BovDB: A data set of stock quotes for Machine Learning on all companies from B3 between 1995 and 2020

Fabian Corrêa Cardoso[1], Juan Malska[2], Paulo Ramiro[2], Giancarlo Lucca[1], Eduardo N. Borges[2], Viviane de Mattos[3], Rafael Berri[2]

[1]Programa de Pós-Graduação em Modelagem Computacional (PPGMC)
Universidade Federal do Rio Grande (FURG), Rio Grande – RS – Brazil

[2]Centro de Ciências Computacionais (C3)
Universidade Federal do Rio Grande (FURG), Rio Grande – RS – Brazil

[3]Instituto de Matemática, Estatística e Física (IMEF)
Universidade Federal do Rio Grande (FURG), Rio Grande – RS – Brazil

{fabiancorrea, juanandreyvmalska, paulojr2016canaa, giancarlo.lucca, eduardoborges, vivianemattos, rafaelberri}@furg.br

Abstract. *Stock markets are responsible for the movement of huge amounts of financial resources around the world. This market generates a high volume of transaction data, which after being analyzed are very useful for many applications. In this paper we present BovDB, a data set that was built considering as source the Brazilian Stock Exchange (B3) with information related to the years between 1995 and 2020. We have approached the events' impact on the stocks by applying a cumulative factor to correct prices. The results were compared with public data from InfoMoney and BR Investing, showing that our methods are valid and in accordance with the market standards. BovDB data set can be used as a benchmark for different applications and is publicly available for any researcher on GitHub.*

1. Introduction

Stocks are securities that represent properties of a company by the shareholders [Wang 2021]. The Stock Exchange (SE) allows companies to raise financial resources in exchange for the sale of shares and corporate bonds. Thus, a stock market is where buying and selling transactions take place.

There are different stock markets around the world. We highlight the NYSE (New York Stock Exchange), the Chinese SSE (Shanghai Stock Exchange), the SZSE (ShenZhen Stock Exchange) and the B3 (Brasil, Bolsa, Balcão)[1]. The latter is the market that trades stocks of companies in Brazil.

The stock markets are responsible for the movement of huge amounts of financial resources around the world [Harris 1997]. Due to the immense number of transactions performed by a SE, there is a large generation of data about the stocks available for trading. An important question that arises is related to the amount of data generated by these transactions and how to make it available to the public.

[1]Access the B3 website at: `http://www.b3.com.br`

The data on Brazilian stocks are made available digitally to the public. It is worth pointing out that this data can be accessed daily on B3's website[2]. Among the information available we highlight: short company name, currency used, opening price, closing price, lowest price, highest price and volume traded[3].

Data are provided in text files, separated by year and in their raw form (without any formatting). The two aspects pointed out to make it difficult to extract knowledge from this data and to understand the distortions caused by the events on the stock prices. As consequence, it may be difficult to understand and it may present a high complexity for people not connected to the financial market.

It is easily noticeable that the use of this data is of utmost importance and crucial for the understanding and study of stock markets. Therefore, the number of studies using stock data presented in the literature is enormous, and it is clear to see the importance of this area. For different case studies found in the literature, data sets of the most diverse types are used, there are several data sources and time windows of analysis are quite variable, which can range from minutes to decades [Sezer et al. 2020].

Operations with stocks are classified according to the execution period. When they are executed on the same day (minutes or hours) they are short-term or daytrade. When they are executed in more than a day, a week or a few months they are medium-term (swing trade). When they are executed after many months, years or even decades (rare) they are long-term (buy and hold). Generally, daytrade and swing trade are linked to technical analysis and buy and hold are linked to fundamental analysis [Vachhani et al. 2019].

To assist research in finance, we sought to make available the daily stock data of all companies listed on B3 between the years 1995 and 2020, presenting this data preprocessed ready for data mining. We provide the data set called BovDB and make it available to serve as a benchmark for future applications, such as stock price forecasting, among others.

This paper is organized as follows. Section 2 presents the related work. Section 3 discusses possible applications for the database. Sections 4, 5, and 6 discuss in detail the processing performed, challenges of creating BovDB and the description of the database, respectively. As mentioned, the data set is publicly available and the full reference is available in Section 7. The conclusions of the paper are in Section 8.

