AESOP - Data descriptor

Release 0.1

Juliane

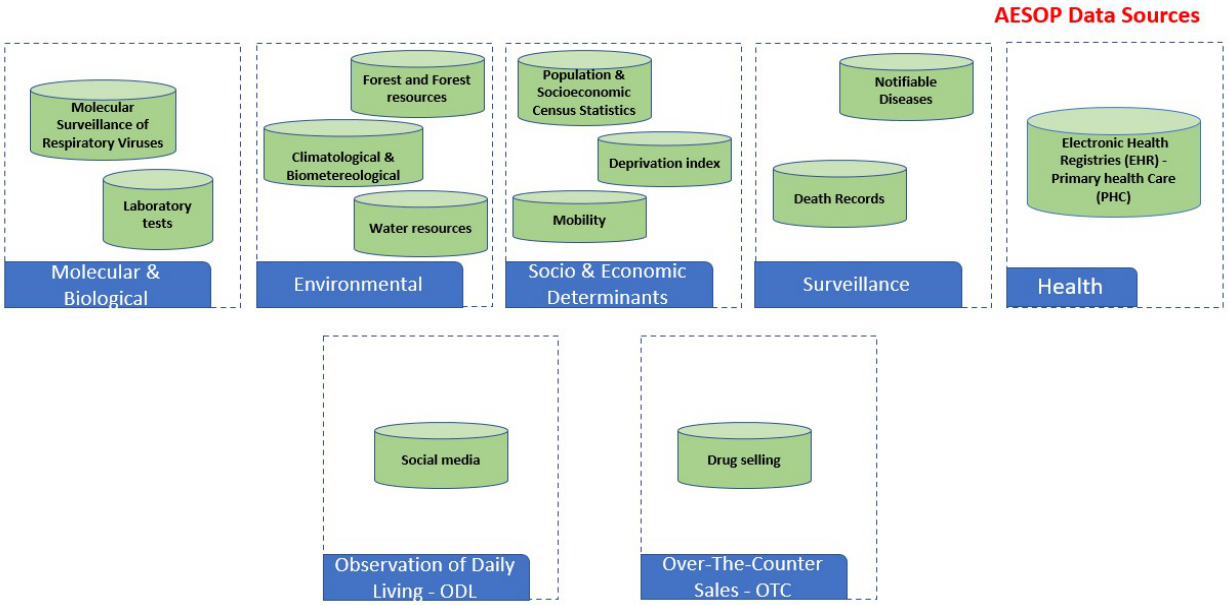
May 01, 2024

CONTENTS

This documentation provides a description of all databases used in the Alert-Early System of Outbreaks with Pandemic Potential (AESOP) project.

ÆSOP aims to be a data-driven surveillance system for the early alert of infectious disease outbreaks and their potential threats. To achieve this goal, the system will be built on the integration of existing multimodal data routinely collected from different sources, as well as on a set of analytical models for pandemic detection and transmission forecast.

ÆSOP relies on databases that report and describe health, biological, pharmaceutical, environmental, climate, human mobility, social media, and socioeconomic features of Brazil.



Schematic overview of the AESOP data source.

In the next sections, we explain each database associated with this project, including the repository they are stored in, access permissions and an overview of the data files and their formats. In addition, we provide detailed descriptions of the methods used to collect the data, the computational processing, and the potential for data reuse.

Links, references, codes, program, or data processing workflow is provided to facilitate understanding or use the data.

Note: This documentation is under active development.

CHAPTER

ONE

CONTENTS

1.1 Health data

Updated: 2024-04-29

1.1.1 Primary Health Care (PHC)

Description

Worldwide, various countries and health systems adopt the use of Primary Health Care (PHC) data integrated with traditional health surveillance systems to enhance sentinel surveillance and early warning of disease outbreaks capabilities (1,2).

Brazil's National Information System on Primary Health Care (SISAB) harbors data on all publicly funded PHC encounters in the country, coded by either the International Classification of Diseases (ICD-10) or the International Classification of Primary Care (ICPC-2) (3).SISAB is a decentralized system maintained by the Ministry of Health (MoH), with every Brazilian municipality running a version of it. Since 2016, its use has become mandatory for receiving federal funds for PHC.

It is estimated that the Brazilian national PHC system covers around 75% of the population (4). In 2023, the system registered an average of 39,6 million encounters per month, and all of the 5,570 municipalities registered encounters in the system.

Methods of data collection

The AESOP Project maintains a SISAB database ranging from January 2017 up to now, with regular weekly updates through a specific VPN connection. Data is obtained under the permission of the MoH, and data collection is approved by the Ethical Review Board of Oswaldo Cruz Foundation - Brasília Regional Office, CAAE 61444122.0.0000.0040.

