
E	0.32	0.23
GP	0.32	0.23
ME	0.23	0.21
PO	0.25	0.01
Y	0.38	0.24

Ratio of missing observations of each characteristic and data source.

In Figure 3 below we have a different visualization of the missing patterns of our data. In each quadrant of Figure 3, we have a matrix that represent a quarter in our sample, with stocks as rows and firms characteristics as columns. The Quantum data is located in the right column and the Economatica in the left column. In these matrices, if we observe the firm characteristic for one stock, we fill that row/column combination with 1, if otherwise we are missing the information for that stock-firm characteristic combination, we fill the respective spot with 0. We can see that characteristics are not missing at random, with some stocks and some firms being more prone to have missing data.

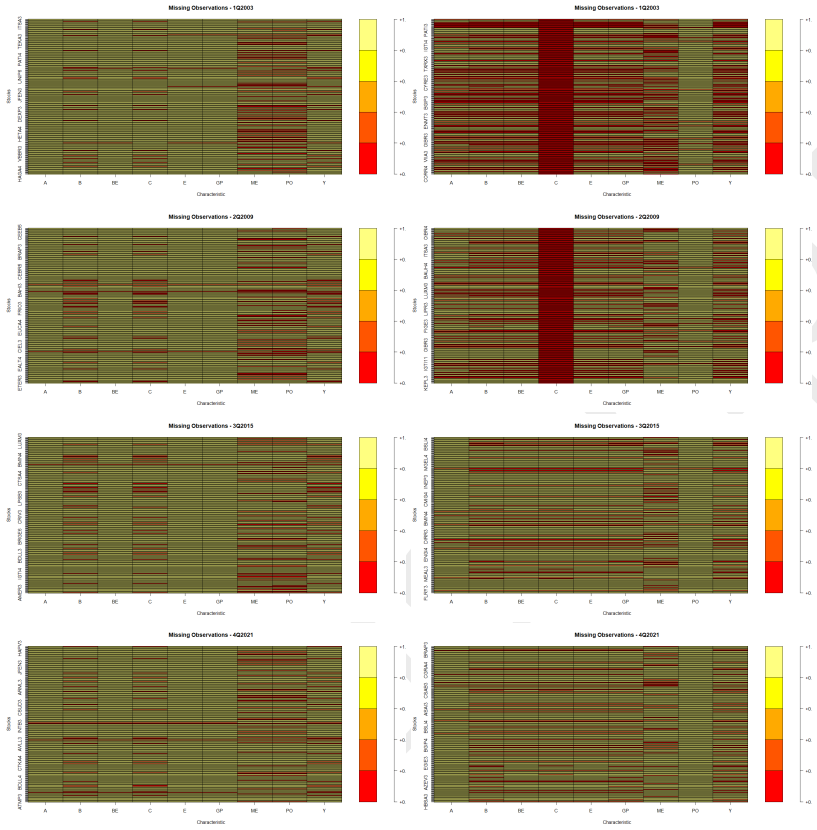

Figure 3
Missing Pattern Quantum/Economatica

We would also like to highlight another problem when using these data sources "straight off the shelf" to study Brazilian firms. In Figure 4 bellow we have the evolution of each characteristic l for one single stock i from, the pharmacy chain RaiaDrogasil SA (RADL3). We can see that, with some missingness, we have data for this stock from the Quantum data source since the late 90s. From Economatica we only start observing some characteristics in 2014, while ME (Market Equity) and PO (Net Payout) are observed, like Quantum, from the late 90s. Also, most characteristics for this stock seem to go through a jump in level around 2010. Investigating further we found that two different companies Raia and Drogasil, merged at the second semester of 2011 with The Administrative Council of Economic Defense (CADE) unan-

imously approving the merge in the first quarter of 2012. It appears that the Quantum data source uses one of the firms as base for the data in quarters before the merge and creation of the RADL3 ticker, while Economatica only does this for two characteristics. While this is just one anecdotal evidence of more problems with these data sources, we must think of strategies to deal with this kind of situation.