2. Related Work

Pre-processed public data sets on stocks of several SE around the world can be freely found on specialized sites such as InfoMoney[4] or Yahoo Finance[5]. For those using the R programming language, Yahoo Finance can provide data with a simple download command, within the code itself. However, some data was noticed to be missing from this data set for B3 stock values. The free and public existence of a pre-processed academic data set specific to Brazilian stocks is not of our knowledge.

The data provided by sites like those cited are not raw, but pre-processed with

[2] Download the B3 historical series raw data at: `https://syr.us/tAQ`

[3] For which data are provided in the files made available by B3, access: `https://syr.us/79e`

[4] Available at: `https://www.infomoney.com.br`

[5] Available at: `https://finance.yahoo.com`

adjustments based on price changes for events that do not occur during normal trading, and there is no transparency in how these adjustments are made. Besides, they make data available in Microsoft Excel spreadsheets, CSV or even TXT files, for instance, and, sometimes, the range date someone can select is limited, specially for over a decade. On the other hand, in Kaggle[6], looking for stocks, is possible find 983 data sets that make available data from the most varied types and places. So, to make calculations and analyses, the academic researchers need to look for, download, and create theirself the database, to make it able for use.

Among some related works, we can cite [Efimov et al. 2020] that use data sets from American Express for risk modeling. They also list data sets available "which are used to benchmark and validate Machine Learning algorithms developed by various researchers, academic groups and companies". The authors assumed that their model of Generative Adversarial Networks replicated with good accuracy the relationship between the target variables and the data features.

In [Guo et al. 2018], a Shanghai Stock Exchange data set was used for intraday analysis using the adaptive Support Vector Machine Regression (SVR) method for high frequency stock price prediction by 5-min, 30-min and daily basis. This study suggests that "the improved SVR with dynamic optimization of learning parameters and particle swarm optimization can get a better result than other compared methods including SVR and back-propagation neural network".

An analysis of the features used for forecasting was done by del Angel [Del Ángel 2020]. Precisely, this study considers the closing price for time series forecasting. The author provided a comparison between Backpropagation and Resilient Backpropagation Machine Learning algorithms, having used stock market indices from Europe, Asia and North America in the period from 2010 to 2019. "Instead of prediction itself, the scientific objective was to evaluate the relative importance of characteristic variables that allow prediction" [Del Ángel 2020]. An analysis of the features used for forecasting was done.

The study provided by Sowinska and Madhyastha uses data set of text from Twitter as samples and the stock return information as labels to predict the impact on stocks with 4 labels (one, two, three and four-day returns). The authors allow the download of their scripts and data sets from the internet on GitHub. Moreover, according to the authors [Sowinska and Madhyastha 2020], this study is "well-suited for building models for long-term, fundamental investing".

Having into consideration the data set provided by the Yelp Data set Challenge[7] in [Rafay et al, 2020], different prediction models based on Machine and Deep Learning where applied. They chose the 100-dimensional vector of pre-trained Global Vector (GloVe) word embeddings of the Yelp Dataset available on Kaggle website[8] to classify. The authors stated that the best classifier for binary and multi-class classification was C-LSTM obtaining overall good scores.

Among the data sets found, not only in the above mentioned articles but also in

[6]To look for the stock data sets, access: `https://www.kaggle.com/datasets`

[7]For more information about this data set, access: `https://www.yelp.com/dataset`

[8]Access this data set at: `https://www.kaggle.com/yelp-dataset/yelp-dataset`

others, it is very common to observe that the calculations with stocks are usually performed mainly with the daily closing value of the stock. Although, some work use the daily opening values, the highest/lowest values per day, the traded volumes per day or the stock indexes [Sezer et al. 2020].

Find works that use stock market data, although it is not ordinary find their data sets available on the internet. So, there are related works that make calculations with data sets of stock exchange data, but few of them provide their data sets (even when the study is granted by public resources). In the work presented in this paper, we pre-process this data and make it available for free access, to allow the use of the entire time series of stocks and to minimize the interference of events on B3's assets.

3. Application

The stock data from the market could be used for academic researchers to make analysis [Nti et al. 2019], comparisons [Rahat et al. 2019], forecasts [Bustos and Pomares-Quimbaya 2020], etc. Making the stock data available is a way to encourage researches in this field. Thus, the great advantage is that it is supplied a pre-processed, reliable and organized information in a data set. It can be used as the basis of many applications. This section aims to describe some of them.