The data is aggregated and non-identified. Each line represents the weekly count of PHC encounters per reason of encounter (coded by either ICD-10 or ICPC-2), city of encounter, gender, and age group.

For establishing the early warning system based on syndromic surveillance, we grouped a broad set of codes into three categories for defining each of the following syndromes: influenza-like illness (ILI), dengue-like syndrome (DLS) and diarrhea. We aimed at guaranteeing enough sensibility for outbreak detection. The list of codes can be found at: https://github.com/cidacslab/AESOP-Data-Documentation/tree/main/DataPipeline/documentation

Data access information

Although the AESOP Project has access to weekly updates following an exclusive permission by the MoH, open access to monthly data is available through the following websites, maintained by the MoH:

- https://sisaps.saude.gov.br/painelsaps/
- https://sisab.saude.gov.br/

Data Quality Index (DQI)

Since SISAB is an administrative database designed and maintained for governance and accountability purposes, evaluating the quality of the available data when used for early warning of outbreaks is crucial to ensure accurate interpretation of epidemiological analysis results.

We established a Data Quality Index (DQI) based on the quantitative assessment of the completeness, timeliness, and consistency of SISAB. The DQI is continuously monitored in an 8-week rolling window baseline. In this context, these three quality dimensions are defined as follows:

Completeness: defined as the proportion of weeks in each 8-week rolling window with registries of PHC encounters, provided there is no Consecutive Missing in the last two weeks of the 8-week rolling window.

Timeliness: represented by the number of weeks between the PHC encounter and the date the data was entered into the system. According to SISAB primary use, PHC encounters may be registered with up to a 4-week lag. For the early warning system purpose, such lag is unacceptable as it would yield lack of opportunity for outbreak detection.

Consistency: defined as a minimum number of PHC encounters registered weekly. As administrative health care databases often present with artificial aberrations in time series analysis due to events affecting health services availability and patient behavior, we considered the data to be consistent if the number of registered encounters were within the range of two standard deviations (SD) from the average number of encounters in the last 7 weeks.

In 2023, 4,458 (80%) municipalities showed completeness above 85% for at least 80% of epiweeks and 3,969 (71.2%) showed timelines above 75% for at least 80% of epiweeks. These results yielded a total of 3,713 (66.6%) with a valid DQI for issuing early warning during 80% of 2023 epiweeks.

Data-specific information

Data is one dataset where each line is corresponds to one Brazilian city in one month of the studied period. The variables are city population, location, number of PHC teams and facilities and the number of PHCE in each group of interest on that month .

Limitations of PHC dataset

SISAB does not carry the entire patient's electronic health record and, for each encounter, only one diagnostic code is retrieved in the database, leading to imprecision in classifying cases.

Moreover, detailed information on symptoms, tests, and treatments is unavailable.

Variable list for PHC database

References

(1) Bagaria J, Jansen T, Marques DFP, Hooiveld M, McMenamin J, de Lusignan S, Vilcu AM, Meijer A, Rodrigues AP, Brytting M, Mazagatos C, Cogdale J, van der Werf S, Dijkstra F, Guiomar R, Enkirch T, Valenciano M, I-MOVE-COVID-19 study team. Rapidly adapting primary care sentinel surveillance across seven countries in Europe for COVID-19 in the first half of 2020: strengths, challenges, and lessons learned. Euro Surveill. 2022;27(26):pii=2100864. doi:10.2807/1560-7917.ES.2022.27.26.2100864.

(2) Prado NMBL, Biscarde DGDS, Pinto Junior EP, Santos HLPCD, Mota SEC, Menezes ELC, Oliveira JS, Santos AMD. Primary care-based health surveillance actions in response to the COVID-19 pandemic: contributions to the debate. Cien Saude Colet. 2021l;26(7):2843-2857. doi: 10.1590/1413-81232021267.00582021.

(3) Cerqueira-Silva T, Oliveira JF, Oliveira VA, Florentino PTV, Sironi A, Penna GO, Ramos PIP, Boaventura VS, Barral-Netto M, Marcilio I. Early warning system using primary healthcare data in the post-COVID-19-pandemic era: Brazil nationwide case-study. Pre-print available at medRxiv: doi: 10.1101/2023.11.24.23299005

(4) Sellera PEG, Pedebos LA, Harzheim E, Medeiros OL de, Ramos LG, Martins C, D'Avila OP. Monitoramento e avaliação dos atributos da Atenção Primária à Saúde em nível nacional: novos desafios. Ciênc Saúde Coletiva. 2020;25(4):1401–12. doi:10.1590/1413-81232020254.36942019

Contributors

George Barbosa	Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
Izabel Marcilio	Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
Juracy Bertoldo	Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
Pilar Veras	Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
Vinicius Oliveira	Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil

1.2 Surveillance

1.2.1 Respiratory infections

Description

Acute Respiratory Infections are the third cause of mortality worldwide and are usually caused by virus or bacterial agents.

