BovDB: a data set of stock prices of all companies in B3 from 1995 to 2020

Fabian Corrêa Cardoso[1], Juan Andrey Valverde Malska[2], Paulo Junior Ramiro[2], Giancarlo Lucca[1], Eduardo Nunes Borges[2], Viviane Leite Dias de Mattos[3], Rafael Alceste 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 vast amounts of financial resources worldwide. This market generates a high volume of transaction data, which after being analyzed are very useful for many applications. In this article, we present BovDB, a data set that was built considering a source of 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 follow the market standards, based on the proposed factor. BovDB data set can be used as a benchmark for different applications and it is available in open access for any researcher on GitHub.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: B3, Data set, Stocks, Time series

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 by selling stocks and corporate bonds. Thus, a stock market is where buying and selling transactions take place. We highlight that there are different stock markets around the world: the NYSE (New York Stock Exchange)[1], the Chinese SSE (Shanghai Stock Exchange)[2], the SZSE (ShenZhen Stock Exchange)[3], and the B3 (Brasil, Bolsa, Balcão)[4]. The latter is the market that trades stocks of companies in Brazil. B3 is a private company, created in 1895 and named as "Bolsa de Títulos de São Paulo" (São Paulo Stock Exchange). After, it was renamed to "Bolsa de valores de São Paulo" (Bovespa) and since march of 2017 is B3 [CVM 2019].

[1] Access the NYSE website at `http://nyse.com`
[2] Access the Shanghai website at `http://english.sse.com.cn/`
[3] Access the Shenzhen website at `http://www.szse.cn/English/index.html`
[4] Access the B3 website at `http://www.b3.com.br`

One way to invest in stocks is using the fundamentalist approach [Vachhani et al. 2019]. By considering this kind of approach, investing requires knowledge about the companies. Since 2000 was created the "New Market", which has established a new standard of collaborative governance that goes beyond what is required by law [Anderson Jr 2003]. Companies must voluntarily exhibit more transparency in this stock market and may only issue ON (ordinary stocks), with voting rights, and no more PN (preferred stocks), without voting rights. The difference is that in the first one, the investor becomes a partner of the company and in the last one is a creditor of the company, a kind of financier that provides a loan [CVM 2019].

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. These big data need to be analyzed, compared, foreseen, calculated, etc. Considering the usage of these analyses, another way of investing is the technical approach.

An important question that arises is related to this amount of data and how to make it available to the public. The data on Brazilian stocks are made available digitally to the public and can be accessed daily on B3's website[5]. Among the information available, we highlight a short company name, the currency used, opening price, closing price, lowest price, highest price, and volume traded[6]. Although the data are available, they are provided in text files, separated by year and in their raw form (without formatting). This makes it difficult to extract knowledge from the B3's data set and to understand the distortions caused by the events in the stock prices. Consequently, it may be challenging to understand and complex for people not connected to the financial market.

Following the technical path, it can be noticed that the quality of information is fundamental in any data set, and this is essential for constructing forecast models that portray reality in the most reliable way possible. In time series, the most frequent problems are characterized by lack of data, level change, and sharp peaks or falls [Cao and Wang 2019]. In the last two cases, an observation or a set of observations is very different from the others, and it may arouse the suspicion that a different mechanism was generated by them. Its occurrence can be linked to technical-operational problems, such as typos and measurement errors, or variations in the context in which the variable is inserted [Pellegrini 2000], this is the case of a split, inplit, ex-bonus, and other events which occur with stocks.

Using some techniques to eliminate or minimize these problems in time series modeling can be quite pertinent. According to [Koehler et al. 2012], these occurrences can exert significant influence on the estimates and significance of the model parameters and, consequently, on its predictions. For [Melo and Castro 2013], it is essential to identify and treat these occurrences since the basic statistics can be biased by the interference of these values, leading to wrong conclusions. They agree with [Chandola et al. 2009], those who point out that data quality directly affects forecast quality.

It is easily noticeable that using this data is of utmost importance and crucial for understanding and studying 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, diverse types of data sets are used; there are several data sources and also time windows that 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 day trade. They are medium-term (swing trade) when executed for more than a day, a week, or a few months. They are long-term (buy and hold) when executed after many months, years, or even decades (rare). Generally, day trade and swing trade are linked to technical analysis and buy and hold are linked to fundamental analysis [Vachhani et al. 2019].

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

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

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 these data pre-processed ready for data mining. We provide the data set called BovDB and make it available to serve as a benchmark for future applications in technical analysis, such as stock price forecasting, among others. An essential aspect that this study provides is related to its usage in others Stock Markets that allow public access to their stock data. Furthermore, the factor (proposed in this article) can also be applied if the same stock events are available in that Stock Market.

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

2. RELATED WORK

On specialized sites such as InfoMoney[7] or Yahoo Finance[8] can be freely found pre-processed public data sets on stocks of several SE around the world. Yahoo Finance can provide data with a simple download command, within the code for those using the R programming language. However, some data was 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 to our knowledge.

The data provided by sites like those cited are not raw but pre-processed with adjustments based on price changes for events that do not occur during regular 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, especially for over a decade. On the other hand, in Kaggle[9], looking for stocks, it is possible to 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 their self database, to make it able for use.

Among some related works, we can cite [Efimov et al. 2019] that use data sets from American Express for risk modeling. They also list data sets available, as they say, “used to benchmark and validate Machine Learning algorithms developed by various researchers, academic groups and companies.” The authors assumed that their Generative Adversarial Networks model replicated the relationship between the target variables and the data features with reasonable accuracy.

In [Guo et al. 2018], was used a data set of the Shanghai Stock Exchange for intraday analysis using the adaptive Support Vector Machine Regression (SVR) method for high-frequency stock price prediction on a 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 backpropagation 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 was done of the features used for forecasting.

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

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

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

The study provided by Sowinska and Madhyastha uses a data set of text from Twitter as samples and the stock return information as labels to predict the impact on stocks with four 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."

Taking into account the data set provided by the Yelp Data set Challenge[10] 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 the Kaggle website[11] to classify. The authors stated that the best classifier for binary and multi-class classification was C-LSTM obtaining good overall scores.

Among the data sets found, not only in the articles mentioned above but in others, it is widespread to observe that the calculations with stocks are usually performed mainly with the stock daily closing value. Although some works 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 is not tricky, although it is not ordinary to find the data sets they used 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 public resources grant the study).

In the work presented in this article, we pre-process B3 data and make them 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, making a complete description of BovDB's data set. Thus, we have three main differences between our data set and the commercial ones. The first is related to open access, e.g., anyone can download it and use it without any charge. Moreover, the second difference is that commercial data sets (Infomoney and Yahoo), as previously discussed, did not provide how they calculate their factor. Finally, the last difference related to the academic data set is that we proposed our factor and the use of B3 data.

