Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.3570

Workflow for the acquisition, processing, and dissemination of Brazilian public data focused on education

Abílio Nogueira Barros [**Universidade Federal Rural de Pernambuco** | ***abilio.nogueira@ufrpe.br***]
Aldéryck Félix de Albuquerque [**Cesar School** | ***derycck@gmail.com***]
Andrêza Leite de Alencar [**Universidade Federal Rural de Pernambuco** | ***andreza.leite@ufrpe.br***]
André Nascimento [**Universidade Federal Rural de Pernambuco** | ***andre.nascimento@ufrpe.br***]
Ibsen Mateus Bittencourt [**Universidade Federal de Alagoas** | ***ibsen@feac.ufal.br***]
Rafael Ferreira Mello [**Universidade Federal Rural de Pernambuco** | ***rafael.mello@ufrpe.br***]
Received: 23 July 2023 • **Published:** 5 April 2024

Abstract This article aims to demonstrate the process of creating public databases focused on the educational and population areas. It describes the process of obtaining data from official government sources such as INEP (National Institute for Educational Studies and Research) and IBGE (Brazilian Institute of Geography and Statistics), the procedures for data adaptation and optimization to create their historical series, as well as the best practices followed for their development and the generated metadata. Highlighting the specificities between the themes of education and population, reporting their challenges and peculiarities of each dataset. It also reports the results that can already be directly obtained from each dataset and how, when combined, they can track indicators of the National Education Plan, one of the largest Brazilian public policies focused on education.

Keywords: Datasets,Open data, Public data, Education data, Smart government

1 Introduction

The pursuit of improving the quality and efficiency of public education policies is a constant challenge for governments and institutions. In this context, accurate and comprehensive data play a fundamental role in developing metrics and indicators that underpin the construction of robust educational plans and in evaluating the effectiveness of implemented actions. Among the essential data for this purpose, information about student enrollments in different municipalities, coupled with knowledge of population volumes segmented by age, provides a valuable source of information to guide educational strategies. It also helps in devising metrics and indicators that serve as a starting point for building models supporting the development of educational plans and monitoring their implementation.

These informations enable a better understanding of the demand and supply of education in different regions, identifying areas with a higher concentration of students in specific age groups and where educational needs are more pressing. Furthermore, such data enables the analysis of the distribution of the school-age population, assisting in projecting future student contingents and preparing the educational system to accommodate these demands. The segmentation by municipalities also allows for a more localized approach, considering regional specificities and formulating specific strategies to meet the educational needs of each locality. Based on these robust analytical data, the government can make informed decisions and allocate resources more efficiently, ensuring the implementation of more effective educational policies and thereby contributing to the advancement and improvement of the country's educational system.

The present study aims to construct educational and population datasets for Brazil, segmented by municipality and age through a detailed process of transformation and structured modeling to facilitate data consumption. In public educational data sources, a collection is observed that requires translation treatment based on a data dictionary, as well as an enrichment of this collection to contextualize the data and enable future processes of aggregation, filtering, selection, and other operations needed to extract value from this material.

In light of the above, the main source of data used in this study was established with the Educational Census coordinated by the Brazilian National Institute for Educational Studies and Research (INEP), which is regulated by normative instruments that establish the obligation, deadlines, responsibilities, and procedures for the entire data collection process. [1]

This article is an extended version of the paper Elaboration of the aggregated dataset of the basic education census Barros *et al*. [2022], published in DSW-2022. As an extended version, new analysis of the use of educational data added to population data was made, in order to demonstrate its impact on public policies, which in the case of this project is the National Education Plan, one of the major Brazilian educational public policies. Additionally, this extended version was also added the process of a work Dataset of population estimates disaggregated by municipality and age 2014-2020 Albuquerque *et al*. [2022], fundamental to the new analysis performed.

The article is organized as follows: Section 2 presents some related works on the use of educational and population data in the construction of public databases and their applications, as well as discusses methods for estimating the

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar

population; Section 3 presents the process through which the educational and population data were developed, from their acquisition to the provision of the databases along with their metadata; Section 4 demonstrates some of the results that can already be achieved with the individual use of these data and then delves into a case study of a national plan; and finally, Section 5 provides the final considerations and future work that may improve and optimize the work already done.

2 Related Works

To create the knowledge base of previous works, it was necessary to search for general public data usage, including educational, population, and health data, and particularly how these data have been applied for monitoring public policies.