Applications with these data set can be made for adjustment and validation of statistical forecasting of stock time series [Alhnaity and Abbod 2020], for neural network classification of stock buy/sell transactions [Schierholt and Dagli 1996], for technical analysis of stocks and investments [Rousis and Papathanasiou 2018].

For applications of statistical analysis, investigation on the data generating mechanism could be studied, characterizing the behavior of a series with the identification of periodicities, for example [Thomaz et al. 2021]. This would allow reliable and accurate predictions of stock performance, see, for example [Zhang 2021]. Also, the application of statistical regression, which seeks to predict future stock price values [Upadhyay et al. 2012].

In the financial area it is possible to forecast lots of scenarios using these data set, for instance, calculations of investment risk [Basak et al. 2019], predictions of stocks [Hu et al. 2021] [Thomaz et al. 2021] and technical analysis [Li and Bastos 2020] of stocks. The use of neural networks as stock classifiers makes it possible to predict the behavior of one or more stocks and suggest the purchase or sale of a given asset at a given moment, indicating a high/moderate uptrend, stability, or moderate/ sharp drop.

4. Data Set Design

The data set is stored in a relational database management system (RDBMS) [Garcia-Molina 2008] know as SQLite [Allen and Owens 2010], which is written in C programming language. In contrast to many other database management systems, SQLite is not a client–server database engine. Rather, it is embedded into the end program. Additionally, it does not have any prerequisites nor it requires to be downloaded, which is why we decided to use it, since we will be using it locally.

Precisely, our data set is composed by 5 different tables, which are:

- **Company** - Stores data referring to companies;

- **Ticker** - Stores data of a given stock;
- **Price** - Stores the trading data of a specific stock on a specific date;
- **Event** - Stores information on the different event types;
- **EventPrice** - Stores data of a specific event on a specific stock on a given date.

In Figure 1 we provide the schema of the database. It shows the relationships among tables, their fields, Primary Key (PK), Foreign Keys (FK) and data types that can be textual (text), numerical (real or integer) or date. Some fields have a Unique integrity constraint.

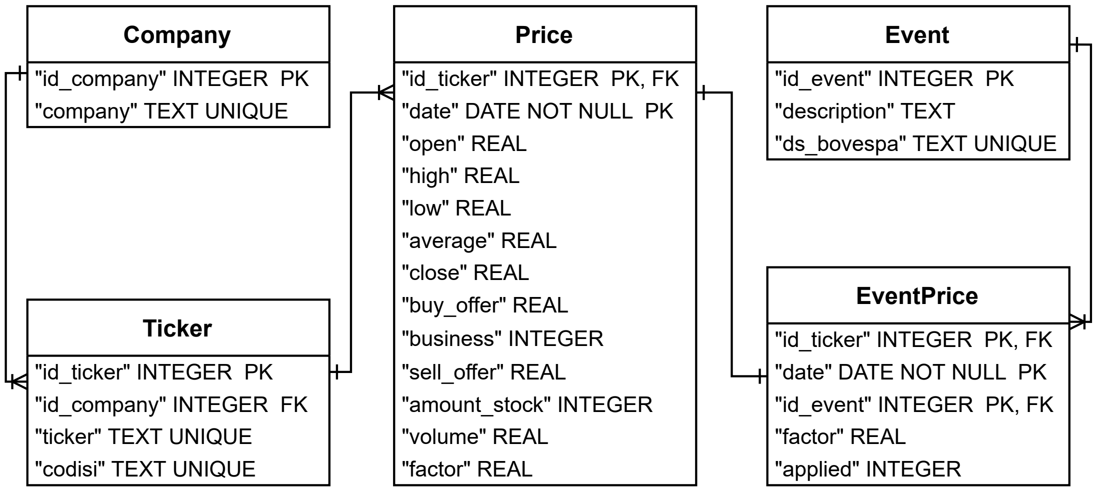

Figure 1. Relational model of the proposed data set BovDB.