3. DATA SET DESIGN

The data set is stored in a relational database management system (RDBMS) [Garcia-Molina 2008] known 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. Instead, it is embedded into the end program. Additionally, it does not have any requisites, nor is it required to be downloaded, so 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 data on the different event types;
—**EventPrice** - Stores data of a specific event on a specific stock on a given date.

In Fig. 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.

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

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

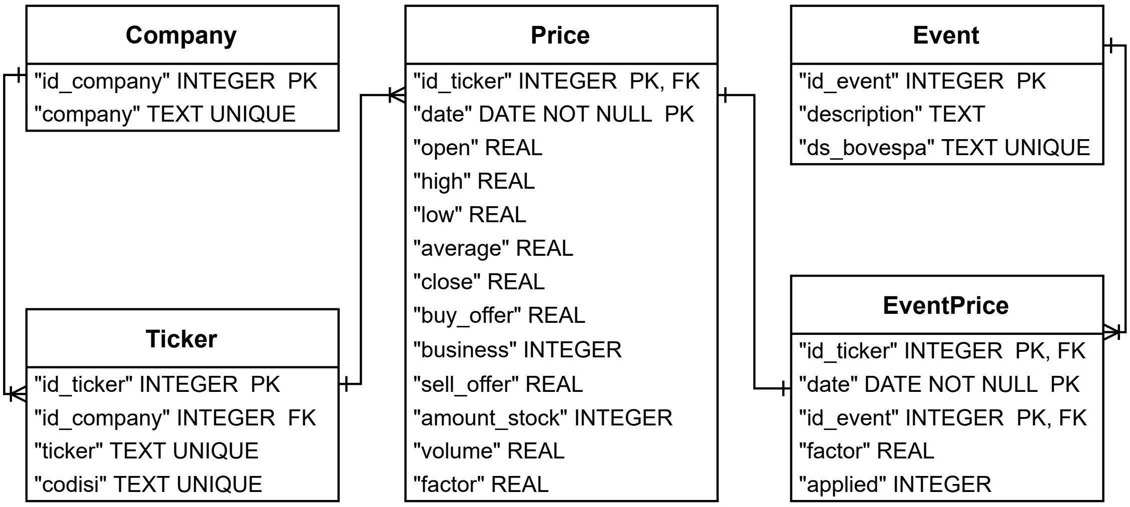

Fig. 1: Relational model of the proposed data set BovDB.

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 is the company's name.

—**Ticker.id_ticker** – It is 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 is the ticker/stock's name.

—**Ticker.codisi** – This is the stock code (of B3).

—**Price.date** – It represents the stock trade date, it is also used, along with the id_ticker, to identify a given ticker on a specific date, thus forming a composite primary key. The EventPrice 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 previous 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 is 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 is 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.

4. DATA SET DESCRIPTION

This section discusses 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. We also demonstrate the usefulness of factor and our new database's potential to support stock analysis.

The Table I summarizes the principal amounts in the data set: companies, events, stocks, and tickers. We can observe that there are 12 events possible in the B3 Exchange Market, 1728 are all the companies that are and were listed in B3 from 1995 to 2020, 2540 are all the stocks listed (a company can have 1 or more tickers) and 1,843,399 are all the quotations by day of all companies.

Table I: Summarizing the principal amounts of the data set

Tables	Number of occurrences	Description
event	12	Table that informs the types of events.
company	1,728	Table that stores information about companies.
ticker	2,540	Table that stores information about stocks.
eventprice	26,967	Table that stores information about events.
price	1,843,399	Table that stores information about stock movements.

Fig. 2 shows the numbers of companies throughout the years. Every bar represents the total number of companies until that year. Each slice of the bar has the following information: the black part are the companies that stopped operating in comparison with the last year, the red part is the total companies that stopped operating from the begging, the light green are the new companies that started appearing in that year, and the green part are the currently operating companies.

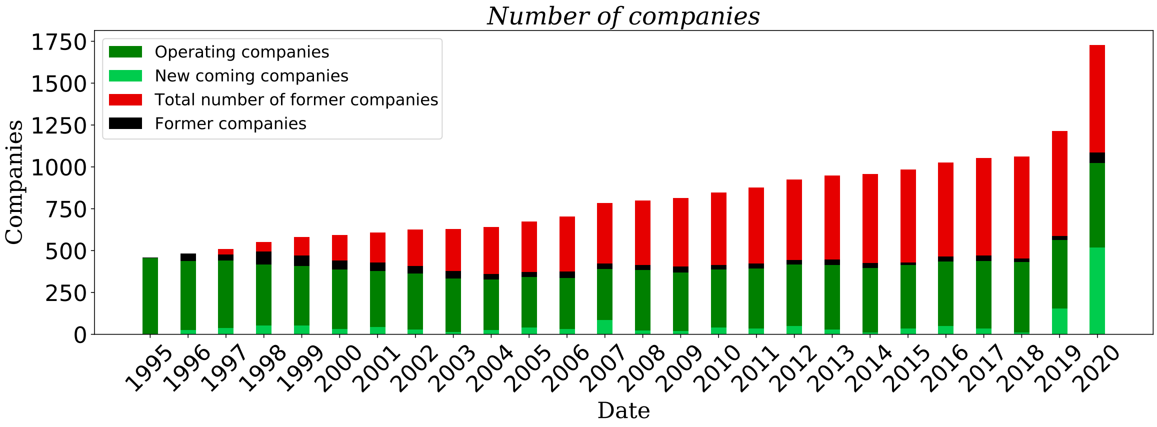

Fig. 2: Companies by year of B3.

Analyzing this figure, it is observable that, since 1995, many new companies appeared. It occurs due to some being created after this date, some companies merging, or some going on hiatus. It is also noticeable that many companies have become inactive through the years. It is due to some companies being purchased by others. For example, the company Datasul was bought by TOTVS in 2008, or some closing. We highlight that all these cases can be tracked using our proposed data set. It is also possible to infer, by looking at Fig. 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.

Fig. 3 shows the relations between numbers of stocks and years. Every bar represents the total number of stocks until that year. Each slice of the bar has the following information: the black part are the stocks that stopped being traded in comparison with the last year, the red part is the total stocks that stopped being traded since the begging, the light green are the new stocks that started appearing that year, and the green part are the current stocks being traded.