In the context of constructing databases from public sources, an important case demonstrated in Vasconcelos *et al*. [2021] involved collecting and standardizing data directly from the TSE (Superior Electoral Court) portal. After acquiring and processing the raw data, new geographical data were added to enhance the usability for database users.

Transitioning to the field of health, Gonçalves *et al*. [2021] collected, processed, and made available a database on vaccination records registered in the $OpenDataSUS$ platform. The process followed data science principles and was developed interactively until achieving the objective of providing the dataset and data dictionary.

To support the section on population data, it was necessary to search for works that focused on population perspectives and how to work with this disaggregation.

Recent works have explored statistical projection methods for demographic data, essential for constructing public policies. Gonzaga and Schmertmann [2016] aimed to develop an age-based mortality estimation model for small geographic areas in Brazil, using data from the 2010 Demographic Census and the Public Health System (*DataSUS*). The mortality variable plays a crucial role in estimating a location's population in a given year, so the authors proposed a sophisticated statistical model for its projection. However, some observations regarding the timeline for obtaining base data for constructing population estimates in the Brazilian scenario need to be highlighted. The demographic census is updated once every ten years, while actual mortality and birth data are updated annually. Therefore, during the period between the last census and the year prior to the analysis, it is not necessary to estimate mortality data for the population estimate calculation, as this data can be obtained as facts.

It can be observed that the updated factual data obtained from government sources strongly influence the methodological process to generate population data disaggregated by small regions and age. This can even simplify the disaggregation process, as there is no need to estimate certain input variables or establish assumptions. For instance, nominal data on "Births" or "Mortality" at the desired level of disaggregation. If this data is obtained in a timely manner, there is no need for the application of an estimation process for base variables to compose the population estimate calculation.

In addition to the process of constructing population estimates, another relevant concept that underpins the data modeling process is that of population disaggregation in relation to population estimates. The study conducted by González *et al*. [2015] highlights that population projection models are essential tools used in various programs and public policies, especially those related to the sustainability of public pension, health, and education systems. The study distinguishes these concepts, emphasizing that the population disaggregation process starts from pre-existing demographic data provided by government statistical institutions, and subsequently implements a methodology to enhance the granularity of this data, considering factors such as ethnicity, social class, gender, age, among others.

Due to the limited availability of approaches on the population topic, the study's foundational research was more extensive, focusing on which approaches could be applied to this study's process. Based on these concepts and considerations of significant variables, the population disaggregation methodology that proved most suitable for the temporal scope and data acquisition scenario was the methodology proposed by de Atividades Especiais TCE-SC [2021], developed by the Special Activities Directorate of the Court of Auditors of the State of Santa Catarina. In this methodology, the absence of data on migration between municipalities was addressed by considering the variation of population estimates for the respective municipalities over a period of time. Thus, the dataset underlying this methodology consists of census data, birth and infant mortality data, and population estimates for municipalities without age granularity.

Regarding the educational topic, since we already have more consolidated data sources, data mining techniques were applied to construct the necessary foundation to address the main indicators targeted in this study.

3 Data Acquisition and Processing

In this stage of the project, the processing types are separated for their respective areas. For the educational data, the main provider is the National Institute for Educational Studies and Research Anísio Teixeira (INEP)[2], while for geographic data distribution, we have the Brazilian Institute of Geography and Statistics (IBGE)[3]. Given these differences, we will demonstrate the process applied to each data theme in two distinct flows.

The detailed and complete process of the complete processing of the population part can be seen in Albuquerque *et al*. [2022].

3.1 Educational Data

The development process of this dataset was based on the Extract, Transform, Load (ETL) process, as defined in Ferreira *et al*. [2010]. This process involves extracting data from its sources, preferably in its original form, applying the necessary transformations to make the data more accessible and direct for use, and loading it into a more definitive storage location. In the scope of this project, a set of files in CSV

[2] https://www.gov.br/inep/pt-br
[3] https://www.ibge.gov.br

format (Comma-separated values) was used as the final storage format.