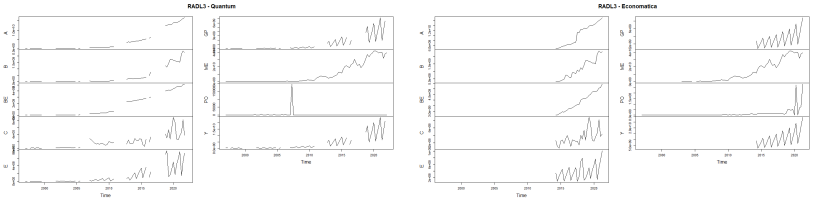

Figure 4
Raia Drogasil SA (RADL3)

We estimate $\hat{\beta}^{l,t}$ by OLS using stacked observed values. That is, we use all $C^l_{i,t}$ with observed $C^l_{i,t-1}$ for our B-XS model. Prior to estimation, we divide our stacked data into in-sample (IS) and out-sample (OS). In our in-sample, we mask, randomly, 10% of the observations of $C^l_{i,t}$ before we estimate our models. We mask these observations in two different ways, completely at random and in random blocks of four consecutive quarters. We than predict these masked observations in our out-of-sample analysis. In Tables 3 and 4 bellow we have the number of masked observations compared to the total number of stacked observations, for each data source.

Table 3
Random In and Out samples

	A	B	BE	C	E	GP	ME	PO	Y
Quantum	25,641	23,295	25,598	23,343	25,486	25,473	33,377	32,699	23,193
10%	2,564	2,330	2,560	2,334	2,549	2,547	3,338	3,270	2,319
Economatica	33,527	30,384	30,849	19,455	30,824	30,575	31,707	40,797	30,364
10%	3,353	3,038	3,085	1,946	3,082	3,058	3,171	4,080	3,036

Table 4
Block Missing In and Out samples

	A	B	BE	C	E	GP	ME	PO	Y
Quantum	25641	23295	25598	23343	25486	25473	33377	32699	23193
10 %	2568	2332	2564	2336	2552	2548	3340	3272	2320
Economatica	33527	30384	30849	19455	30824	30575	31707	40797	30364
10 %	3356	3040	3088	1948	3084	3060	3172	4084	3040

Remember, the B-XS model uses the most information while avoiding any look-ahead bias and the the BF-XS model uses the most information overall,

but it induces look-ahead bias in the imputation. Moreover, the BF-XS can not be estimated on this previously discussed in-sample as it was constructed considering all $C^l_{i,t}$ with observed $C^l_{i,t-1}$ only. Next we present the sample suited for the estimation of the BF-XS model, using all $C^l_{i,t}$ with observed $C^l_{i,t-1}$ and $C^l_{i,t+1}$. Obviously this sample is smaller than the previous one. This means that although the BF-XS model uses more information, compared to the B-XS model, since it includes $C^l_{i,t+1}$ in its estimation, there is a small loss of information due to the smaller sample size, given that the B-XS uses all $C^l_{i,t}$ with observed $C^l_{i,t-1}$ and the BF-XS model requires $C^l_{i,t}$ with observed $C^l_{i,t-1}$ and $C^l_{i,t+1}$.

Table 5
Random In and Out samples

	A	B	BE	C	E	GP	ME	PO	Y
Quantum	22259	20225	22225	20360	22131	22131	32734	32221	20132
10%	2226	2022	2222	2036	2213	2213	3273	3222	2013
Economatica	32903	29813	30262	18929	30232	29996	30598	40242	29790
10%	3290	2981	3026	1893	3023	3000	3060	4024	2979

Both the B-XS and BF-XS models can be estimated in the samples presented in Table 5, however only the B-XS model can be estimated in the samples presented in Tables 3 and 4. We argue that this is one more reason why the better model for our application is the B-XS model. It does not induce look-ahead bias and it uses the most $C^l_{i,t}$ in the estimation process.

We evaluate our different models based on their out-of-sample RMSE (root-mean-squared error). We consider the RMSE for each characteristic separately as

$$RMSE_l = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2} \tag{7}$$

We also consider the RMSE averaged over all stocks, time periods and characteristics as

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}\left(C_{i,t,l} - \hat{C}_{i,t,l}\right)^2} \tag{8}$$

We then compute the modified Diebold-Mariano test form Harvey et al. (1997) and the model confidence set (MCS) from Hansen et al. (2011) to compare the forecast accuracy of our models and identify our best forecasting model.

