BrStats: a socioeconomic statistics dataset of the Brazilian cities

J. M. Toledo[1,2], Thiago J. M. Moura[1], R. D. A. Timoteo[2]

[1] Federal Institute of Paraíba (IFPB) – Avenida Primeiro de Maio, 720, Jaguaribe – João Pessoa – Paraíba – Brazil – CEP 58015-435

[2]Ministério do Trabalho e Emprego – Esplanada dos Ministérios, Bloco F - Brasília – DF.

jefferson.m.toledo@gmail.com, thiago.moura@ifpb.edu.br

Abstract. *Brazil is the fifth largest country in the world and is one of the most populous. These characteristics make it very important to summarize up-to-date data from all the Brazilian regions for decision-makers, in the public or private sectors. In this work, we obtain a unified dataset with statistical data for all cities in the country, integrating data related to population, economy, employment, education, and health. We show the process of extraction of the data from different public sources, the processing, and the generation of the dataset. We discuss the possible uses of the dataset, analyze the limitations of the proposed methodology, and discuss its possible evolutions.*

Resumo. *O Brasil é o quinto maior país do mundo e um dos mais populosos. Essas características tornam muito importante resumir dados atualizados de todas as regiões brasileiras para os tomadores de decisão, no setor público ou privado. Neste trabalho, obtemos um conjunto de dados unificado com dados estatísticos para todas as cidades do país, integrando dados relacionados à população, economia, trabalho, educação e saúde. Mostramos o processo de extração dos dados de diferentes fontes públicas, o processamento e a geração de uma única tabela. Discutimos os possíveis usos do conjunto de dados, analisamos as limitações da metodologia proposta e discutimos suas possíveis evoluções.*

1. Introduction

Recently, we are observing growth in the use of data in planning and decision-making processes both in the public [Mergel et al. 2016, Maciejewski 2017] and in the private sectors [Wang and Zong 2023, De Laat 2018]. The increase in the storage and the power of computer processing is allowing improvement in data mining. The consequent availability of data is bringing benefits to society, increasing the accuracy of the decision-making process, accelerating the planning stages in companies, and reducing costs [Maciejewski 2017].

Brazil has 5,570 municipalities [IBGE 2023a] spread in an area of more than 8 million squared quilometers[IBGE 2023d]. In this vast territory, the country shows a diversity of climates (from temperate to equatorial), biomes (from the Amazon rain forest to semi-arid) and social behaviors[IBGE 2023c, IBGE 2023b]. While the Brazilan largest city (São Paulo) has more than 12 million habitants, some municipalities has as few as a

thousand people living. Given these characteristics, it is very important, both for public and private initiatives, to understand the socioeconomic variables of the country's cities. Although important, it is not a simple task for the same reasons: how can one obtain a summarized dataset that brings together information from different socioeconomic fields?

Considering these aspects, in this work we propose the construction of a unified dataset containing statistical data for all cities in Brazil. We use data acquired by some public organs: the Brazilian Institute of Geography and Statistics [IBGE 2023e], the Institute of Applied Economic Research (IPEA)[IPEA 2023], and the Brazilian Health Ministry[da Saúde 2023]. The data were extracted from public APIs (application programming interfaces) or through CSV (comma-separated values) files downloaded from the institutes' web pages and, then, the tables were joined to obtain the final dataset.

It is noteworthy that several works in literature use socioeconomic indicators proposed in this work to explain and describe various phenomena, for example, the ones related to public health, agroindustrial production, and education in Brazil [Fischer et al. 2007, Santos and Barbosa 2017, Jaen-Varas et al. 2019] and in other countries [Tang et al. 2022, Zhang et al. 2022, Rodríguez-Rueda et al. 2021]. Thus, the present work can also contribute to scientists and searchers in several fields.