The influenza A(H1N1) pandemic of 2009 highlighted the importance of collecting information about disease severity in a standardized manner and having historical data available for countries to assess current influenza seasons in the context of previous ones[1]. Thus, since 2009, notification of all hospitalized cases of Severe Acute Respiratory Infection (SARI) is mandatory in Brazil.

Data on hospitalized SARI cases are inserted and collated in a centralized system, the "Sistema de Informação de Vigilância Epidemiológica da Gripe" (SIVEP-Gripe), which is under governance of the Ministry of Health (MoH).

[1] PAHO. Operational Guidelines for Sentinel Severe Acute Respiratory Infection (SARI) Surveillance. September 2014. https://www.paho.org/hq/dmdocuments/2015/2015-cha-operational-guidelines-sentinel-sari.pdf

Following the Covid-19 pandemic, mild Influenza-like illness cases started to be reported on e-SUS Vigilância Epidemiológica (e-SUS-VE), a new national COVID-19 reporting system, also centralized and ran by the MoH. Although currently in use, reporting of mild cases will probably be discontinued as mitigation of the public health emergency caused by Covid-19 goes on.

Both e-SUS-VE and SIVEP-Gripe include suspected and confirmed cases as reported by public health and private services, and case definitions for notification are:

Hospitalized SARI cases

A hospitalized patient presenting with the acute onset of fever and cough OR sore throat AND with one of the following: acute respiratory distress, dyspnea, or O2 saturation < 95%. Any patient with the above symptoms and died, regardless of hospitalization.

Influenza-like illness mild cases

Any patient presenting with the acute onset of fever (ut to 5 days) and cough OR sore throat. All notified cases should go under laboratory investigation in order to ascertain the causative agent for confirming or discarding infectious diseases.

Based on the previous description, we collected two databases: the Flu syndrome database (FSdb) and the Severe Acute Respiratory Infection database (SARIdb) to perform studies on AESOP research.

Data access information

The FSdb and SARIdb data are licensed under a Creative Commons Attribution License cc-by (version 4.0)[2] and[3]. Additionally, the databases are publicly available and published by the Ministry of Health of Brazil. Therefore, no approval by an ethics committee is required to use this data, according to Resolutions 466/2012 and 510/2016 (article 1, sections III and V) from the National Health Council (CNS), Brazil.

Methods of data collection

A python code is available on our PAMEpi's GitHub directory to download the FSdb and SARIdb data from the OpenDatasus, see details in[4].

Alternatively, FSdb can be downloaded direct from the OpenDatasus links, for the years 2020 to 2022:

- https://opendatasus.saude.gov.br/dataset/notificacoes-de-sindrome-gripal-leve-2020,
- https://opendatasus.saude.gov.br/dataset/notificacoes-de-sindrome-gripal-leve-2021,
- https://opendatasus.saude.gov.br/dataset/notificacoes-de-sindrome-gripal-leve-2022.

And SARIdb can be downloaded from OpenDatasus links as shown in https://github.com/cidacslab/AESOP-Data-Documentation/blob/main/Data%20Collection/SRAG.ipynb, for the years 2009 up to 2022.

[2] Ministério da Saúde. Open Datasus. Notificações de Síndrome Gripal. Retrieved 08 25, 2021, from https://opendatasus.saude.gov.br/dataset/casos-nacionais

[3] Ministério da Saúde. Open Datasus. Banco de dados SRAG. Retrieved 04 25, 2021, from https://opendatasus.saude.gov.br/dataset/bd-srag-2021/resource/42bd5e0e-d61a-4359-942e-ebc83391a137, https://opendatasus.saude.gov.br/dataset/bd-srag-2021

[4] Platform For Analytical Modelis in Epidemiology. (2022). PAMepi/PAMepi-scripts-datalake: v1.0.0 (v1.0.0). GitHub directory: https://github.com/PAMepi/PAMepi_scripts_datalake.git. Zenodo. . https://doi.org/10.5281/zenodo.6384641. Accessed: February 25, 2022.