Finally, we performed a couple of transformations on our data before we start estimation of our model, which we describe next. We tried to manually

pre-impute some missing data in Outstanding Shares before computing PO (Net Payout), as the number of Outstanding Shares of a firm is not expected to vary much from quarter to quarter. The idea being that if we have one (up to three) missing values where the previous and next observed values where the same, than it is likely that the number of Outstanding Shares did not vary for the one (up to three) quarter that we have missing. For that we used the following algorithm to try to impute this cases of missing data:

1. If $C^l_{i,t-1} = C^l_{i,t+1}$, then $\hat{C}^l_{i,t} = C^l_{i,t-1} = C^l_{i,t+1}$
2. If $C^l_{i,t-1} = C^l_{i,t+2}$, then $\hat{C}^l_{i,t} = C^l_{i,t-1} = C^l_{i,t+2}$
3. If $C^l_{i,t-1} = C^l_{i,t+3}$, then $\hat{C}^l_{i,t} = C^l_{i,t-1} = C^l_{i,t+3}$

No cases were found in any of our two data sources and so we did not imputed data this way. Following Gonçalves (2021), we set as missing any strange values in our characteristics.[7] Finally, we apply to our data a rank normalization to the interval $[-0.5, 0.5]$, with 0 being the cross-sectional median.

$$C^{t\ rn}_{i,l} = \frac{C^{t\ r}_{i,l} - \min\left(C^{t\ r}_{l}\right)}{\max\left(C^{t\ r}_{l}\right) - \min\left(C^{t\ r}_{l}\right)} - 0.5 \tag{9}$$

After imputing the data as ranked normalized characteristics based on the forecasts of our model we can map the rank normalized data back to raw characteristics simply using the empirical density function of each characteristic as suggested by Bryzgalova et al. (2022). However, this implies that we only obtain imputed raw values that were already in the distribution of raw characteristics before the imputation. We can improve the accuracy of our final dataset by using some sort of interpolation to obtain raw characteristics from their rank normalized counterpart. Preliminary results of our estimations are presented and discussed in the next Section.

4. Results

To start our results discussion we look at the number of K latent factors to be used in the model discussed in Section 2. In Figure 5 bellow we see that the explained variance starts at around 40% if we use one factor and rapidly goes to close to 90% with 7 factors. We also see that the average magnitude of each factor rapidly decreases and stabilizes at around 10 factors. Next we

[7] Gonçalves (2021) sets any non-positive A , BE, ME, and Y; any negative B and C; any B, BE, and C higher than A; and any BE higher than 50 × ME or below (1/50) × ME to missing.

present the results for models that include 7 and 10 factor, but we tested all models that include from 1 through 10 factors.

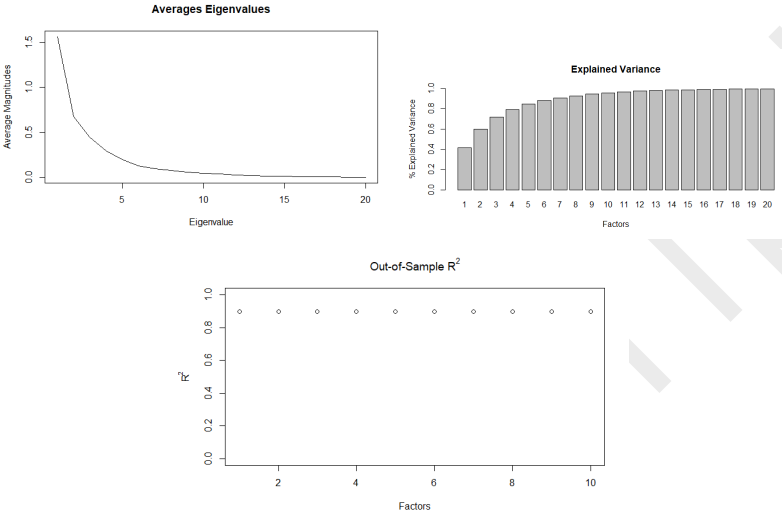

Figure 5
Number of Latent Factors

We applied the procedure presented in Section 2 to our data presented in Section 3 and performed forecasts for each of our nine firm characteristics. We evaluate seven competing models using their estimated RMSE. In Table 6 below we have the $RMSE_l$ from the forecasts generated by our models for each of the nine firm characteristics averaged over all stocks and time periods.