Taking into account the scheme of the database, in what follows we provide a deeper analysis of the considered fields:

- **Company.id_company** – It is an auto incremented integer that represents the company identifier, therefore it is the Company's primary key, it is also used as a foreign key on the Ticker, to reference the former.
- **Company.company** – It's the Company's name.
- **Ticker.id_ticker** – It's an auto incremented integer that represents the Ticker identifier, therefore it is a primary key, it is also used as a foreign key on the EventPrice and the Price tables, to reference the former.
- **Ticker.ticker** – It's the Ticker/stock's name.
- **Ticker.codisi** – This is the stock code (from B3).
- **Price.date** – It represents the date of trade of a stock, it is also used, along with the id_ticker, to identify a given ticket on a specific date, thus forming a composite primary key. On the EventPrice, it is a composite primary key, along with the id_ticker and the id_event, that indicates the date on which a certain Event occurs.
- **Price.open, high, low, average, close, buy_offer, business, sell_offer, amount_stock, and volume** – These fields represent, respectively, the opening price, the highest price, the lowest price, the average price, the closing price, the best offering price, the number of trades carried out with the paper, the best selling price, number of stocks traded on this paper, the total volume of titles traded on this paper. All of the prior columns are of a given date.

- **Price.factor** – It is the cumulative impact of events from newest to oldest until a specific date is reached.
- **Event.id_event** – It's an auto incremented integer that represents the Event identifier, therefore it is a primary key, it is also used as a foreign key on the EventPrice, to reference the former.
- **Event.description** – It's the Event description.
- **Event.ds_bovespa** – The Event abbreviation, as it is shown in the files provided by B3.
- **EventPrice.factor** – It is the impact of the event on a specific stock and day.
- **EventPrice.applied** – Represents the Events that we take in consideration as a 1, and 0 the ones we don't consider.

5. Data Set Contextualization

In this section, we describe the main steps related to the study. Precisely, we describe some of the challenges and obstacles we encountered, while extracting and manipulating the data.

As mentioned before, the raw daily values are taken from B3's website, and they are available in TXT format inside a ZIP file. The first problem we encountered was the fact that some years' file data were corrupted, meaning their format was incorrectly saved and they weren't standardized. For example, the year 2000 file is named "COTAHIST.A2000", resulting in the file extension .A2000 instead of .txt. So we had to filter every file and create our own standardized file names and formats.

Another problem found was related with stocks that had their name changed, for example, Vale's stock VALE3 was formerly named VAL 3. This problem resulted in the data set having duplicate stocks. To overcome this we used the stocks codes in the ISIN (International Securities Identification Number), but this wasn't available until late 1995, so we had to adapt and use a jointly method to overcome this issue.

One more problem was the fact that the stocks' daily data was outdated, for example, values from the year 2000 had no correction applied, ignoring all the events that had happened until 2020, even though B3 provides the dates were events appeared, they don't give their current or retroactive impact, so we had to calculate our own. Then another problem came up when we realized that the dates where the events appeared were inconsistent, so we had to use the paper distribution number alongside the dates. A deeper explanation of how this calculation was made can be found in Section 6.

6. Data Set Description

In this section, we discuss the differentials of the presented data set and its utilities. To do so, we analyze some interesting cases where the data set could be used along with some related technologies.

Figure 2 shows the numbers of companies throughout the years. Every bar represents the total number of companies until that year, and each slice of the bar is related to additional information: the black part represents the companies that stopped operating in comparison with the last year, the red part represents the total companies that stopped operating from the begging, the light green represents the new companies that started appearing that year, and the green part refers to the currently operating companies.