All processing was designed and executed with the ultimate goal of consolidating a basic education census dataset, with the chosen time frame for this project being from 2010 to 2021. This created a dataset that can be analyzed and integrated with other public databases. To correctly design the data processing flow, readings on the subject were conducted, and, most importantly, an understanding of the data dictionaries published by the publisher. This resulted in the following process, as illustrated in Figure 1

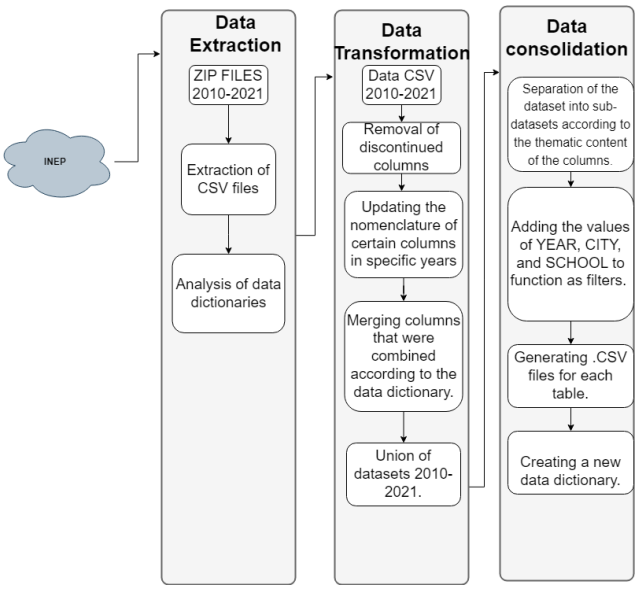


Figure 1. Processing pipeline of the used data.

3.1.1 Data Extraction

The original data for this project was entirely retrieved from the INEP website. These data are provided in a compressed format to facilitate downloading to the local machine. The data is originally provided annually, with a CSV file containing the data and an excel spreadsheet data dictionary, along with other metadata, such as questionnaires applied to institutions and a user manual indicating information for use in specific software.

3.1.2 Data Transformation

After reading the data dictionaries, the necessary changes were identified to unite all the years targeted in the project. It was observed that a large part of the columns had already been standardized through updates made by INEP over the years. However, there were still some columns that would hinder this unification, and some others were discontinued in more recent years.

1. **Removal of discontinued columns**: Discontinued columns from 2019 onwards were removed as the most current data no longer reflects them and, therefore, may not provide up-to-date information for monitoring the educational network structure. Out of the total of 370 available variables, 36 were removed, leaving 334 for further processing.
2. **Update of column names**: During the processing of the remaining columns, certain changes were applied to adapt the nomenclature and variables to the applied time frame. This need was identified through the examination of the data dictionaries provided with the files.
3. **Combining columns**: Some columns had their names changed after 2018, and it was necessary to standardize these columns to the most current names. These columns had their names altered, but not their values. This process followed the standard indicated in the data dictionary. Other columns had to be combined starting from the year 2019, and the strategy of using the OR was applied. Since these columns represent specifications of the same response, if any of these columns have a positive value, the new column will be assigned a value of True. If none of the columns have a positive value, the value False will be assigned to the new column.
4. **Data consolidation**: The final step of this process aimed to merge all the data that was processed annually into a single data table. In contrast to the approach adopted by INEP, a separation of columns was proposed, grouping them by theme, as can be observed in the table 1, to allow for loading only specific groups of columns. The data was separated by themes such as enrollment, teachers, classes (previously separated by INEP in earlier versions of the census), and results.The removed, combined, and remapped columns are available in the metadata, which can be found alongside the database at `https://zenodo.org/record/6666613#.YrioP3bMLnY`.

3.2 Population Data

The data for the chosen population disaggregation methodology was obtained from IBGE and the Ministry of Health. As highlighted in the technical note that formalizes the methodology, the obtained data includes:

1. Obtained from IBGE:
   - Census data segmented by year and age; [4]
   - Annual population estimates of municipalities. (No age segmentation available.) [5]
2. Obtained from the Ministry of Health:
   - Data from the Live Birth Information System; [6]
   - Data from the Mortality Information System. [7]