Table 6
Out-of-Sample $RMSE_l$

Quantum	A	B	BE	C	E	GP	ME	PO	Y
B-XS 7	0.039136	0.083950	0.034136	0.095276	0.150669	0.098898	0.036256	0.093570	0.072901
B-XS 10	0.039131	0.083907	0.034153	0.095107	0.150598	0.098959	0.036241	0.093529	0.072889
XS 7	0.273715	0.276298	0.279531	0.280228	0.287138	0.276689	0.277190	0.270820	0.271434
XS 10	0.273656	0.276325	0.279437	0.279740	0.287198	0.276707	0.277073	0.270837	0.271212
B	0.039269	0.084874	0.034161	0.096078	0.151027	0.099025	0.036223	0.094268	0.072949
PV	0.039408	0.085708	0.034194	0.097253	0.156755	0.100954	0.036252	0.095672	0.073706
XS M	0.287799	0.289947	0.290857	0.291383	0.292547	0.285970	0.288467	0.290031	0.284187
Econometrica	A	B	BE	C	E	GP	ME	PO	Y
B-XS 7	0.025056	0.042264	0.043673	0.067529	0.118322	0.059356	0.040410	0.241529	0.038098
B-XS 10	0.025050	0.042237	0.043659	0.067568	0.118210	0.059346	0.040388	0.241572	0.038042
XS 7	0.277789	0.270622	0.273989	0.275918	0.278848	0.278739	0.274847	0.278310	0.271206
XS 10	0.277753	0.270588	0.273675	0.275604	0.278798	0.278620	0.274984	0.278347	0.271158
B	0.025056	0.042298	0.043726	0.067669	0.118926	0.059343	0.040480	0.249970	0.038078
PV	0.025080	0.042406	0.043832	0.068087	0.121171	0.059660	0.040568	0.287721	0.038182
XS M	0.291280	0.285320	0.289034	0.283710	0.290574	0.291485	0.289914	0.289905	0.283676

First, we see that for all nine characteristics including 7 or 10 factors in the B-XS model affects little the forecasting results of this model. This is confirmed by our Diebold-Mariano test results in Table 8. Given that there is little to no gain in accuracy including more factors after the 7th, when using this model to impute data, one should use the B-XS model with 7 factors to be more parsimonious. We can also point that there is no clear best model across all characteristics in this first analysis. Both the B and PV models get forecasting errors really close to our B-XS model with 7 factors. The worst $RMSE_l$ overall were those of the XS-M, which uses only the cross-sectional mean in it's forecasts. We can also see that the characteristics Net Payout (PO) and Income Before Extraordinary Items (E) had the higher $RMSE_l$ across all models. In Table 7 below we have the *RMSE* for the forecasts of each model averaged over all stocks, time-periods and characteristics.

Table 7
Out-of-Sample *RMSE*

	Combined	Quantum	Economatica
B-XS 7	0.076724	0.078310	0.075137
B-XS 10	0.076699	0.078280	0.075119
XS 7	0.276295	0.277005	0.275585
XS 10	0.276206	0.276909	0.275503
B	0.077412	0.078653	0.076172
PV	0.080367	0.079989	0.080745
XS-M	0.288672	0.289021	0.288322

The B-XS model with 10 factors appears to produce forecasts not much better than the model with 7 factors. We see than the B and PV models next, with forecasting errors a bit larger. We can also highlight that all models except the PV generate better forecasts for the Quantum data source than the Economatica. Next, to properly evaluate which model is better given that some of the errors are very close to each other, we present in Table 8 bellow the results of our Diebold-Mariano tests comparing the forecasts of different models.