It is also essential to observe that the features used in this work are periodically updated by the responsible institutes. It is worth calling attention to the fact that the sources used are not uniform, which enlarged the effort in treating and integrating described below. Thus, the obtained dataset has the highest territorial granularity in the Brazilian Republic and can be frequently updated, representing, then, an essential contribution to data science in the country since it provides multiple data that can be used for developing studies.

This work is organized as follows. We briefly review the related works in Section 2. In Section 3, we list the data sources used in the project and describe the process of integration and acquisition of the result dataset, which is analyzed in Section 4. The examples of uses are described in Section 5, while the limitations and future works are analyzed. Finally, we present the conclusions in 7.

2. Related works

It is possible to find in the literature some recent works that aimed to obtain datasets with characteristics of Brazilian cities, especially for specific sociodemographic fields.

In which concerns the population estimate, [de Albuquerque et al. 2022] obtained a dataset that contains this data for Brazilian cities. The authors obtained the estimated population for sex and groups of ages in small geographic areas in the country filling a lack in the public data [de Albuquerque et al. 2022].

With respect to the data related to basic education in Brazil, [Barros et al. 2022] obtained a unified dataset for the years 2020 and 2021, using public data made available by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). The authors improved and updated some predecessor initiatives like the one presented in [Conte 2019].

We can also find in the literature the achievement of datasets related to public health in the country. This was done, for example, by [Clarindo et al. 2020], who built

a dataset named QualiSUS. The authors obtained data from DATASUS [1], standardizing the variables, suppressing invalid data, and aggregating information to the QualiSUS. On the other hand, focusing on the COVID-19 pandemic, [Gonçalves et al. 2021] obtained a dataset with vaccination data, which was cross-checked with other sources to guarantee its correctness.

In order to analyze the efficiency of public spending in Brazilian cities, [Davis 2022] built a dataset containing data from different fields like financial data, education, public security, and public health. The authors analyzed the effectiveness of public governance by evaluating the correlation between socioeconomic variables and the corresponding expenditures by the cities.

In our work, we aimed to build an up-to-date dataset with Brazilian cities variables, coming from several socioeconomic areas, just keeping quantities that are periodically updated by those responsible. On the other hand, we used only variables that are calculated for the Brazilian cities, to obtain only territorial granular data. Thus, we can conclude that this work represents an essential contribution since it integrates data from different sources and fields, providing a unified dataset with data for all Brazilian cities and simplifying the work of data scientists in the country.

3. Sources and ETL process

In this section, we describe the sources used to build the BrStats dataset of Brazilian cities statistics and the methods used to obtain it.

3.1. Data sources

The data sources used in this work are briefly described in what follows. It is worth calling attention to the fact that we only considered public sources online available, such that the development of this work finds no obstacle in Brazilian laws.

3.1.1. IBGE

IBGE is the main provider of statistical data in Brazil [2]. The main objective of the institute is to offer a complete view of the county, through the production, analysis, and consolidation of statistical and geographic information.

The institute maintains an API called SIDRA (Sistema IBGE de Recuperação Automática) [3], from which one can obtain the data from its numerous surveys and research. In order to simplify the search and the use of SIDRA API, IBGE gives a corresponding code to identify each table and variable. In table 1, we list the table codes, variable names, and the corresponding variable code. We also inform the year of the last information on SIDRA.

While IBGE table n. 6579 brings the estimated population contingent for the Brazilian cities, table n. 6449 has information obtained in a continuous statistical pol

[1] https://datasus.saude.gov.br/

[2] https://www.ibge.gov.br/acesso-informacao/institucional/o-ibge.html

[3] https://apisidra.ibge.gov.br/

Table code	Table subject	Variable	Variable code	Last information
6579	Resident population	Population	6579	2021
6449	Companies and other organizations	Working staff	707	2021
		Salaries	662	2021
		Companies	2585	2021
		Salaried staff	708	2021
1301	Surface area	Area	615	-
5938	Public finance	Gross Domestic Product	37	2020
5457	Agriculture	Cultivated area	8331	2021
		Harvested area	216	2021
		Agricultural production	215	2021
74	Livestock	Livestock production	215	2021

Table 1. Description of the data extracted from IBGE.

named ”Cadastro Central de Empresas”, in which the institute obtains data about employment in the country (like the number of companies, total employed persons, salaried employed persons, and wages). Tables n. 1301 and 5938, respectively, bring information about the surface area of Brazilian cities and their Gross Domestic Product (GDP). Finally, tables n. 5457 and 74 deal with agriculture and livestock production in the country, bringing data about cultivated and harvested area and total production.