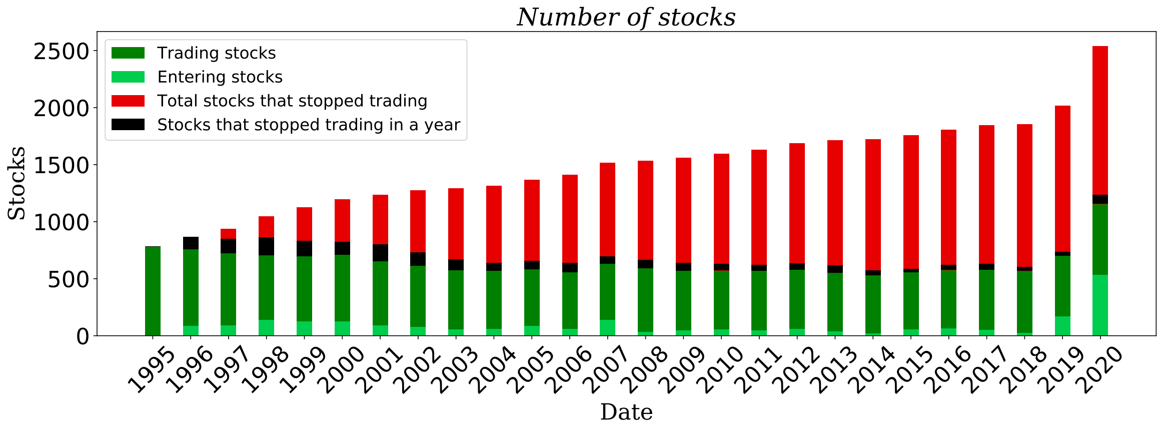

Fig. 3: Total number of stocks.

Analyzing our data set is possible to observe, alongside the graph in Fig. 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 years and, due to many different reasons, such as fusion, started working with a different name or acronym, thus changing their ticker and name. Alternatively, some companies need to be closed for a wide range of reasons, resulting in their stock no longer being traded. We can observe more than 2500 total stocks present in our data set, looking closer at the same figure.

4.1 The impact of the usage of Factor

The most significant differential of this data set are the events and how we approach them, their impact and importance will be made clear throughout this section.

Let d be the previous day on which an event occurs, the *factor_ isolated* and *factor_ cumulative* are defined by equations 1 and 2, respectively. We calculate the events' impact on the stocks by dividing the closing price from the previous day, close(d), by the current day's opening price, open(d+1), on which the event appeared. It is the so-called "factor." Then we calculate the cumulative and regressive (from 2020 to 1995, stock by stock) 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.

$$factor\_isolated(d) = \frac{close(d)}{open(d+1)} \tag{1}$$

$$factor\_cumulative(d+1) = \frac{factor\_cumulative(d)}{factor\_isolated(d)} \tag{2}$$

To use the "factor" in our data set, for each record of the table (see Fig. 1) is necessary to divide the columns open, high, low, average, close, buy_offer, business, and sell_offer by the factor of that day (*factor_ isolated*), and multiply the amount_stock by the *factor_ cumulative*. To better understand how we calculate the factor, we will use the opening and closing prices from 2007-08-29 to 2007-09-05 from VALE3 to simulate our algorithm's process.

In Table II, it is possible to observe in the fourth row an event EB. When an event appears, we use the closing price of the day before the first day on which it appears and divide it by the first day's opening price. The result is stored on the day before which the event first appeared, as observed in the column *factor_ isolated*.

Afterward, as seen in the column *factor_ cumulative*, the factor is multiplied from last to first and stored accordingly. The value of 1.86000 observed in rows 4, 5, and 6 were obtained in other factor calculations.

Table II: An example of how to calculate an individual factor

Row	Date	Open	Close	Event	Factor_Isolated	Factor_Cumulative
1	2007-08-29	91.71000	93.80000	/	1.00000	3.72000
2	2007-08-30	93.40000	94.20000	/	1.00000	3.72000
3	2007-08-31	96.51000	96.98000	/	1.99547	3.72000
4	2007-09-03	48.60000	48.96000	EB	1.00000	1.86000
5	2007-09-04	49.10000	50.00000	/	1.00000	1.86000
6	2007-09-05	49.00000	48.85000	/	1.00000	1.86000

The graph in Fig. 4 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 compared to some of Brazil's most highly regarded data sets of stock providers. It proves that our methods are valid and that our data set follows the market standards.

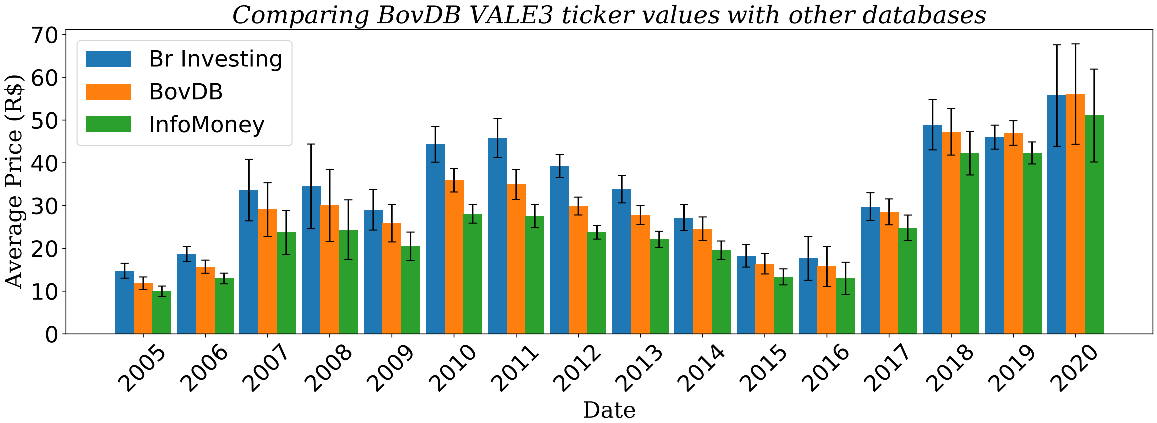

Fig. 4: Average and Standard Deviation of VALE3 comparison among BovDB, InfoMoney, and BR Investing from 2005 to 2020

Another evidence of the factor's impact is present in Fig. 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 is represented an up and down standard deviation.

The data without the factor (original trading prices from the 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 directly compare temporal with the most current ones since today's stock can represent more or less than stock in the past. The data adjusted with the "factor", on the other hand, have a minor standard deviation and a mean that is comparable (in terms of quantity) over the entire time series.