Table 8
Diebold-Mariano

Quantum	A	B	BE	C	E	GP	ME	PO	Y
B-XS 7 vs B-XS 10	0.170741	1.254085	-0.599922	2.491869	1.099965	-1.037857	0.978199	1.102476	0.552736
	0.864441	0.209937	0.548612	0.012776	0.271451	0.299435	0.328047	0.270336	0.580497
B-XS 7 vs XS 7	-43.416239	-36.298144	-49.044506	-38.238004	-27.965464	-34.945822	-53.522991	-36.555401	-38.124811
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
B-XS 7 vs B	-1.885495	-2.560661	-0.564128	-2.901200	-1.075830	-0.522230	1.095630	-1.573462	-0.196709
	0.059476	0.010510	0.572716	0.003752	0.282105	0.601555	0.273320	0.115709	0.844073
B-XS 7 vs PV	-2.628182	-2.954735	-0.672942	-4.018098	-5.442624	-3.809944	0.085123	-2.911976	-2.014358
	0.008635	0.003161	0.501045	0.000061	0.000000	0.000142	0.932169	0.003616	0.044087
Economatica	A	B	BE	C	E	GP	ME	PO	Y
B-XS 7 vs B-XS 10	0.442614	0.774647	1.136452	-0.580190	0.810233	0.156933	1.325605	-0.894015	1.302795
	0.658074	0.438608	0.255856	0.561854	0.417869	0.875308	0.185066	0.371367	0.192743
B-XS 7 vs XS 7	-45.488212	-42.908684	-50.144993	-34.577332	-38.605900	-45.699407	-52.179380	-14.281442	-44.401494
	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
B-XS 7 vs B	-0.033451	-1.271835	-1.709706	-1.870174	-3.041846	0.473157	-1.730569	-6.314807	0.756429
	0.973317	0.203529	0.087421	0.061610	0.002371	0.636135	0.083626	0.000000	0.449451
B-XS 7 vs PV	-0.896222	-2.234268	-2.062772	-3.420992	-5.718850	-3.040080	-3.063063	-14.319983	-1.372975
	0.370199	0.025538	0.039218	0.000637	0.000000	0.002385	0.002209	0.000000	0.169861

We can see that overall the forecasting errors from the B-XS model with 7 and 10 factors are statistically not different. The B-XS model, which includes both past and cross-sectional information, is strictly better than the XS, which includes only cross-sectional information, with the same 7 factors. The analysis of which model is best gets unclear when comparing the B-XS to the B and PV models. For quite a few characteristics we can not say that the the forecasting results of the B-XS model are statistically different than the other two models.

Finally, while we follow Bryzgalova et al. (2022) and use rank normalized data, we opted for a different method of mapping this rank normalized data back to raw characteristics data using interpolation to obtain the final raw imputed data.

5. Conclusion

In this study we presented a few interesting (and concerning) aspects of data from Brazilian firm fundamentals from different sources of data. We analyzed the prevalence of missing data in nine characteristics, the same ones used by Gonçalves (2021) in his application to the US market. We find evidence that the adoption of the IFRS in 2010 had a considerable result on the quality of the firm fundamentals data in Brazil. We applied the novel imputation method form Bryzgalova et al. (2022) to a Brazilian data-set and performed forecasts comparisons with other benchmark models. Finally we propose the use of three different data sources, Quantum, Economatica and Compustat to generate one combined Brazilian firm characteristic dataset using the chosen imputation method. We are hopeful that this novel dataset can improve the quality of analysis in both Asset Pricing and Corporate Finance research using this kind of data due to the increased volume and quality of the resulting dataset compared to the initial three datasets. Using this imputed

dataset reduces the selection bias of using only fully observed data from one single data source, reduces the omitted variable bias of naive imputed datasets while also not incurring in look-ahead bias.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Andrews, S. and Gonçalves, A. (2020). The bond, equity, and real estate term structures, *Kenan Institute of Private Enterprise Research Paper Forthcoming* .

Boudoukh, J., Michaely, R., Richardson, M. and Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing, *The Journal of Finance* **62**(2): 877–915.

Bryzgalova, S., Lerner, S., Lettau, M. and Pelger, M. (2022). Missing financial data, *Available at SSRN* **4106794**.

Davis, J. L., Fama, E. F. and French, K. R. (2000). Characteristics, covariances, and average returns: 1929 to 1997, *The Journal of Finance* **55**(1): 389–406.

Freyberger, J., Höppner, B., Neuhierl, A. and Weber, M. (2022). Missing data in asset pricing panels, *Available at NBER* **30761**.

Gonçalves, A. S. (2021). The short duration premium, *Journal of Financial economics* **141**(3): 919–945.

Hansen, P. R., Lunde, A. and Nason, J. M. (2011). The model confidence set, *Econometrica* **79**(2): 453–497.

Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors, *International Journal of forecasting* **13**(2): 281–291.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium, *Journal of financial economics* **108**(1): 1–28.

Xiong, R. and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference, *Journal of Econometrics* **233**(1): 271–301.