3.1.2. IPEA

IPEA is a Brazilian public institution that provides support to the federal government regarding fiscal, social and economic public policies [4]. The institute publishes more than 250 studies annually, aiming to improve the efficiency of government decisions and, as a consequence, help the social, economic, and structural country's development.

The institute also provides a public API to simplify the data extraction[5], which was used in this work. We summarize, in Table 2, the data extracted from the API, listing the name of the table, the variable, and the year of the last information.

Table name	Variable	Last information
EXPORTACAO	Exports	2021
IMPORTACAO	Imports	2021
RECTOTCH	Revenue	2021
RTRCORTOM	Current transfers	2021
RTRKTOM	Capital transfers	2021

Table 2. Description of the data extracted from IBGE.

The tables named EXPORTACAO and IMPORTACAO bring information about

[4]https://www.ipea.gov.br/portal/categorias/110-conheca-o-ipea/13764-who-we-are
[5]http://www.ipeadata.gov.br/api/

the foreign trade of Brazilian cities, which is correlated with their economic production and welfare. Tables RECTOTCH, RTRCORTOM, and RTRKTOM are related to municipal finance, informing, respectively, the cities' total revenue, current and capital transfers. The last two variables are part of the balance of payments of public accounting and, while the capital transfers are related to the changes in changes in cities' ownership of assets, the current transfer quantifies the net incomes in the public sector[Bandy 2018].

3.1.3. DATASUS

The Brazilian Health Ministry maintains a web application called DATASUS which allows us to extract public data[6]. The data was downloaded directly from the web page in a CSV file.

In this work, we considered the information about born children and child death, such that we can estimate child mortality in Brazilian cities. The correlation between child mortality and a variety of socioeconomic variables has been studied in the literature[Fischer et al. 2007]. So, this quantity can be related to the health condition of a city and, as a consequence, it helps to measure the human development of a locality [Fischer et al. 2007]. Therefore, it is an important feature to compose the dataset obtained in this work.

3.2. Extract, transform and load (ETL)

In Fig 1, we depict the extraction and transformation process of the data used in this work, which was developed using Python programming language [Van Rossum et al. 1995].

In the left part of the figure, we show the extraction of the data from multiple origins. As discussed above, while the IBGE data is extracted from the Sidra web API[7] and the IPEA data is obtained through Ieadata API[8], the data acquired from DATASUS is downloaded in CSV files.

After the extraction process, the tables have been aggregated by their origins: the first group is the data coming from IBGE, the second group is the data from IPEA and the third one represents the public health data.

Finally, the three tables were integrated to obtain the final dataset. In this stage, some treatment needs to be executed, due to the lack of standardization of data arriving from different sources. It is necessary to mention that the IBGE adopts a numeric code for each city in Brazil, which contains seven digits, the last one being a check digit. The data coming from the DATASUS only contains six digits and, then, some transformations needed to be done.

4. The resulting dataset

After the ETL process described in Sec. 3, we obtain a dataset aggregating statistics for all Brazilian cities from 2012 to 2021. The obtained dataset has 36,315 rows and 21 columns related to socioeconomic variables. The BrStats dataset is

[6](https://datasus.saude.gov.br/informacoes-de-saude-tabnet/)