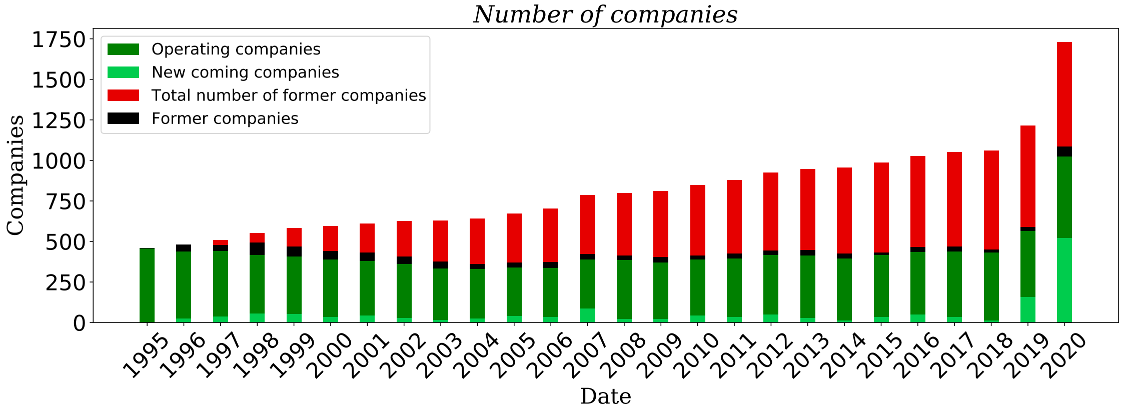

Figure 2. Companies by year of B3.

Analyzing this figure, it is observable that, since 1995 a lot of new companies appeared. This occurs due to some being created after this date, some companies merging, or some going on hiatus. It is also noticeable that there were a significant number of companies that became inactive through the years, this is due to some companies being purchased by another or some closing, for example the company Datasul that was bought by TOTVS in 2008. We highlight that all these cases can be tracked by using our proposed data set.

It is also possible to infer, by looking at Figure 2, that in 2020 a significant number of companies started operating and that there are more than 1750 total companies present in our data set.

Figure 3 shows the relations between numbers of stocks and years. Every bar represents the total number of stocks until that year, and each slice of the bar is related to additional information: the black part represents the stocks that stopped being traded in comparison with the last year, the red part represents the total stocks that stopped being traded since the begging, the light green represents the new stocks that started appearing that year, and the green part are the concurrent stocks being traded.

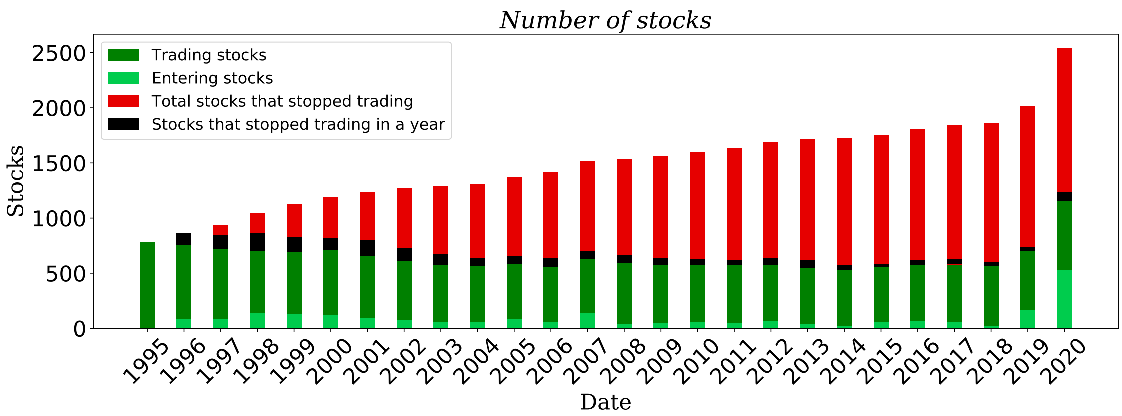

Figure 3. Total number of stocks.

Analyzing our data set it is possible to observe, alongside the graph in Figure 3, that throughout the years there were a significant number of stocks that seized their operations. For example, an enterprise that worked under a specific name for some years and then, due to many different reasons, such as merging, started to work with a different name or acronym, thus changing their stock code (ticker) and name. Or some companies needing to be closed for a wide range of reasons, resulting in their stock no longer being traded. Taking a closer look, at the same figure, it is possible to observe that there are more than 2500 total stocks present in our data set.

The biggest differential of this data set is the events and how we approach them, its impact and importance will be made clear throughout this section. The way we calculate the events' impact on the stocks is by dividing the closing price from the previous day by the opening price of the current day on which the event appeared. This is our "factor". Then we calculate the cumulative impact of all the events by multiplying all factors of the same stock until a given date is reached, thus changing the factor based on all the previous factors.

Figure 4 compares the prices between the data with and without the factor applied from a stock called PETR4 during a period where an event called ex-bonus occurs. This is evidence that there is a significant difference between the raw data (Figure 4, left) and the data when the "factor" is considered (Figure 4, right). Therefore, the Factor can be used to obtain more precise and interesting classifiers, since the gap caused by the ex-bonus could be otherwise seen as a depreciation of the stock, when in reality it was a price adjustment made by the company, in order to make the stock more accessible to market participants. This event did not alter the market value (total) of the company, it simply increased the number of available stocks for trading in B3.