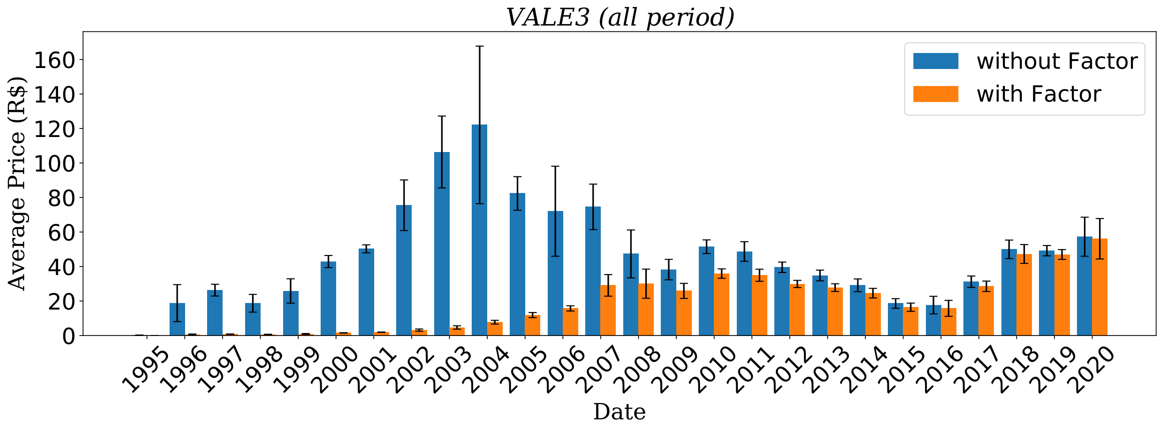

Fig. 5: Average and Standard Deviation of VALE3 between 1995 and 2020, with and without factor

Fig. 6 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. In this figure, on that day, the green candle means that was growth and the red candle means that was a fall in the stock price. See [Nison 2001] for more details. There is a significant difference between the raw data (Fig. 6, left) and the data when the "factor" is considered (Fig. 6, right). Therefore, the factor can be used to obtain more accurate and interesting classifiers since the gap caused by the ex-bonus could be, otherwise, seen as a stock depreciation. The company, in reality, made a price adjustment to make the stock more accessible to market participants. This event did not alter the company's market value (total), and it simply increased the available stock number for trading in B3.

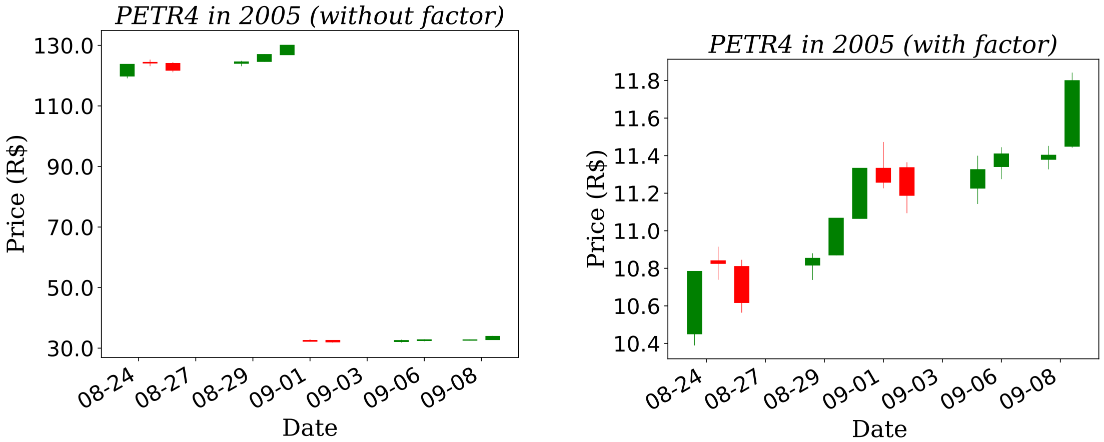

Fig. 6: PETR4 comparison between 2005-08-24 and 2005-09-10 without the factor (left) and using the prices (right).

The treatment performed on the data by factor aims to "standardize", for instance, the daily opening, and closing prices. The objective is to eliminate possible discrepancies, outliers and problems that can interfere with the calculations or that can affect interpretations and analyses.

Fig. 7 shows the raw data of the daily behavior of the opening and closing prices of ABEV3[12] stocks[13] in the year 2010 on B3 (called Bovespa at that time). This figure shows that they are similar,

[12] A company named Ambev, more information can be found at `https://www.ambev.com.br/`

[13] In 2010 the ticker was changed to AMBV3

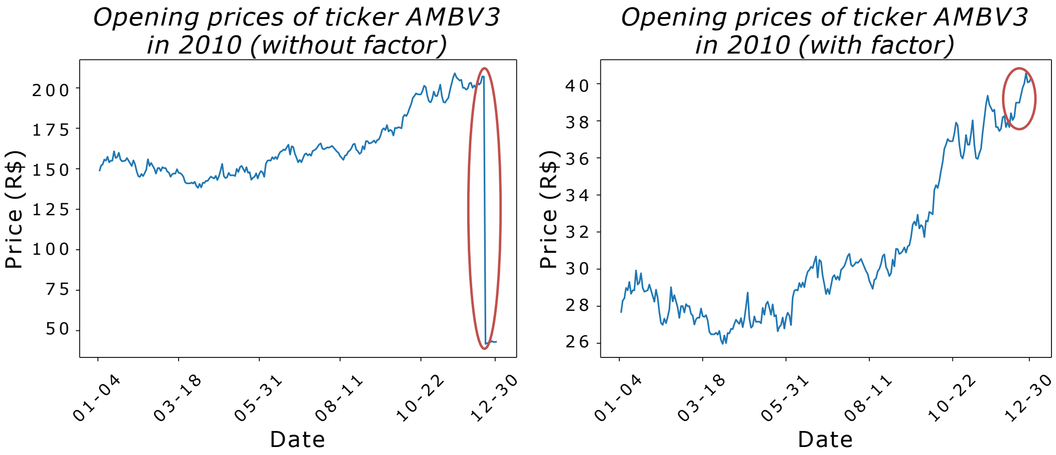

Fig. 7: Opening prices of AMBEV Stock in B3 in the year of 2010 with and without factor

but they present a sharp drop at the end of the analyzed period. In this case, ex-bonus was an event that changed the opening and closing price in the last month of 2010, reducing the previous closing value by approximately one-fifth. This occurrence makes the modeling process difficult.

Factor can contribute to obtaining more interesting classifiers using an actual (representative) time series, as can be seen in the gap caused by the ex-bonus event (Fig. 7) could be identified as an abrupt stock depreciation. Although it was only an adjustment effectuated by the company, the objective was to make the stocks more accessible for buying. This event has not altered the company's market value but has increased the stock number available for negotiation in B3.