[7]https://apisidra.ibge.gov.br/

[8]http://www.ipeadata.gov.br/api/

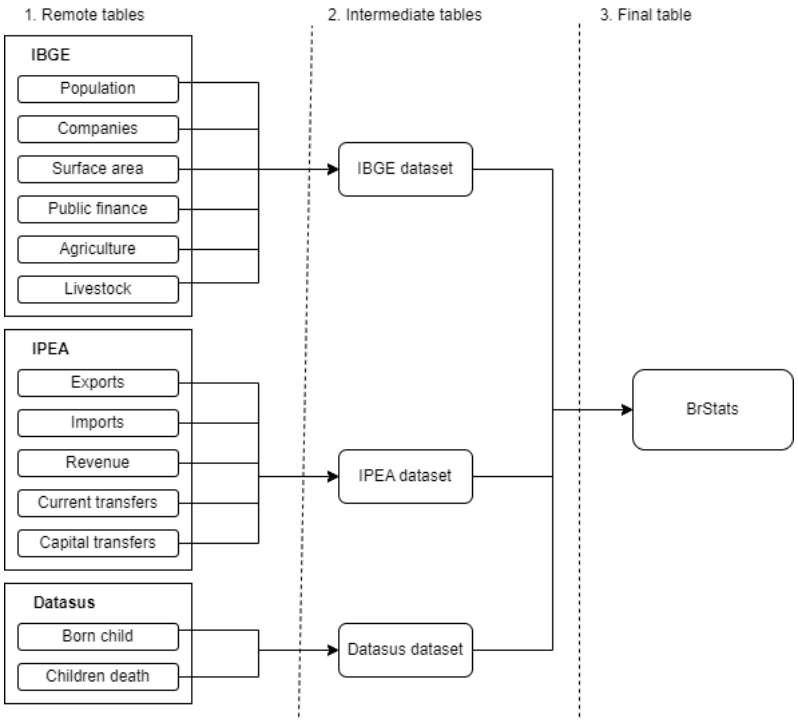

Figure 1. Process of extraction and integration to obtain the resulting dataset.

publicly available through the address `https://drive.google.com/file/d/1HBI054CnCCAX7kzAmEUUH2olhix2eiqe/view?usp=sharing`.

4.1. Data dictionary

The data dictionary of the obtained dataset is represented in Table 3, in which we list the final table columns, the corresponding data type, the unit of measurement, the maximum and minimum values of the variables, and, finally, a synthetic description. All the variables are important indicators of socioeconomic and demographic aspects of the Brazilian cities.

It is important to notice that the IPEA data (exports, imports, revenue, current transfers, and capital transfers) are not available for all cities in the considered period and, thus, the BrStats user needs to observe the possibility of using or not these data.

5. Examples and possible uses

The BrStats dataset can be used in a large number of proposes. Since the variables extracted may reflect the local culture and economy in Brazilian cities, they can be used to perform data analysis or as features in machine learning models.

Initially, we can observe that the dataset itself has a large number of variables that describes the socioeconomic profile of country regions. These variables can be used to generate dashboards and statistics of the municipalities or can be used to obtain an even larger number of consequent variables. In Fig. 2, for example, we calculated the number of salaried workers per company for all the cities and depicted this information in a filled map of Brazil. This information is an important signal of companies' automation and the

Column name	Type	Unit	Min value	Max value	Null count	Description
Ano	int	-	2016	2021	0	year
CDMunicipio	str	-	-	-	0	IBGE code of the city
Populacao	int	-	771	1.23×10^7	30	City population
PessoalOcupado	int	-	20	7.32×10^6	30	Working staff
PessoalAssalariado	int	-	3	6.55×10^6	30	Salaried staff
VrSalarios	int	10^3 R$	35	3.21×10^8	30	Total salaries sum
PIB	int	10^3 R$	11679	7.63×10^8	5595	Gross Domestic Product
QtEmpresas	int	-	3	638246	30	Number of companies
AreaPlantada_h	int	$10^4 m^2$	0	9483	72	Cultivated Area
AreaColhida_h	int	$10^4 m^2$	0	9483	72	Harvested area
VlProducaoAgricola	int	10^3	0	9975	72	Agricultural production
VlProducaoPecuaria	int	10^3	0	8379	36	Livestock production
Area	float	km^2	3.6	159533.4	30	Surface area
Povoamento	float	person/km^2	0.03	14656.55	30	Nr. of people by km^2
Importacoes_US$	float	US$	1	1.52×10^{10}	20375	Total values of imports
Exportacoes_US$	float	US$	3	1.31×10^{10}	21392	Total values of exports
Receitas_R$	float	US$	9.36×10^6	6.48×10^{10}	11173	Total revenue
Transferencias_correntes_R$	float	R$	0	2.29×10^{10}	243	Current transfers
Transferencias_capital_R$	float	R$	0	8.37×10^8	243	Capital transfers
NrNascimentos	int	-	0	169299	0	Nr. of children born
NrObitosInfantis	int	-	0	1894	0	Nr. of children deceased