Data-specific information

FSdb is organized according to each federal unit (i.e., state) of the Brazilian federation, while SARSdb files are organized annually (one per year). FSdb is available in .csv format, with a total of 30 variables and a size of around 15 GB in December 2021. In turn, SARSdb, also available in .csv format, contains 161 variables and a size of around 2 GB in December 2021

The FSdb dataset has a total of 30 columns and showed a total of 48,288,827 registries (rows) and a size of 14 GB in the last update of August 19th, 2021. The SRAIdb dataset has a total of 162 columns and showed a total of 1,264,480 registries (rows) and a size of 694 MB in the last update of August 19th, 2021 (informations regarding the files from 2020 up to 2021).

Limitations of the dataset

Several cases in the FSdb lack final classification. Every case in FSdb will have a final classification given by the epidemiological surveillance teams of the Secretariat of Health. However, given the number of registered patients, they may not be classified on time (some even closed and not analyzed anymore). To overcome such a difficulty, our team will apply a classification algorithm to give a pre-diagnosis of the cases in FSdb that have no final classification. The algorithm is applied to the information of symptoms that is available in the dataset.

Every case in SARI will have a final classification given by the epidemiological surveillance teams of the Secretariat of Health. However, given the number of registered patients, they may not be classified on time (some even closed and not analyzed anymore). To overcome such a difficulty, our team will apply a classification algorithm to give a pre-diagnosis of the cases in SRAG that have no final classification. The algorithm is applied to the information of symptoms that is available in the dataset.

Variable list

Given the extensive number of variables, we refer to https://pamepi.rondonia.fiocruz.br/en/sg_en.html and https://pamepi.rondonia.fiocruz.br/en/srag_en.html for a whole descriptions of the FS and SARI databases [5]and[6].

1.2.2 Arboviruses infections

Description

Arbovirus (arthropod-borne virus) infection is an infection caused by a viral spread to humans (and/or other vertebrates) through the bite of a blood-feeding arthropods (eg. flies, mosquitoes, ticks, etc). There are more than 250 species of arbovirus, including dengue, Zika, chikungunya, West Nile, Yellow fever, and others. An Arbovirus catalog is described here.

For the purposes of the AESOP project, we collected data of suspected Dengue, Zika and Chikungunya infections that were reported and are available in the Notifiable Diseases Information System (SINAN). The notification of every suspected case of these 3 diseases is mandatory, and case definition are as follows:

Dengue:

Suspect case: Any patient residing in (or having traveled to in the previous 14 days), an area with dengue or Aedes aegypti occurrence, and who presents with acute onset of fever (lasting up to 7 days) and 2 or more of the following symptoms: nausea/vomiting, rash, myalgia/arthralgia, headache, retro-orbital pain, petechiae, positive tourniquete test, leukopenia

Chikungunya

[6] da Silva, N.B., Valencia, L.I.O., Ferreira, A., Pereira, F.A., de Oliveira, G.L., Oliveira, P.F., Rodrigues, M.S., Ramos, P.I. and Oliveira, J.F., 2022. Brazilian COVID-19 data streaming. arXiv preprint arXiv:2205.05032.

Suspect case: Any patient presenting with sudden onset of high fever (>38.5°C) and acute onset of arthralgia or severe arthritis not explained by other conditions, residing in (or having visited in the previous 14 days) areas with chikungunya transmission, or who has an epidemilogic link to a confirmed imported case

Zika

Suspect case: Any patient presenting with pruritic maculopapular rash and one of the following: fever, conjunctival hyperaemia/non-purulent conjunctivitis, arthralgia/polyarthralgia, Periarticular edema.

In each disease, a patient will be assigned as a confirmed case of dengue, Zika or chinkungunya infection when a laboratory test (PCR, serology, virus isolation) is confirmed OR, when laboratory analysis is not possible, when the case is compatible with clinical presentation AND with epidemiologic link to a confirmed case AND for which no other diagnosis was confirmed.

Note: the Platform for analytical models in epidemiology - PAMEpi offers support in the documentation and collection of this database. More details in[Page 6, 4],[5] and[Page 7, 6].

References

1.3 Molecular Data

1.3.1 Description

Here, we detail data that will be collected from human clinical samples subjected to molecular approaches (including next-generation sequencing [NGS]) to characterize the genetic information on microorganisms present in these samples. We will use biosensors based on DNA sequencing to leverage microbial and viral identification and quantification in clinical samples, informed by the first-level signals coming from the surveillance of respiratory syndromes. Using molecular characterization we can accelerate the identification of an (re)emerging pathogen, also allowing for the identification of unknown microorganisms and viruses using shotgun-based approaches.

The potential of this data generated by the methodology based on DNA sequence (NGS) is very very good, since we could unravel, at the same time, uncultivable viruses and microorganism leading us to realize a very good characterization of the clinical or enviromental microbiota.