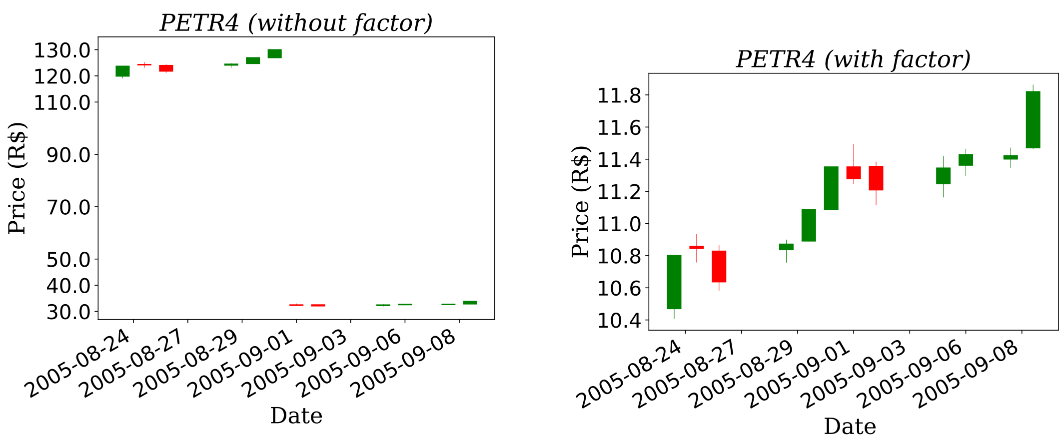

Figure 4. PETR4 comparison between 2005-08-24 and 2005-09-10 without the factor (left) and using the prices (right).

Another evidence of the factor's impact is present in Figure 5, where we present the average and standard deviation comparison between the stock VALE3 with (orange bars) and without (blue bars) the factor applied throughout all years from 1995 to 2020. On the top part of all annual bars, an up and down standard deviation is represented.

The data without the factor (original trading prices from that years without any correction) have a higher standard deviation due to events that unnaturally adjusted the stocks prices in their respective year. This adjustment makes it so that old prices do not

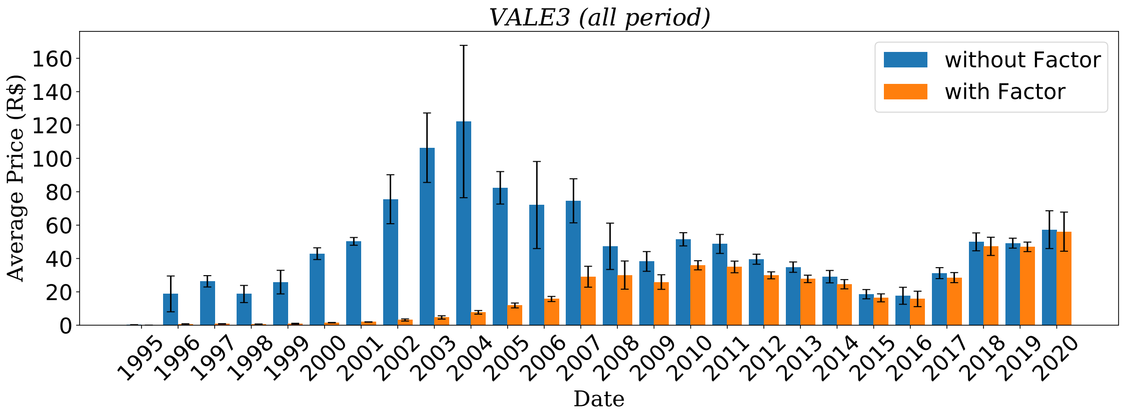

Figure 5. Average and Standard Deviation of VALE3 between 1995 and 2020, with and without factor

have a direct temporal comparison with the most current ones, since a stock today can represent more or less than a stock in the past. The data adjusted with the "factor", on the other hand, have a smaller standard deviation and a mean that is comparable (in terms of quantity) over the entire time series.