Table 3. Data dictionary.

Figure 2. Filled map of the salaried workers per company in Brazil states and in the cities of the Brazilian Northeast region.

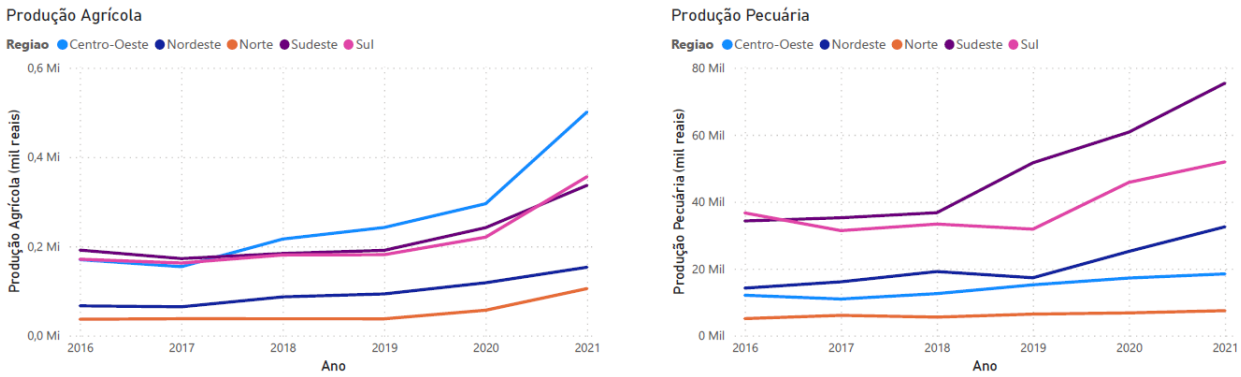

Figure 3. Line graph for agricultural activity in Brazil.

use of personal labor, and these data can be used, for instance, in the distribution of labor inspections in the country.

Also based on the BrStats content, we can analyze the profile of agricultural activity in Brazil, which is an important component of the country's economy. In Fig 3, we represent time series charts, respectively, for agricultural and livestock production for all Brazilian regions. We can observe, for example, that the monetary earnings from agriculture in the Midwest region have grown in past years and become the highest in Brazil. On the other hand, the Southeast and South regions are the largest livestock producers in the country.

In Figure 4, we represent bar graphs for GDP per working person and for mean salary per working person in Brazilian regions in 2021. We can observe that, while the mean salary per laborer is higher in Midwest and Southeast regions, the GDP per working person is higher in the North and Midwest regions. We so can state that wages are lower in the North region, but it composes a large part of the region GPD.

In addition to the examples presented, it is possible to carry out a large number of analyses from the BrStats, which are beyond the scope of this work.