Several studies used molecular/metagenomic data to detect pathogens in human and environmental samples. Plase find a non extensive list of examples below (see References[1] to[5]).

[5] Platform for analytical models in epidemiology - PAMEpi. https://pamepi.rondonia.fiocruz.br/en/index_en.html. Accessed: February 25, 2022.

[1] Tschoeke, Diogo Antonio, Louisi Souza de Oliveira, Luciana Leomil, Amilcar Tanuri, and Fabiano Lopes Thompson. "Pregnant women carrying microcephaly foetuses and Zika virus contain potentially pathogenic microbes and parasites in their amniotic fluid." BMC Medical Genomics 10, no. 1 (2017): 1-5.

[5] Oranger, Annarita, Caterina Manzari, Matteo Chiara, Elisabetta Notario, Bruno Fosso, Antonio Parisi, Angelica Bianco et al. "Accurate detection and quantification of SARS-CoV-2 genomic and subgenomic mRNAs by ddPCR and meta-transcriptomics analysis." Communications biology 4, no. 1 (2021): 1-10.

1.3.2 Data access information

A specific data use agreement covering the transfer of raw sequence data outside AESOP team members is being created. Aggregated data tables containing the number or proportions of microorganisms in panel/sequencing data across samples will be made available to federation members as data is produced.

AESOP team will be able to access the raw and processed metagenomic data in the AESOP HPC facility. The bioinformatics pipelines, data analysis and visualization codes will be available at the AESOP GitHub repository.

We already have the necessary permits to perform the metagenomic sequencing focused in the microbial and viral community within the subjects. All genetic material sequenced in Brazil will be registered in SisGen platform. SisGen was implemented following the implementation of Law 13.123/2015 ("Biodiversity Law"), that regulates access to components of the genetic heritage, protection of and access to associated traditional knowledge and the fair and equitable sharing of benefits for the conservation and sustainable use of Brazilian biodiversity.

1.3.3 Methods of data collection

When a potential outbreak is identified, we request local health authorities the immediate sample collection of 100 patients who meet the flu case definition with <5 days of symptoms onset through a systematic convenience sampling process. The samples will be individually submitted to RT-qPCR for SARS-CoV-2, Influenza A (FluA), Influenza B (FluB), and RSV detection. In parallel, pools of 10 samples (500 ul of each) will be prepared and stored in aliquots for pathogen detection using NGS and future validation.

The biological samples will be stored to the closest regional unit of the Fiocruz Genomics Network where an outbreak is identified. The DNA sequence data will be stored in CIDACS HPC Cluster.

The samples will be collected based on outbreak alerts given by the integration of i) respiratory infection symptoms, ii) OTC, iii) social media, iv) environmental, and v) socioeconomic data. We expect that, despite being irregular, the sampling might have seasonal behavior.

1.3.4 Data-specific information

The molecular sensor of AESOP will generate dozens of terabytes of data. The data types will be:

- **Raw DNA short reads sequences data (.fastq): Fastq files have the DNA sequences per se and have each base pair quality score. This file type possesses the DNA sequences in one line, the sequence identifier, the following line, and the quality score for each base. In AESOP, we will generate 2 million sequences for a pool of 100 individuals. We will use the following sequencing approaches:**
    i) Respiratory Pathogen ID/AMR Enrichment Panel (RPIP); and
    ii) Metatranscriptomics. The first several well-known pathogens, such as viruses, bacteria, fungi, and antimicrobial resistance genes (AMRs) are targeted (Table 1). In the latter, we will be able to identify new pathogens.
- Processed DNA short reads sequences data (.fasta): Artifacts from the sequencing process and host sequences will be removed. In fasta files, only the DNA sequences are present. This file type possesses in one line the sequence identifier the following line the DNA.
- Assembled sequences (.fasta): Complete and draft genomes or contigs composes of these files. This file type possesses in one line the sequence identifier the following line the DNA. The difference from the short reads file is that assembled sequences fasta files has fewer sequences but longer.
- Raw annotation files (.csv, .txt, .tsv): Tabular data originated from the annotation software. One file is for taxonomic, and another for functional annotation. Usually, this type of file has four columns and hundreds of thousands of rows. All taxonomic or functional levels are provided in a single file.

- Processed annotation files (.csv, .txt, .tsv): Processed tabular data to perform statistical analysis and visualization. Usually, this type of file has hundreds to dozens of thousands of columns and hundreds or thousands of rows, depending on the taxonomic or functional level of the analysis. Each taxonomic or functional level generates a single file.
- Statistical analysis outputs (.csv, .txt, .tsv): Tabular data containing the results from several statistical analyses. It depends on the performed analysis, but files usually have dozens of columns and hundreds of rows.
- Visualization products (.png, .jpg, .tiff): Figures generated, e.g., stacked bar, scatter plots, heatmaps, multivariate analysis biplots (PCA, nMDS, CCA).