3.2.1 Data Acquisition and Processing

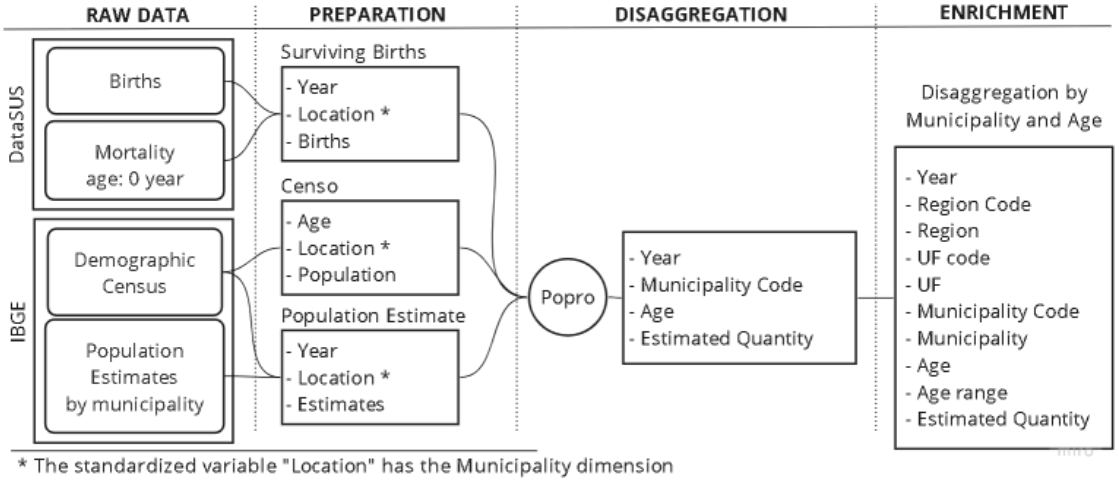
The process illustrated in Figure 2 follows a flow where input data from the previous section are cleaned and normalized, as some of them are provided in spreadsheets with textual observations in cell values and refer to municipalities in a

[4] `https://sidra.ibge.gov.br/tabela/200`
[5] `https://ftp.ibge.gov.br/Estimativas_de_Populacao`
[6] `https://datasus.saude.gov.br/nascidos-vivos-desde-1994`
[7] `https://datasus.saude.gov.br/mortalidade-desde-1996-pela-cid-10`

Table 1. Tables generated by theme.

Table	Description
TB_GEOGRAFICA	Geographical data where the educational institution is located.
TB_TURMA	Information about the classes of a specific educational institution.
TB_DOCENTE	Quantitative data about the teachers belonging to the educational institution.
TB_MATRICULA	Enrollments made at the educational institution.
TB_ESCOLA_ADMINISTRATIVO	Data about the educational institution.
TB_ESCOLA_ESTRUTURA	Data about the infrastructure of the educational institution.
TB_ESCOLA_FUNCIONARIO	Data about non-teaching staff at the educational institution.
TB_ESCOLA_PEDAGOGICO	Data about the pedagogical structure of the educational institution.

Figure 2. Data flow for generating the Population Dataset



textual format instead of their unique code standardized by IBGE.

In general, the first common challenge in processing the input data for all the analyzed reports lies in the presence of explanatory texts in the header and footer of the CSV reports extracted from both *DataSUS* and *IBGE*, with no standardized number of comment lines in these two locations.

Going into detail, the ”Births” and ”Mortality with age of zero year” reports extracted from *DataSUS* in CSV format have only two columns. The first column is labeled ”Município” and contains text that concatenates the municipality code standardized by *IBGE* with the municipality name. Besides this concatenation characteristic requiring treatment, there are a few lines in each report where the municipality code is not valid when queried in the IBGE municipality code table. However, the total population quantity of rows with invalid municipality codes represents an insignificant fraction, around 1 in every 30,000 individuals counted (0.0033%).

In the Census Demographic report provided by *IBGE*, the ages are arranged in different columns, and municipalities are presented in different rows, forming a data matrix with thousands of rows and dozens of columns. It is observed that for some rare combinations of municipalities and age groups, the population quantity is not provided, and a null value is given. Considering the population volume for each age group in the municipalities where these cases occur, it seems unlikely that the cause of these null values is the absence of people of that age in the respective municipality. Instead, the cause is likely related to the data collection or processing process during the report construction, which can be resolved by the responsible party for its development.

Lastly, the Population Estimates report, also provided by *IBGE*, like the previous ones, contains textual information in the header and footer. Additionally, in the quantity column, there are specific markings indicating footnotes that clarify some methodological details leading to the reported numbers. These markings consist of numbers within parentheses, altering the structure of the column, making it non-purely numeric. This, once again, requires transforming the report data into a tabular format with consistent data types in each column.

With the completion of the cleaning process, three normalized auxiliary tables are constructed. Along with the metadata indicating the year to which the census data relates, these tables will be used as input variables in the population disaggregation tool presented in Section 3.2.2.