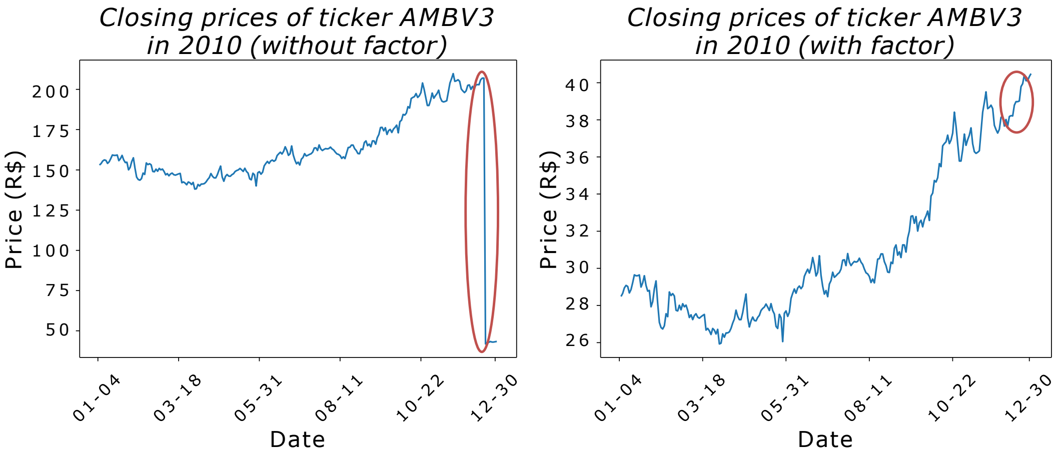

Fig. 8: Closing prices of AMBEV Stock in B3 in the year of 2010 with and without factor

Fig. 8 shows the behavior of these same series after treatment with the factor, showing the elimination of the identified problem. Also can be observed that with factor, the variation is reduced from 50 to 250 (raw data) to 26 to 40 ("factorized") because factor threats the data and reduces the range of prices to a range of "factors".

From now on, in this subsection, we exemplify other events in B3 data set to justify the usage of factor instead of raw data. When an inplit[14] event occurs, as can be seen in Fig. 9, the stock price changes a lot in just one day. In this figure, we provide an example of GOAU3[15], where the stock price was multiplied by 10, or, in other words, the price, by the usage of factor (Fig. 8) can be comparable with the 2020 price and is noticeable that the variation is almost imperceptible.

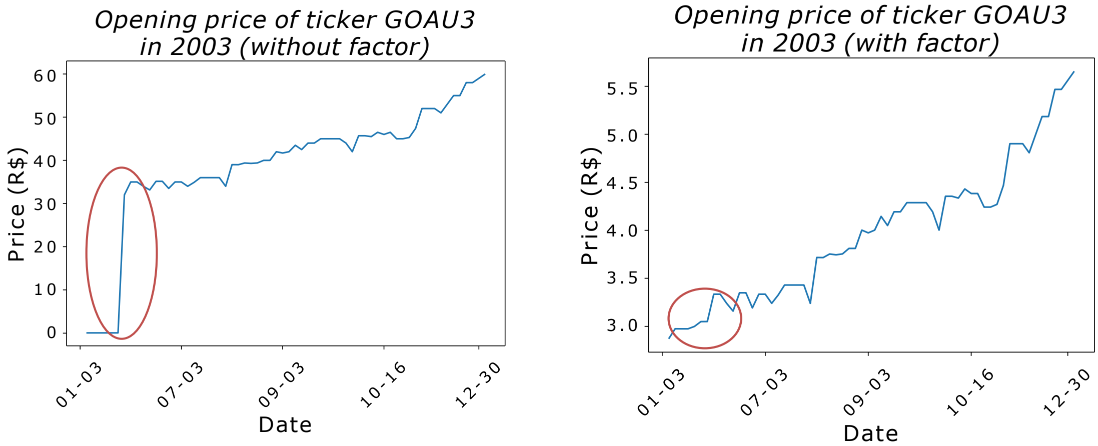

Fig. 9: An Inplit event which occured in B3 with the stock GOAU3 in 2003

Another event that we exemplify, which occurs in the data stored in the B3 data set, is the payment of dividends[16]. As can be seen in Fig. 10, the stock price showed a fall on February 2, 2006. Nevertheless, this fall was not in the price and just because of this event. With the factor, this does not affect the price of the stock.

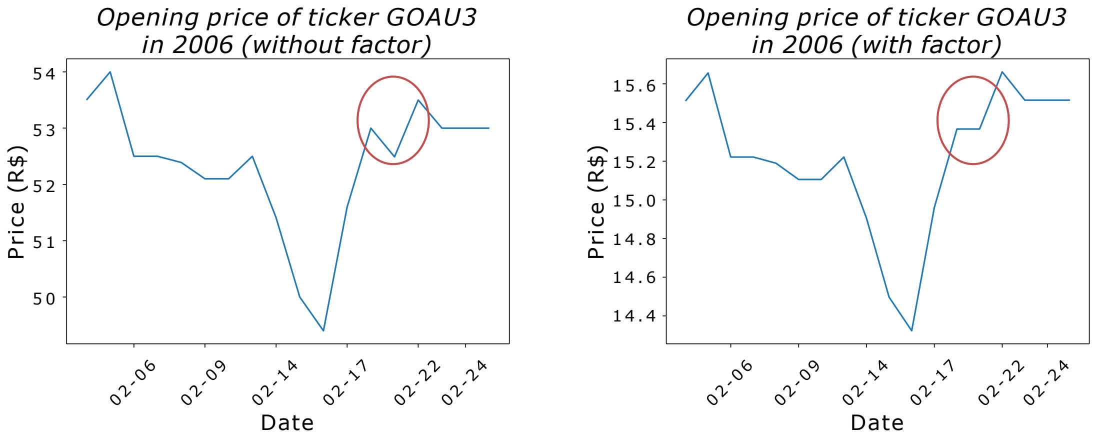

Fig. 10: A Dividend event occurred in B3 with the stock GOAU3 in 2006

[14] An inplit event happens when two or more stocks are grouped in one stock

[15] A company named Gerdau, more information can be found on: `https://www2.gerdau.com.br/`

[16] The payment of dividends refers to a portion of the companies' net profit which is destined to remunerate the shareholders without payment of income tax

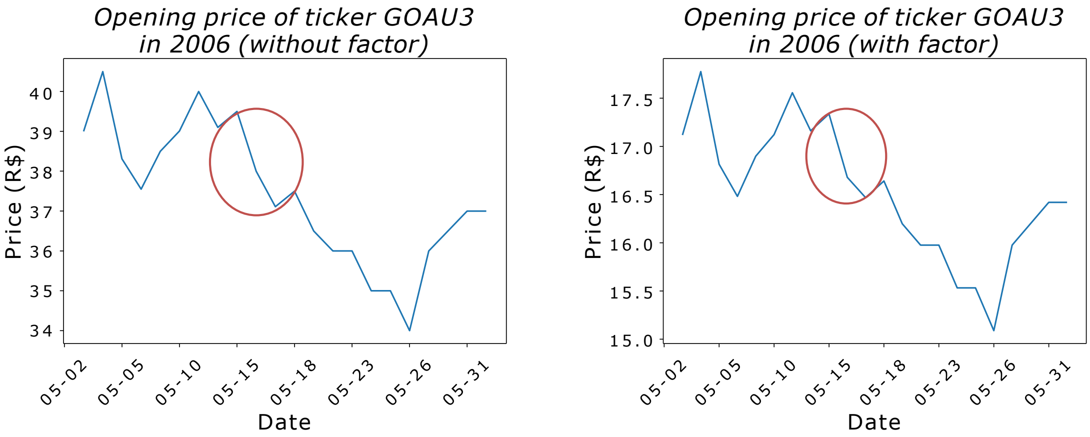

Fig. 11: An Interest on Equity event occurred in B3 with the stock GOAU3 in 2006

The payment of interest on equity[17] is another event that occurs in B3. Fig. 11 shows when this event occurred in 2006 with GOAU3, with and without application of factor. As can be seen in the graphic, with factor the event can be better perceived.