To use the "factor" in our data set, for each record of the table (see Figure 1) is necessary to divide the columns open, high, low, average, close, buy_offer, business and sell_offer by the factor of that day, and multiply the amount_stock the factor.

The graph in Figure 6 compares the average price throughout the years (Date) of our data set (orange bar) with two other data sets (InfoMoney (green bar) and BR Investing (blue bar)). Each bar also has an up and down standard deviation. It is noticeable that our values are positioned between the two other data sets in almost all years of the data period. It demonstrates that the values we came up with are neither too high nor too low when compared to some of the most highly regarded data set providers in Brazil, thus proving that our methods are valid and that our data set is in accordance with the market standards.

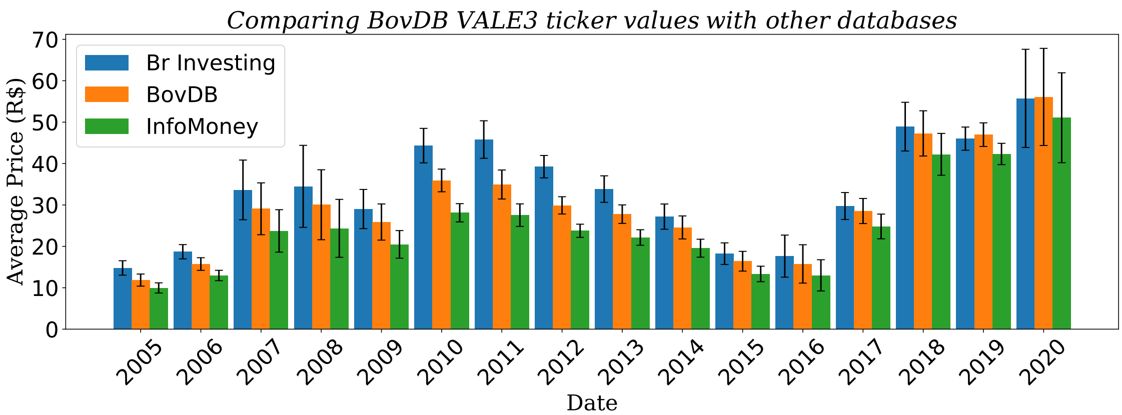

Figure 6. Average and Standard Deviation of VALE3 comparison among BovDB, InfoMoney, and BR Investing from 2005 to 2020

7. Download and Citation Request

The database as well as the sample files are available for download in the repository named "BovDBrepository[9]", licensed under CC BY-NC 4.0. In Table 1, by columns we provide the arrangement of the data files and their relative content. We highlight that there are two main folders available: the first, called "Codes", holds examples of accessing the base and the second, "DataBase", contains the project's database (See Section 4). All files related with the project are accessed by these folders and instructions to execute a test example is provided in the readme file.

Table 1. The overview of the provided git repository.

Directory	Content
BovDBrepository	Repository root directory
BovDBrepository/Codes	Folder with examples of access/use of the database
BovDBrepository/DataBase	Database directory

In case the BovDB is used for scientific or academic purposes, include a citation to this paper.

8. Conclusion

The goal of building and making available a data set from the Brazilian Stock Exchange (B3) that would be able to assist researchers in the fields of finance, computing, and statistics has been achieved. More than that, the daily raw data between 1995 and 2020 of all stocks that were listed on the old Bovespa and are listed (from 2017 onwards) on new B3, have been pre-processed and are ready for download.

BovDB is unprecedented in Brazil and can be used as a tool and benchmark for stock studies, such as Machine Learning algorithms that can be used to predict accurately fall and rise in prices.

The difficulties in building this data set were mainly related to inconsistencies and price corrections due to unnatural events in the market that we had to fix and work around to achieve a consistent data set that took in consideration the cumulative impact that the events had in the stock market. The pre-processing and adjustment calculations performed have been made transparent in this paper, which is not the case in other data sets made available by private companies on the internet.

Our data set is managed and maintained by the Information Management Research Group (GInfo)[10] of the Center for Computational Sciences (C3) from the Federal University of Rio Grande (FURG) with data made available on the internet by B3.

Finally, we point out that the potential of this data set is great since it simplifies a lot of the research work about B3 stocks. Moreover, it also facilitates the use of the data and its understanding, which will allow future work that needs to use B3 stocks. All of this without any hard work on this data and by saving hours in its usage.