All data will be stored in the AESOP HPC facility, and the raw sequences data will be held in other servers in Fiocruz. All the codes to perform the bioinformatics analysis, including the pipeline implementation, the statistical analysis, and the data visualization, will be maintained in the AESOP GitHub repository.

Table 1 - Major pathogens and AMRs targeted in RPIP sequencing approach.

Pathogen type	Number of strains/genes	Examples
Viruses	42	Coxsackievirus A Human adenovirus B Influenza A viruses Rhinovirus SARS coronavirus SARS-CoV-2 (2019-nCoV)
Bacteria	187	Nocardia nova Ochrobactrum anthropi Pseudomonas stutzeri Prevotella melaninogenica Streptococcus agalactiae Treponema denticola Yersinia pestis
Fungi	54	Alternaria alternata Candida auris Exophiala dermatitidis Purpureocillium lilacinum Schizophyllum commune Trichosporon asahii
AMRs	1218	Antibacterials (Aminoglycosides, Carbapenems, Fluoroquinolones) Antimycobacterials (First-line: Isoniazids, Pyrazinamides. Second-line: Ethionamides, Aminoglycosides) Antivirals (Oseltamivir, Zanamivir, Peramivir, Laninamivir, Baloxavir)

1.3.5 Limitations of Biological dataset

Due to logistics, the most significant limitation will be assessing remote areas in Brazil to collect biological samples. Difficult-to-access regions, which may be the origin centers of outbreaks, will be monitored using other AESOP data. However, we will focus efforts on collecting patient samples in larger city centers close to those locations. The sampling location choice will consider how connected these areas are, including information about the road, airports, and fluvial networks.

References

Contributors

Pedro Milet Meirelles	Institute of Biology, Federal University of Bahia, Salvador, Brazil National Institute for Interdisciplinary and Transdisciplinary Studies in Ecology Evolution (IN-TREE), Salvador, Brazil

1.4 Pharmaceutical data

1.5 Social media data

1.6 Socioeconomic determinants

1.6.1 Brazilian deprivation index (IBP)

Description

It provides deprivation measures for each Brazilian municipality based on the 2010 Brazilian census data. It is used to evaluate health inequalities across the country. The 2010 Brazilian Census social and economic estimatimations are the basis for calculating the deprivation measure, available at[1] and[2].

The IBP index combines three factors:

1. the percentage of families with income per capita below half of the minimum wage;
2. the percentage of illiterate people older than 7 years old;
3. the percentage of people without adequate access to drinkable water, sewage, garbage collection, bathroom, or shower;

A complete documentation about the construction of the index is presented in[Page 11, 2]. The original data source and visualization can be found in[1] and[2].

Data access information

The data and documentation are available at the University of Glasgow[2]. The data and this report are distributed under the Creative Commons Share-Alike license (CC BY-SA 4.0) and can be freely used by researchers, policymakers, or members of the public.

[1] CIDACS. (2020, 09 01). IBP. Retrieved October 07, 2022, from https://cidacs.bahia.fiocruz.br/ibp/painel/.

[2] Allik, M., Ramos, D., Agranonik, M., Pinto Júnior, E.P., Ichihara, M.Y., Barreto, M.L., Leyland, A.H. and Dundas, R., 2020. Developing a small-area deprivation measure for Brazil. See https://researchdata.gla.ac.uk/980/.

Methods of data collection

The data can be download in[1] or[Page 11, 2]. The database describe the deprivation index for each municipality in 2010. No other update are available in the institutions that created it.

Data-specific information

The IBP for Brazilian municipalities dataset has a total of 15 columns (variables) and shows a total of 5566 registries (rows) and a size of 734 KB.

Limitations of IBP dataset

The IBP dataset includes only 3 measures of deprivation excluding others like employment, crime, health, education, and access to public services. It is also exclusively related to the year 2010 making comparison along time impracticable. The different population and ethnic groups (eg, indigenous peoples, quilombolas, riverside populations) are also not considered and there are some biases inherent to rural areas.