The auxiliary tables consist of: ”Surviving Births” (Year, Location, Births), ”Real Population from the last census segmented by age and location” (Age, Location, Population), and ”Population Estimate by Municipality” (Year, Location, Estimate).

3.2.2 Population Disaggregation

The population disaggregation process follows two main approaches, depending on the target year for analysis and the birth year of the age group under consideration. In summary, the methodology starts with a population base and ages it

based on the fluctuation of the population estimate for the municipality under analysis over the required period.

This population base can be from two sources, depending on the birth year of the age group under analysis. The population base is census-based when the population set for a specific target year has a birth year equal to or earlier than the last Census execution. The population base is derived from the volume of surviving births for other cases, i.e., when the population set for a specific target year has a birth year subsequent to the last Census execution. The mathematical model and a more detailed discussion can be found in the technical note de Atividades Especiais TCE-SC [2021], which formalizes this methodology.

The disaggregation process was carried out using the Popro tool Albuquerque [2022], a population disaggregation library developed by the authors in the Python programming language. The tool is expandable to different methodologies, and the statistical model of the mentioned methodology was implemented as a plugin for generating the current dataset.

3.2.3 Resulting Dataset

Upon completion of the data extraction, processing, disaggregation, and enrichment steps, the resulting dataset is capable of supporting action plans and guiding the establishment of goals in public policies. Due to the implemented disaggregation methodology having a decreasing level of confidence as the estimated age increases, it is essential to emphasize that the provided database is more suitable for supporting plans involving age groups of children and adolescents, such as primary and secondary education policies.

The resulting *Dataset* has granularity in the variables YEAR, MUNICIPALITY, and AGE. Important data additions were made for future analysis and grouping, including the municipality code, name of the federative unit, and the name of the geographic region, each with their respective codes standardized by *IBGE*.

3.2.4 Best Practices for Dataset Dissemination

This step, being common to both datasets, followed the same procedure, albeit producing slightly different artifacts, aiming to comply with a set of data publication best practices [8].

By achieving the goals outlined in Table 2, we were able to provide a dataset that is traceable both in terms of its source data (BP5) and the version being used and publication date (BP7), along with updated metadata for coherent data utilization (BP1, BP2, BP3).

The data and metadata responsible for applying these best practices are available on the ZENODO platform [9] :

- Educational:`https://zenodo.org/record/6666613#.YrioP3bMLnY`
- Populational: `https://zenodo.org/record/6689160#.YrdBeHbMK5c`.

[8] https://ceweb.br/media/docs/publicacoes/1/fundamentos-publicacao-dados-web.pdf

[9] https://zenodo.org/

The files are made available under the CC-BY-4.0 license, which allows others to download, adapt, and build upon this work, even for commercial purposes, as long as proper attribution is given to this project as the source.

4 Applications of the datasets

When addressing the theme of outcomes regarding the creation of public datasets, we should first consider the progression or regression of certain age groups, access to education, and the quantity of teachers in specific locations. All these data, even in isolation, already provide informative insights and can contribute to the formation of a collaborative society. Another perspective is to cross-reference these data to monitor plans and public policies. These datasets can offer a way to observe them at the municipal level, encompassing all Brazilian municipalities, thus creating a plural and inclusive approach to monitoring.

Below are demonstrations of uses in these two areas, both using the datasets individually and in combination, to monitor the goals of the National Education Plan (PNE), one of the main public policies in the educational field in Brazil.

4.1 Utilization of the datasets for educational monitoring

Individually, the data can already provide us with an overview of the presence of certain factors during the historical series covered in the studies.

In the educational field, there are countless possibilities for applying these data. For example, in Balbinot and Haubert [2017], the application of these census microdata was used to monitor indicators related to special education in Rio de Janeiro. Similarly, in Balbinot and Haubert [2015], a temporal analysis of enrollments in the state of Paraná was conducted. These are some examples of works already carried out based on these educational data.

Individually, we can use data visualization tools like Plotly [10], which, combined with Python [11], allows us to create visualizations about the educational situation of a specific municipality, helping us understand the situation reflected over the time covered by the dataset.

In Figure 3, we can observe the decrease in the number of teachers in the municipality of Moreno, PE. The teachers from the three education levels (Blue representing basic education, red representing elementary education, and green representing high school) show that in that locality, the largest number of available teachers is in basic education. We can also observe if the number of enrollments follows the same pattern.