4.2 Analyzing the data set

This subsection shows the power of our data set. To do so, we exemplify some possibilities that can be reached when the available data are considered. Precisely, we produced some graphics to demonstrate how we can analyze data of stocks and sectors of the Brazilian economy that are listed in B3.

Fig. 12 shows events over the years in B3. They have increased mainly over the last decade but especially last year. It is possible to see that in 2020 the number of dividend events was multiplied by more than three, considering the average of these events from 1995 to 2019. Also can be observed that negotiations with BDR-DRN, which are stocks of other countries negotiated in B3, have increased 207,95% from 2019 to 2020.

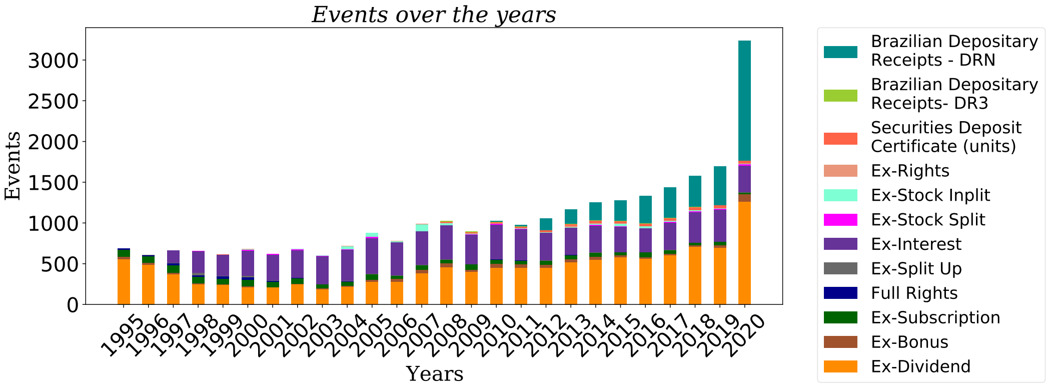

Fig. 12: Events which occurred in B3 from 1995 to 2020

Fig. 13 shows the behavior of essential stocks in B3 from 1995 to 2020. We highlight that these

[17] The payment of interest on equity refers to the profit obtained by the company in the previous years destined to remunerate the shareholders, with incidence of 15% income tax payed for the shareholders

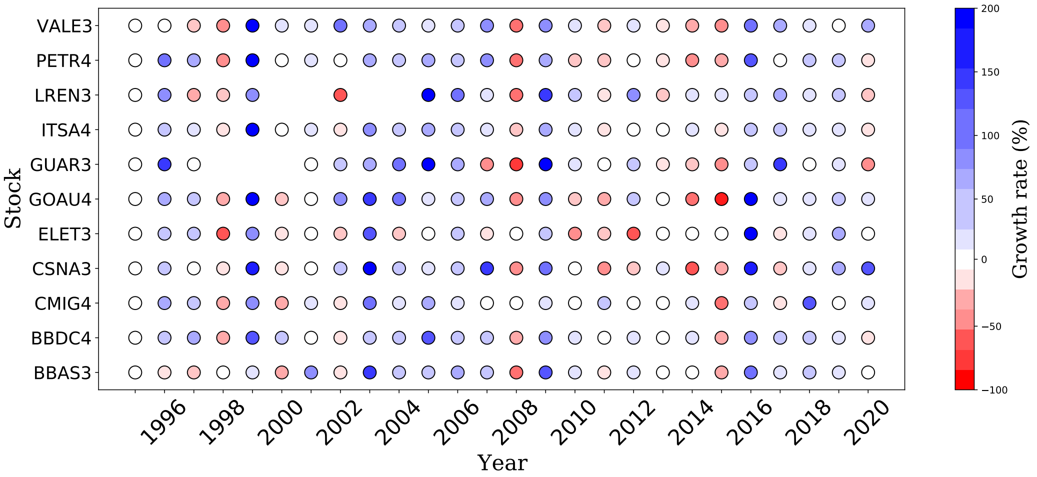

Fig. 13: Comparison of important stocks in B3

stocks are related to some of the most important companies in Brazil, thus having huge participation (percent) in the Bovespa Index. This figure presents the value of the stocks for each company (rows) per year (columns). Observe that each value is related to a scale of color, where red is the lowest and blue is the highest.

We can see, for instance, the company GUAR3[18] was not listed from 1998 to 2000. It is also possible to see that 2003 was one of the best years for almost all companies presented in this graphic. In 1998, 2008, and 2015, on the other hand, almost all companies decreased in value or, in the best case, stayed stable but recovered in 1999, 2009, and 2016.

In a similar schema to the previous analysis, we provide, in Fig. 14, the behavior of the stocks but consider different sectors of the economy. These sectors are used for B3 to classify the stocks and are similar to Global Industry Classification Standard (GICS), developed by Morgan Stanley Capital International (MSCI) and Standard & Poor's [Standard & Poor's 2020]. We highlight that the values presented in this figure are the average among some stocks per sector and that in Fig. 13 and Fig. 14 the blue color means that was increased (in percentage) and the red color means that was decreased (in percentage).

In the Information Technology sector, it is possible to see that it started in 2007 in B3. It is a sector that had a fall in the 2008 crisis, like almost all the other sectors. In the Healthcare sector, we can see that it started in B3 in 2005.

Generally, we can observe that after the 2008 crisis, the sectors have not returned to have the gain (dark blue in the graphic) experienced before. In 2009 was a good recovery in all sectors, but after it showed more losses and stability. Also can be seen that the Financial sector had an average performance in these twenty-six years, as the Public utility and Consumer Staples had.