Variable list for IBP database

Original field name	Field label	Type	Category	Description
ip_cd_d	Localization code, in this case it is the same as the city.	Number	Uncategorized	Localization code, in this case it is the same as the city.
ip_cd_m	City code	Number	Uncategorized	city code
ip_nm_f	State name	String	Uncategorized	State name
ip_nm_r	Region name	String	Uncategorized	Region name
ip_cd_f	State acronym	String	Uncategorized	State acronym
ip_vl_f	State code	Number	Uncategorized	State code
ip_vl_p	Population according to IBGE (Brazilian Institute of Geography and Statistics) 2010 census	Number	Uncategorized	Population according to IBGE 2010 census
ip_nm_m	city name	String	Uncategorized	City name
ip_vl_n	Value of the deprivation index	Number	Uncategorized	Value of the deprivation index
ip_dcl_	Decile of the deprivation index	Number	Uncategorized	Decile of the deprivation index
ip_qntl_n	Quintile of the deprivation index	Number	Uncategorized	Quintile of the deprivation index
ip_prcnt_r	Percentage of people with per capita income below 1/2 minimum wage	Number	Uncategorized	Percentage of people with per capita income below 1/2 minimum wage
ip_prcnt_d	Percentage of illiterate people over 7 years old	Number	Uncategorized	Percentage of illiterate people over 7 years old.
ip_prcnt_m	Percentage of population in inappropriate homes.	Number	Uncategorized	Percentage of population in inappropriate homes.
ip_cd_c	empty/unknown			

To facilitate visualisation, we have also provided a data explorer that allows users to view the first rows the dataset along with metadata, including column descriptions, variable type, and variable harmonisation where applicable. This also allows broader re-use of this dataset, particularly since the original descriptors and data dictionaries are usually only available in Portuguese. Please, see https://pamepi.rondonia.fiocruz.br/en/ibp_en.html.

Note: the Platform for analytical models in epidemiology - PAMEpi offers support in the documentation and collection of this database. More details in[3],[4] and[5].

[3] Platform for analytical models in epidemiology - PAMEpi. https://pamepi.rondonia.fiocruz.br/en/index_en.html. Accessed: February 25, 2022.
[4] GitHub directory - PAMepi/PAMepi-scripts-datalake: v1.0.0 (v1.0.0). Zenodo. . https://doi.org/10.5281/zenodo.6384641. Accessed: February 25, 2022.
[5] da Silva, N.B., Valencia, L.I.O., Ferreira, A., Pereira, F.A., de Oliveira, G.L., Oliveira, P.F., Rodrigues, M.S., Ramos, P.I. and Oliveira, J.F., 2022. Brazilian COVID-19 data streaming. arXiv preprint arXiv:2205.05032.

References

1.7 Environmental data

1.7.1 Description

1.7.2 Data access information

1.7.3 Data-specific information

1.7.4 Limitations of Environmental dataset

References

1.8 Human Mobility

1.8.1 Road and Fluvial network

Description

In 2016, the Network Agency of the IBGE released a database to assess the accessibility of Brazilian municipalities through public transportation[1]. The primary focus of the data is to identify the most and least accessible areas within the country.

The data, provided by the IBGE, includes information on weekly trip frequency between pairs of municipalities, vehicle types, travel durations, and ticket costs. Much of this information was obtained from bus companies through questionnaires completed at bus terminals. For municipalities without bus stations, alternative contact points were used, such as ticket offices in commercial establishments, bus stops under municipal administrations, city halls, and direct communication with companies. However, relying solely on formal bus companies was insufficient to represent the true accessibility of cities, as there are municipalities where such companies are absent and no bus lines operate. As a result, informal and alternative modes of transportation (vans, station wagons, mini-buses, etc.) were also included in the research. These informal/alternative carriers typically step in to provide transportation in areas where official companies are unavailable. The data provided by these alternative transportation modes is of declaratory nature, with the database categorized into those registered in the National Register of Legal Entities (CNPJ) and those that do not disclose such information.

Recognizing the diversity of transportation options and the uneven distribution of road networks, data on water transport, predominantly found in the North Region of Brazil, was also gathered. Similar to road transport, formal companies operating at waterway terminals, boat cooperatives, and individual boat operators, with varying levels of formalization, were included in the data collection process.

[1] Ligações rodoviárias e hidroviárias: 2016 / IBGE, Coordenação de Geografia. - Rio de Janeiro: IBGE, 2017. 79p. ISBN 978-85-240-4417-5. Retrieved July 03, 2023, from https://biblioteca.ibge.gov.br/visualizacao/livros/liv100602.pdf.

Data access information

The data is openly available on the IBGE website and documentation is available at the 2016 roads and fluvial connections report[Page 14, 1].