Continuing to use the dataset, in Figure 4, with the same color and legend scheme (Blue representing enrollments in basic education, red representing enrollments in elementary education, and green representing enrollments in high school), we can see that the majority of enrollments belong to the basic education level. Similarly to the number of teachers

[10] https://plotly.com/

[11] https://www.python.org/

Table 2. Achieved best practices for the datasets

Code	Definition
BP1	PROVIDE METADATA
BP2	PROVIDE DESCRIPTIVE METADATA
BP3	PROVIDE STRUCTURAL METADATA
BP5	PROVIDE DATA PROVENANCE INFORMATION
BP7	PROVIDE VERSION INDICATOR

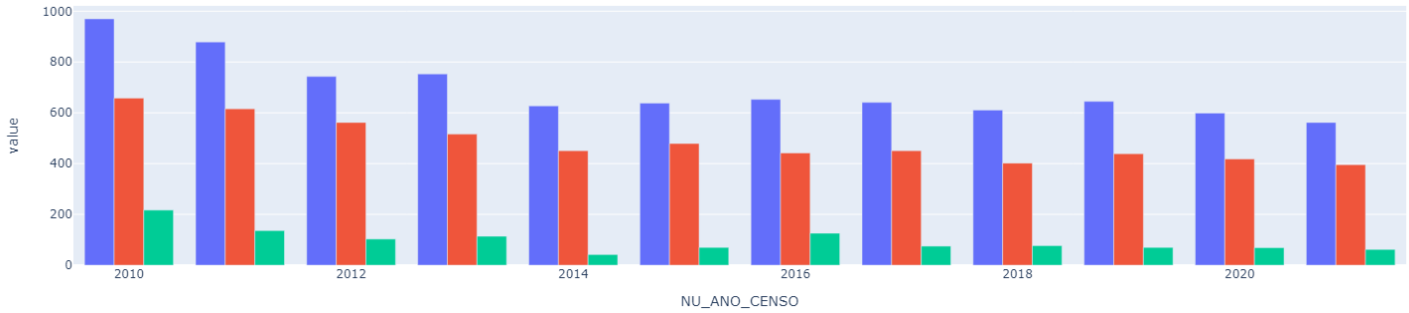


Figure 3. Number of Teachers per year and grade in the Municipality of Moreno-PE

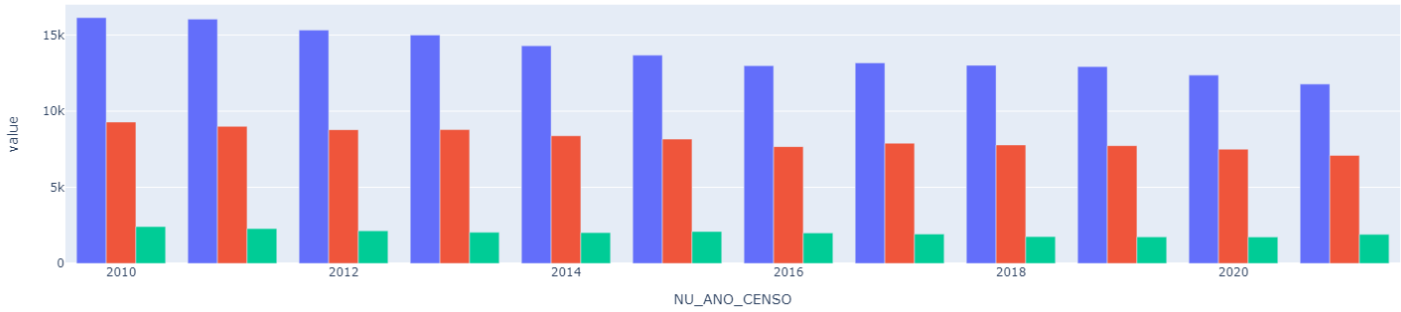


Figure 4. Number of Enrollments per Year and Grade in the Municipality of Moreno-PE

in the education system, the number of enrollments in this educational level has been decreasing, but it still remains higher than the others.

4.1.1 Monitoring age groups over the years

The use of a population dataset segmented by age in small geographical areas aims to support the development of public policy planning and the measurement of metrics in various types of institutions, both public and private.

For instance, in the private educational sector, a basic education institution, using this dataset and its own database of enrolled students, can measure the representation of its students compared to potential students in the municipality where the institution is located. This allows for the calculation of the market share indicator, as discussed in the study Ozkan *et al*. [2022], which is essential to guide strategic decisions, such as focusing on attracting new students, retaining enrolled students, or improving the quality of services offered.