Finally, our last study demonstrates another important feature that can be considered in our data set. A week-by-week analysis (Fig. 15). To do so, we focus on the last year available in the B3 data set, 2020. We provide an analysis of the stocks related to the companies which represent some sectors of the Brazilian economy:

—ITUB4: a bank, named Itaú

—MGLU3: a network of department stores, named Magazine Luiza

[18] A company named Guararapes Confecções, more information can be found on: `https://www.guararapes.com.br`

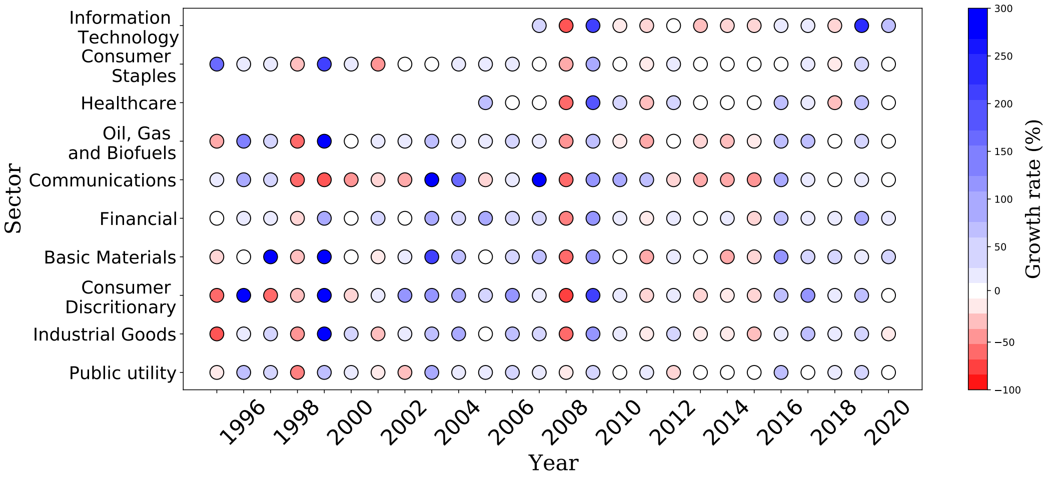

Fig. 14: Comparison of stocks by sectors in B3

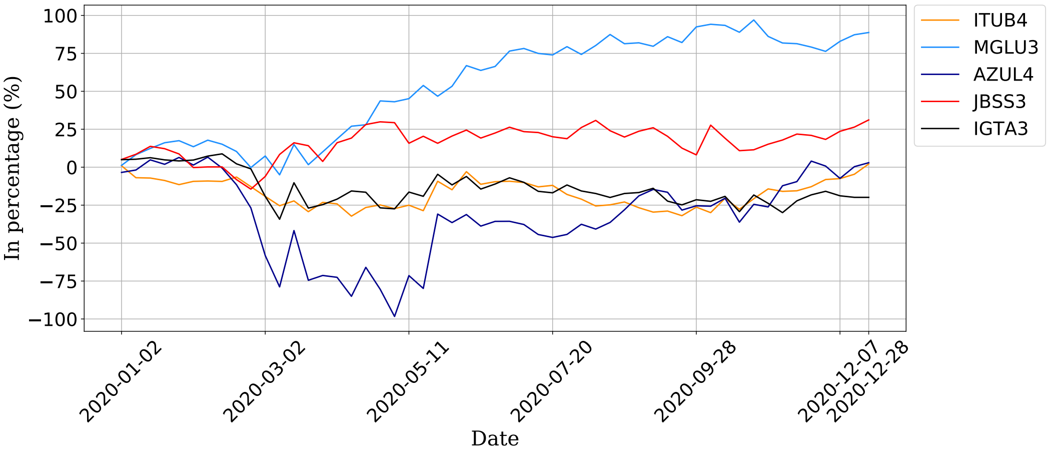

Fig. 15: Comparison of stocks in B3 in 2020

—AZUL4: an airline company, named Azul
—JBSS3: a slaughterhouse company, named JBS
—IGTA3: a network of shopping centers, named Iguatemi.

We can see that Covid-19 has affected some stocks, like IGTA4 and AZUL4, which have not recovered their value up to the end of 2020. The stocks of Itaú returned to the same level, and Magazine Luiza and JBS have increased their value. In Fig. 15 also can be observed that all stocks oscillated negatively in early March 2020, when the pandemic of Covid-19 started.

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 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. Thus, after downloading the data, we tried to unzip the data automatically. The first problem we encountered was the fact that some years' file data were corrupted, meaning their format was incorrectly saved and they were not standardized. For example, the year 2000 file is named "COTAHIST.A2000", resulting in the file extension .A2000 instead of .txt. So, as a second step, we had to filter every file and create our standardized file names and formats.

Another problem found was related to stocks that had their ticker changed. For example, Vale's stock VALE3 was formerly named VAL 3. This problem resulted in the data set having the same stock duplicated tickers. To overcome this, as a next step, we used the stocks codes in the ISIN (International Securities Identification Number), but this was not available until late 1995, so we had to adapt and use jointly a method to overcome this issue.

One more problem arose when we realized that the dates where the events appeared were inconsistent, so we used the paper distribution number alongside the dates. Then, another problem was 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 where events appeared, they do not give their current or retroactive impact, so we had to calculate our own.

We detail the main pre-processing steps to build BovDB in Table III.

Table III: Pre-processing steps to build BovDB

Steps	**Problems identified**
Download of the data	Need of download a file per year
Access the files downloaded	Corrupted name file
Stocks tickers identification	Stocks tickers duplicated
Identification of the events	Some events were not listed in B3 data
Calculate factor	Impact of events were not previously applied (see subsection 4.1 for more details)

6. APPLICATION

The market stock data could be used, by academic researchers, to make analyses [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 research in this field. Thus, the great advantage is that pre-processed, reliable, and organized information is supplied in a data set. It can be used as the basis of many applications. This section aims to describe some of them.

Applications with data sets of the stock market 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 of the data generating mechanism could be studied, characterizing the behavior of a series with the identification of periodicity, for example, [Thomaz et al. 2021]. It 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 data sets of the stock market, 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.

All of the BovDB fields can be used, in all the applications listed above, but we highlight that the close price and the factor, calculated as seen in subsection 4.1, are the main fields to be considered.

7. DOWNLOAD AND CITATION REQUEST

The database, as well as the sample files, are available for download in the repository named "BovDBrepository[19]", licensed under CC BY-NC 4.0. In Table IV, 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-named "Codes", holds examples of accessing the base and the second, "DataBase", contains the project's database (See Section 3). All files related to the project are accessed by these folders and instructions to execute a test example are provided in the readme file.

Table IV: The overview of the provided git repository.

Directory	**Content**
BovDBrepository	Root directory repository
BovDBrepository/READ.md	File with general explanations about the repository.
BovDBrepository/Codes	Directory with examples of access/use of the data set
BovDBrepository/Codes/annual_ mean_ and_ standard_ deviation_ graph.py	Example of how create a graph of mean and standard deviation with the data set
BovDBrepository/Codes/daily_ candlesticks_ chart.py	Example of how create a candlesticks chart with the data set
BovDBrepository/DataBase	Directory of databases
BovDBrepository/DataBase_ CSV	Directory with the data set in a CSV file format
BovDBrepository/DataBase_ JSON	Directory with the data set in a JSON file format
BovDBrepository/DataBase/DataBase.db	Project SQLite database.
BovDBrepository/DataBase/DataBase.db.sql	Project SQL database.