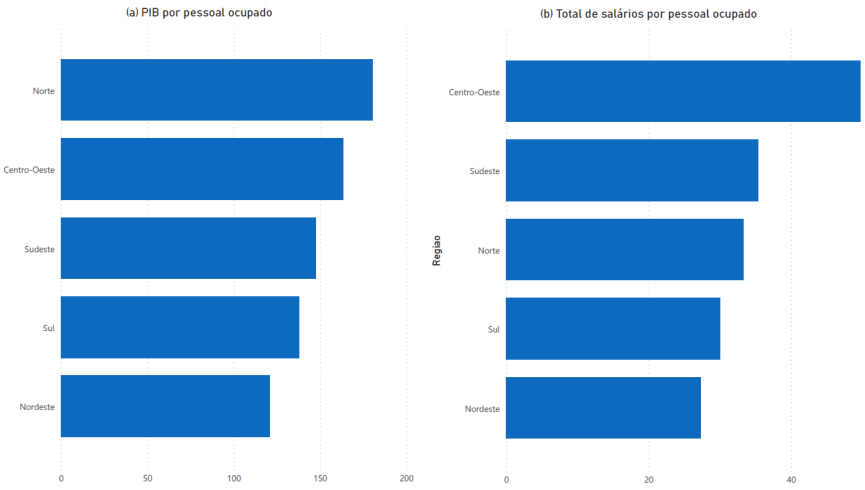

Figure 4. Bar graphs for GDP per working persons and for salaries per working person in 2021.

As discussed, the variables obtained in BrStats can be used as features in machine learning models. Aiming to help improve the effectiveness of the public policies in Brazil, in future contributions, we intend to use BrStats variables, together with other features, to predict labor diseases and accidents in the country.

Given all the above, we can state that the dataset obtained in this work can be used in a variety of projects and studies in multiple areas of knowledge which need systematic information on Brazilian cities.

6. Challenges, limitations, and perspectives

The necessity of data extraction from multiple non-standardized sources brings some challenges to the methodology described above, which, however, can be the subject of future works.

Since we proposed to make available a public dataset of up-to-date features of Brazilian cities and used data from public sources, we need the data sources to be updated by the source institutes. This fact can be impacted in some situations, such as the COVID-19 pandemic, during which some data had the update delayed.

Another challenge comes from the extraction process. As we used data from multiple sources and more than one method (API and CSV files), the change in the source layout or URL can impact the initial process in the ETL. Thus, in future contributions, we will need to observe alterations introduced by the Brazilian statistics institutes used as sources and update the extraction of the variables.

As discussed, Brazil is a big country with many cities. This fact makes obtaining data in some small regions very challenging. Consequently, some cities are not included in the research by the statistical organs. This fact brings some null variables in the shared dataset, which need to be dealt with by the users.

It is worth calling attention to the fact that this work does not exhaust the possibilities of aggregation of variables for the Brazilian municipalities. On the other hand, a

continued effort is to search for more public variables and to obtain derived features from the ones already obtained.

7. Concluding remarks

In this work, we obtained a single dataset containing socioeconomic variables of the Brazilian cities, coming from multiple sources and covering different areas of knowledge. The problem of non-standardization of columns in the sources was treated in the transform stage ETL, which allowed us to obtain a table with a unique key for each municipality. The final table has data related to the economy, population, and public cities' health.

Brazil is one of the largest countries in the world, full of socioeconomic inequalities and geographic diversity in its territory. The core contribution of this work is, then, to represent such diversity in terms of data, simplifying the data mining process for data scientists, government, and companies.

As can be seen, the data update can be very challenging, since it depends on the research institutes that produce and make the variables available. In the same way, the completeness of the dataset depends on these organs. So, in possible future contributions, it will be needed to verify the modifications in the data sources, update the values, and seek new variables.

The effort for building a data-oriented culture is rising around the world and our work intends, by providing an organized and systematic dataset, to encourage this culture in Brazil.

References

Bandy, G. (2018). *International public financial management: Essentials of public sector accounting*. Routledge.

Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.

Clarindo, J., Fontes, W., and Coutinho, F. (2020). Qualisus: um dataset sobre dados da saúde pública no brasil. *Proceedings of 2nd SBBD DSW*, 2:418–428.