On the other hand, in the public sphere, this dataset can be used, for example, to collect data for analyzing public safety through a methodological approach. It allows the formulation of indicators, such as the crime rate against specific age groups of the population, to assess their behavior over time and identify patterns and trends.

The development of this type of dataset enables the tracking of age group progressions since the IBGE population census is conducted once every decade, creating the need for projections of these other age groups. This understanding is crucial to observe the advancement or decline in the estimated population for each specific age group over the years. We can observe this for the municipality of Recife in Figure 5.

The tables were organized to enable aggregations based on state and region codes. This allows for applying grouping operations with the help of data processing tools or languages. In Figure 6, we can observe the same example applied earlier, this time for the state of Pernambuco.

Figure 5. Population estimate progression by age group in the municipality of Recife-PE.

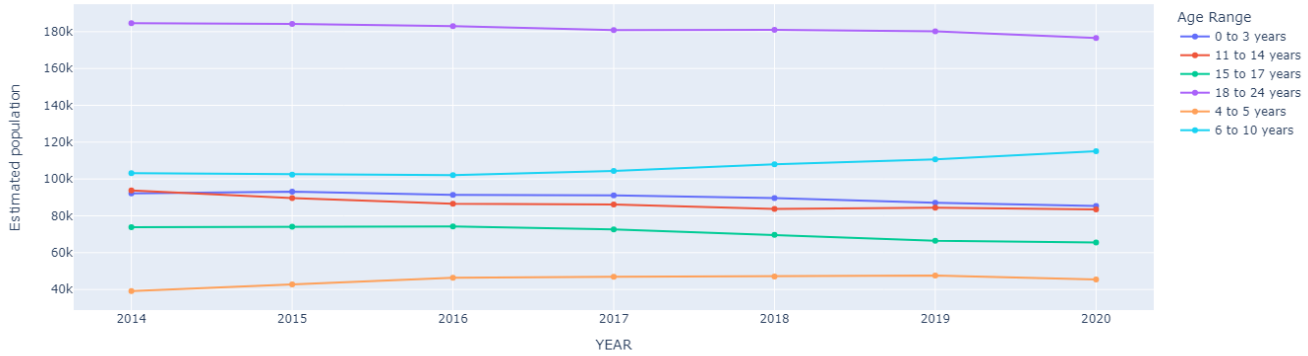
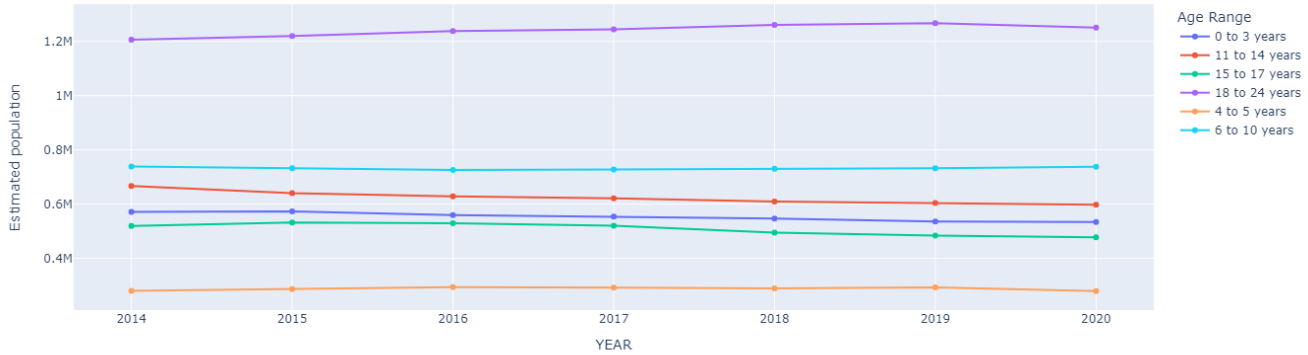


Figure 6. Population estimate progression by age group in the state of Pernambuco.



4.2 Monitoring the PNE

The entire development of this work was based on monitoring one of the most significant public policies focused on education, the PNE. It is a long-term public policy that establishes guidelines and goals for the development of education in Brazil over a ten-year period. It was created to promote quality, inclusive, and equitable education at all levels, from early childhood education to higher education. As such, it is a fundamental tool to guide government and civil society actions in the pursuit of more equitable education capable of driving the country's social and economic development.