In case the BovDB is used for scientific or academic purposes, please include a citation to this work as the following example: "Fabian C. Cardoso, Juan A. V. Malska, Paulo Jr. Ramiro, Giancarlo Lucca, Eduardo N. Borges, Viviane L. D. de Mattos and Rafael A. Berri (2021). BovDB: a data set of stock prices of all companies in B3 from 1995 to 2020. Journal of Information and Data Management. Rio de Janeiro, Brazil".

8. CONCLUSION

The goal of building and making available a data set from the Brazilian Stock Exchange (B3) that would assist researchers in finance, computing, and statistic have 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 the new B3, have been pre-processed and are ready for download.

The difficulties in building this data set were mainly related to inconsistencies and price corrections due to strange events in the market that we had to fix and workaround to achieve a consistent data set that took into consideration the cumulative impact that the events had on the stock market.

BovDB is unprecedented in Brazil and can be used as a tool and benchmark for stock studies, such as Machine Learning algorithms that can accurately predict falls and rises in prices. New database versions can be created and tuples can easily be inserted because it is a relational database and can be migrated to a scalable architecture. Besides, the data set is available in several digital formats: .csv, .db, .json, and .sql.

The pre-processing and adjustment calculations performed have been made transparent in this article, which is not the case in other data sets made available by private companies on the internet. The use of factor becomes easier to classify stocks because it corrects discrepancies and highlights some events, which facilitates calculations. Our data set is complete and powerful and can help calculate and analyze Brazilian stocks listed in B3.

[19] BovDB repository: `https://github.com/Ginfofinance/BovDBrepository`.

BovDB is managed and maintained by the Information Management Research Group (GInfo)[20] of the Center for Computational Sciences (C3) from the Federal University of Rio Grande (FURG) with data made available on the internet by B3.

We point out that the potential of this data set is excellent since it simplifies a lot of the research work about B3 stocks. Moreover, it also facilitates the use of the data and its understanding, allowing future works that need to use B3 stocks without any hard work on these data and saving hours in its usage.

We also point out that it is possible to apply the factor in other Stock Markets that make their data and events available. Also, the usage of Machine Learning techniques in the BovBD is another promising and exciting field of research.

The next steps in this research are to use a machine-learning algorithm to forecast, like reservoir computing ([Budhiraja et al. 2021]) or deep learning ([Lara-Benítez et al. 2021]), making possible an exploratory analysis, and statistical analysis, like Arima ([Domingos et al. 2019]) or Garch ([Thomann 2021].

Acknoledgments

This study was supported by CNPq (305805/2021-5), PNPD/CAPES (464880/2019-00), PIBIC/PIBITI CNPq, and PROBIC FAPERGS.

REFERENCES

Alhnaity, B. and Abbod, M. A new hybrid financial time series prediction model. *Engineering Applications of Artificial Intelligence* vol. 95, pp. 103873, 2020.

Allen, G. and Owens, M. *The Definitive Guide to SQLite.* Apress, USA, 2010.

Anderson Jr, J. W. Corporate governance in brazil: Recent improvements and new challenges. *Law & Bus. Rev. Am.* vol. 9, pp. 201, 2003.

Basak, S., Kar, S., Saha, S., Khaidem, L., and Dey, S. R. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance* vol. 47, pp. 552–567, 2019.

Budhiraja, R., Kumar, M., Das, M. K., Bafila, A. S., and Singh, S. A reservoir computing approach for forecasting and regenerating both dynamical and time-delay controlled financial system behavior. *Plos one* 16 (2): e0246737, 2021.

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

Cao, J. and Wang, J. Stock price forecasting model based on modified convolution neural network and financial time series analysis. *International Journal of Communication Systems* vol. 32, pp. e3987, 05, 2019.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41 (3): 1–58, 2009.

CVM. *Mercado de Valores Mobiliários Brasileiro.* Comissão de Valores Mobiliários, Rio de Janeiro, 2019.

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

Domingos, S. d. O., de Oliveira, J. F., and de Mattos Neto, P. S. An intelligent hybridization of arima with machine learning models for time series forecasting. *Knowledge-Based Systems* vol. 175, pp. 72–86, 2019.

Efimov, D., Xu, D., Kong, L., Nefedov, A., and Anandakrishnan, A. Using generative adversarial networks to synthesize artificial financial datasets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Garcia-Molina, H. *Database systems: the complete book.* Pearson Education India, Upper Saddle River, NJ, 2008.

Guo, Y., Han, S., Shen, C., Li, Y., Yin, X., and Bai, Y. An adaptive svr for high-frequency stock price forecasting. *IEEE Access* vol. 6, pp. 11397–11404, 2018.

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

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

[20] GInfo website is `http://ginfo.c3.furg.br`

Koehler, A. B., Snyder, R. D., Ord, J. K., and Beaumont, A. A study of outliers in the exponential smoothing approach to forecasting. *International Journal of Forecasting* 28 (2): 477–484, 2012.

Lara-Benítez, P., Carranza-García, M., and Riquelme, J. C. An experimental review on deep learning architectures for time series forecasting. *International journal of neural systems* 31 (03): 2130001, 2021.

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

Melo, D. C. and Castro, A. R. Uma nova abordagem para detecção de outliers em séries temporais: estudo de caso em consumo de energia na região amazônica. In Anais do Simpósio Brasileiro de Matemática Aplicada e Computacional 2013. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics* 1 (1): 1–4, 2013.

Nison, S. *Japanese Candlestick Charting Techniques: A Contemporary Guide to the Ancient Investment Techniques of the Far East.* New York Institute of Finance, 2001.

Nti, I. K., Adekoya, A. F., and Weyori, B. A. A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, 2019.

Pellegrini, F. R. Metodologia para implementação de sistemas de previsão de demanda. *Mestrado em Engenharia de Produção-Departamento de Engenharia de Produção e Transportes. Porto Alegre-Universidade Federal do Rio Grande do Sul*, 2000.

Rafay, A., Suleman, M., and Alim, A. Robust review rating prediction model based on machine and deep learning: Yelp dataset. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, pp. 8138–8143, 2020.

Rahat, A. M., Kahir, A., and Masum, A. K. M. 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)*. IEEE, pp. 266–270, 2019.

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

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

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

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

Standard & Poor's, M. Global industry classification standard. *New York: Standard & Poor's*, 2020.

Thomann, A. Multi-asset scenario building for trend-following trading strategies. *Annals of Operations Research* 299 (1): 293–315, 2021.

Thomaz, P. S., de Mattos, V. L. D., Nakamura, L. R., et al. 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, 2021.

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

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

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

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