[9]BovDB repository: `https://github.com/Ginfofinance/BovDBrepository`.
[10]GInfo website is `http://ginfo.c3.furg.br`

Acknoledgments

This study was supported by PNPD/CAPES (464880/2019-00), PIBIC/PIBITI CNPq and PROBIC FAPERGS.

References

[Alhnaity and Abbod 2020] Alhnaity, B. and Abbod, M. (2020). A new hybrid financial time series prediction model. *Engineering Applications of Artificial Intelligence*, 95:103873.

[Allen and Owens 2010] Allen, G. and Owens, M. (2010). *The Definitive Guide to SQLite*. Apress, USA, 2nd edition.

[Basak et al. 2019] Basak, S., Kar, S., Saha, S., Khaidem, L., and Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, 47:552–567.

[Bustos and Pomares-Quimbaya 2020] Bustos, O. and Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156:113464.

[Del Ángel 2020] Del Ángel, R. G. (2020). Financial time series forecasting using artificial neural networks. *Revista Mexicana de Economía y Finanzas Nueva Época REMEF*, 15(1):105–122.

[Efimov et al. 2020] Efimov, D., Xu, D., Kong, L., Nefedov, A., and Anandakrishnan, A. (2020). Using generative adversarial networks to synthesize artificial financial datasets. *arXiv preprint arXiv:2002.02271*.

[Garcia-Molina 2008] Garcia-Molina, H. (2008). *Database systems: the complete book*. Pearson Education India.

[Guo et al. 2018] Guo, Y., Han, S., Shen, C., Li, Y., Yin, X., and Bai, Y. (2018). An adaptive svr for high-frequency stock price forecasting. *IEEE Access*, 6:11397–11404.

[Harris 1997] Harris, R. D. (1997). Stock markets and development: A re-assessment. *European Economic Review*, 41(1):139–146.

[Hu et al. 2021] Hu, Z., Zhao, Y., and Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, 4(1):9.

[Li and Bastos 2020] Li, A. W. and Bastos, G. S. (2020). Stock market forecasting using deep learning and technical analysis: a systematic review. *IEEE Access*, 8:185232–185242.

[Nti et al. 2019] Nti, I. K., Adekoya, A. F., and Weyori, B. A. (2019). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, pages 1–51.

[Rahat et al. 2019] Rahat, A. M., Kahir, A., and Masum, A. K. M. (2019). Comparison of naive bayes and svm algorithm based on sentiment analysis using review dataset. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 266–270. IEEE.

[Rousis and Papathanasiou 2018] Rousis, P. and Papathanasiou, S. (2018). Is technical analysis profitable on athens stock exchange? *Mega Journal of Business Research*, 2018.

[Schierholt and Dagli 1996] Schierholt, K. and Dagli, C. H. (1996). Stock market prediction using different neural network classification architectures. In *IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pages 72–78. IEEE.

[Sezer et al. 2020] Sezer, O. B., Gudelek, M. U., and Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181.

[Sowinska and Madhyastha 2020] Sowinska, K. and Madhyastha, P. (2020). A tweet-based dataset for company-level stock return prediction. *arXiv preprint arXiv:2006.09723*.

[Thomaz et al. 2021] Thomaz, P. S., de Mattos, V. L. D., Nakamura, L. R., et al. (2021). Modeling volatility's long-range persistence and asymmetry effect of bradesco bank stock prices using garch models. *International Journal of Development Research*, 11(03):45532–45543.

[Upadhyay et al. 2012] Upadhyay, A., Bandyopadhyay, G., and Dutta, A. (2012). Forecasting stock performance in indian market using multinomial logistic regression. *Journal of Business Studies Quarterly*, 3(3):16.

[Vachhani et al. 2019] Vachhani, H., Obiadat, M. S., Thakkar, A., Shah, V., Sojitra, R., Bhatia, J., and Tanwar, S. (2019). Machine learning based stock market analysis: A short survey. In *International Conference on Innovative Data Communication Technologies and Application*, pages 12–26. Springer.

[Wang 2021] Wang, J. (2021). The analysis of the financial market in china. *Academic Journal of Business & Management*, 3(2).

[Zhang 2021] Zhang, E. (2021). Forecasting financial performance of companies for stock valuation. *Stanford Projects Spring 2021*.