Conte, V. d. S. (2019). Mineração de dados educacionais para avaliar os fatores que influenciam no desempenho de candidatos do enem.

da Saúde, M. (2023). Datasus. `https://datasus.saude.gov.br/`. Accessed: 2023-06-15.

Davis, P. G. (2022). Indicadores e dados municipais: Um banco de dados para avaliar a eficiência das despesas públicas. In *Anais do IV Dataset Showcase Workshop*, pages 79–90. SBC.

de Albuquerque, A. F., Barros, A. N., Alencar, A., Nascimento, A., Bittencourt, I. M., and Mello, R. F. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34. SBC.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology*, 31(4):525–541.

Fischer, T. K., Lima, D., Rosa, R., Osório, D., and Boing, A. F. (2007). A mortalidade infantil no brasil: série histórica entre 1994-2004 e associação com indicadores socioeconômicos em municípios de médio e grande porte. *Medicina (Ribeirão Preto)*, 40(4):559–566.

Gonçalves, M. V. F., dos Santos, J. S., Ferreira, C. Z., Zavaleta, J., da Cruz, S. M. S., and Sampaio, J. O. (2021). Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. In *Anais do III Dataset Showcase Workshop*, pages 148–159. SBC.

IBGE (2023a). Brasil — cidades e estados - ibge. `https://www.ibge.gov.br/cidades-e-estados`. Accessed: 2023-06-15.

IBGE (2023b). Conheça o brasil - biomas brasileiros. `https://educa.ibge.gov.br/jovens/conheca-o-brasil/territorio/18307-biomas-rasileiros.html`. Accessed: 2023-06-15.

IBGE (2023c). Conheça o brasil - clima. `https://educa.ibge.gov.br/jovens/conheca-o-brasil/territorio/20644-clima.html`. Accessed: 2023-06-15.

IBGE (2023d). Ibge - Áreas territoriais. `https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html?=&t=o-que-e`. Accessed: 2023-06-15.

IBGE (2023e). Instituto brasileiro de geografia e estatística - ibge. `https://www.ibge.gov.br/`. Accessed: 2023-06-15.

IPEA (2023). Instituto de pesquisa econômica aplicada - ipea. `https://www.ipea.gov.br/portal/`. Accessed: 2023-06-15.

Jaen-Varas, D., Mari, J. J., Asevedo, E., Borschmann, R., Diniz, E., Ziebold, C., and Gadelha, A. (2019). The association between adolescent suicide rates and socioeconomic indicators in brazil: a 10-year retrospective ecological study. *Brazilian Journal of Psychiatry*, 41:389–395.

Maciejewski, M. (2017). To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, 83(1_suppl):120–135.

Mergel, I., Rethemeyer, R. K., and Isett, K. (2016). Big data in public affairs. *Public Administration Review*, 76(6):928–937.

Rodríguez-Rueda, P., Ruiz-Aguilar, J., González-Enrique, J., and Turias, I. (2021). Origin–destination matrix estimation and prediction from socioeconomic variables using automatic feature selection procedure-based machine learning model. *Journal of Urban Planning and Development*, 147(4):04021056.

Santos, E. G. d. O. and Barbosa, I. R. (2017). Conglomerados espaciais da mortalidade por suicídio no nordeste do brasil e sua relação com indicadores socioeconômicos. *Cadernos Saúde Coletiva*, 25:371–378.

Tang, W., Wang, H., Lee, X.-L., and Yang, H.-T. (2022). Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, 240:122500.

Van Rossum, G., Drake, F. L., et al. (1995). *Python reference manual*, volume 111. Centrum voor Wiskunde en Informatica Amsterdam.

Wang, P. and Zong, L. (2023). Does machine learning help private sectors to alarm crises? evidence from china's currency market. *Physica A: Statistical Mechanics and its Applications*, page 128470.

Zhang, C., Dong, H., Geng, Y., Liang, H., and Liu, X. (2022). Machine learning based prediction for china's municipal solid waste under the shared socioeconomic pathways. *Journal of Environmental Management*, 312:114918.