Methods of data collection

This data does not need to be updated unless a newer report is published. Data is open access in the IBGE websites, particularly on the following link:

- http://geoftp.ibge.gov.br/organizacao_do_territorio/redes_e_fluxos_geograficos/

Data-specific information

The data gathered by the IBGE provides estimates of the weekly frequency of a standard normalized vehicle capacity. To have a normalized unity of measure of vehicle mobility between cities, first it was estimated the weekly frequency of vehicles by adding the number of weekly departures between each pair of municipalities. For connection pairs with only quarterly or monthly frequency, the sum of the departure frequencies is adjusted by multiplying it by 0.5 and 0.25, respectively, in order to align with the weekly frequency.

Since different types of vehicles have varying transport capacities, they are also assigned weights. Buses are considered as the standard measure (assigned a value of 1), while van and car frequencies are multiplied by 0.25. Regarding waterway vehicles, flying boats have their frequencies multiplied by 0.25, speedboats and catamarans are treated as analogous to buses (value 1), boats are multiplied by 1.5, and ships by 2. This approach enables us to estimate the number of passengers departing from one city to another, based on an average estimate of the number of passengers on a road bus. In this work we assume an average of 60 passengers per bus.

The database provided by IBGE is for the year 2016. It consists of 23 columns and 65,640 records (rows), in xlsx format, with a size of 9.2MB (dataset original name: Base_de_dados_ligacoes_rodoviarias_e_hidroviarias_2016.xlsx). The variables relevant to the project's analysis of road and fluvial mobility are listed in below.

Limitations

Uncertainties surrounding the availability and regularity of transportation services is still a problem in some Brazilian cities. Some municipalities lack public transportation altogether, and their population relies solely on private transport options. In other cases, municipalities have weaker and more tenuous integrated transportation systems, they do not meet the necessary temporal or spatial requirements for inclusion in IBGE research. Consequently, collecting reliable information about their transportation situation becomes challenging.

Another limitation of the data is the representation of connections between municipalities that share the same boundary or that are geographically closed. These municipalities share inter municipal transportations that also operates in another municipality. Examples will be shown later. The use of the NDTI database is one option to overcome this difficulty.

Variable list

List of variables extracted from the IBGE database to describe the intercity road and fluvial mobility in Brazil.

Original field name	Description
ID	Identificador único da ligação (Unique ID identifier)
COD_UF_A	Código da Unidade da Federação do município A do par de ligação (Federation Unit Code of Municipality A of the connection pair)
UF_A	Sigla da Unidade da Federação do município A do par de ligação (Acronym of the Federation Unit of municipality A of the connection pair)
CODMUNDV_A	Código do município A do par de ligação com dígito verificador (Code of municipality A of the connection pair)
NOMEMUN_A	Nome do município A do par de ligação (Name of municipality A of the connection pair)
COD_UF_B	Código da Unidade da Federação do município B do par de ligação (Code of the Federative Unit of municipality B of the connection pair)
UF_B	Sigla da Unidade da Federação do município B do par de ligação (Acronym of the Federation Unit of municipality B of the connection pair)
CODMUNDV_B	Código do município B do par de ligação com dígito verificador (Code of the municipality B of the connection pair)
NOMEMUN_B	Nome do município B do par de ligação (B municipality name of the connection pair)
VAR05	Frequência de saídas de veículos hidroviários no par de ligação (Frequency of waterway vehicles on the connecting pair)
VAR06	Frequência de saídas de veículos rodoviários no par de ligação (Frequency of road vehicles in the connection pair)
VAR07	Frequência total de saídas de veículos no par de ligação (Total frequency of vehicles in the connection pair)
VAR08	Longitude da sede municipal A do par de ligação
VAR09	Latitude da sede municipal A do par de ligação
VAR10	Longitude da sede municipal B do par de ligação
VAR11	Latitude da sede municipal B do par de ligação
Num_pass	Number of passengers in the connection pair. Variables that can be created by multiplying the frequency of vehicles per 60.

Note: the Platform for analytical models in epidemiology - PAMEpi offers support in the documentation and collection

of this database. More details in[2],[3] and[4].

References

[2] Platform for analytical models in epidemiology - PAMEpi. https://pamepi.rondonia.fiocruz.br/en/index_en.html. Accessed: February 25, 2022.

[3] GitHub directory - PAMepi/PAMepi-scripts-datalake: v1.0.0 (v1.0.0). Zenodo. . https://doi.org/10.5281/zenodo.6384641. Accessed: February 25, 2022.

[4] da Silva, N.B., Valencia, L.I.O., Ferreira, A., Pereira, F.A., de Oliveira, G.L., Oliveira, P.F., Rodrigues, M.S., Ramos, P.I. and Oliveira, J.F., 2022. Brazilian COVID-19 data streaming. arXiv preprint arXiv:2205.05032.