The importance of monitoring the National Education Plan lies in the fact that this monitoring is essential to ensure that the established goals are effectively achieved. By periodically assessing the progress of actions, it is possible to identify advances and also potential challenges and obstacles to be overcome. Monitoring allows for adjusting strategies, reallocating resources, and directing efforts to priority areas, ensuring that education reaches the most vulnerable populations in need of investment.

Due to its ten-year duration, its goals and action plans for achieving its indicators need to be understood and continued by all governments operating during this period, regardless of the political party in power at the federal, state, and municipal levels. This requires a firm commitment to cooperation among all government entities throughout the years.

In evaluating all of this, monitoring the PNE is of vital importance, both for its effectiveness as a collaborative government instrument and for its functionality as an overall evaluator of public education and the effectiveness of actions aimed at improving it. When we consider Brazil and its 5,570 municipalities, we can already gauge the difficulty of monitoring all 20 goals with an average of 2 indicators present in the current PNE (2014-2024)[12].

INEP provides biennial reports[13] highlighting results at the federal, state, and some Brazilian capital levels. However, the purpose of creating these datasets is to fully map the quantity of municipalities where data is available from their foundation (given that some municipalities were emancipated after the 2010 census, leading to a lack of data in certain years).

For these reports developed by INEP, some sampling sources such as the Continuous National Household Sample Survey (PNAD) are used, which does not cover all Brazilian municipalities. Therefore, one of the main objectives of creating and merging these datasets is to reproduce the calculation of indicators 1A, 1B, 2A, 3A, and 3B to represent all

[12] `https://pne.mec.gov.br/`

[13] https://pne.mec.gov.br/publicacoes/itemlist/category/4-monitoramento-e-avaliacao

Brazilian municipalities, initially from 2014 to 2020.

With the generated dataset, it is possible to calculate the following indicators for all Brazilian municipalities:

$$1A = \frac{\text{Enrollments for 4 to 5 yo.}}{\text{Estimated population for 4 to 5 yo.}} \times 100$$

$$1B = \frac{\text{Enrollments for 0 to 3 yo.}}{\text{Estimated population for 0 to 3 yo. in school}} \times 100$$

$$2A = \frac{\text{Enrollments for 6 to 14 yo.}}{\text{Estimated population for 6 to 14 yo.}} \times 100$$

$$3A = \frac{\text{Enrollments for 15 to 17 yo.}}{\text{Estimated population for 15 to 17 yo.}} \times 100$$

$$3B = \frac{\text{Enrollments for 15 to 17 yo. in high school}}{\text{Estimated population for 15 to 17 yo.}} \times 100$$

The use of these data to compute these targets allows for their analysis, verification, and monitoring, whether from the same municipality or neighboring municipalities, bringing even more possibilities for collaborative governance. The data enables replication by government officials, research groups in educational or governmental topics, or even citizens who wish to track the situation of their city over the years.

We highlight three readings based on this dataset combination, aiming to perform the analysis of a municipality, which, in this case, is Recife, the capital of Pernambuco. In Figure 7, we can observe variations in the results of only indicator 1A. In 8, we have a comparison of indicators within the same goal, allowing us to assess their situation. In the same location, to compose a goal, two different age groups are required to measure school enrollment in preschool and daycare. Finally, we have a general comparison of all the indicators mentioned in this text 9, illustrating the complete set that can be generated with this data.

5 Conclusion

The datasets presented throughout this work were developed with the aim of addressing the gap in disaggregated educational and population data for the time period of 2014-2021, a specific range for certain projects like the PNE. Monitoring this and other government projects can increasingly lead to public participation in the creation, implementation, and review of public policies.

In future work, we seek continuous improvement of the datasets. Regarding educational data, our objectives include processing the following years, as well as expanding to other datasets such as the Higher Education Census [14] and the Basic Education Assessment System (SAEB) [15]. These datasets have similar structures to the educational dataset generated in this study. Additionally, our future goals involve developing an Application Programming Interface (API) for data delivery using a relational database, facilitating direct queries to specific municipalities without the need to load and select the database, thereby further enhancing access for municipal secretaries.

Regarding the population perspective, we will seek to refine the population disaggregation methodology to increase the precision of estimates. Expanding the disaggregation period through assumptions for estimating the fertility rate is also possible, but in this study, the disaggregated years are limited to factual data reports obtained from the (*DataSus*), specifically Births and Mortality.Another ongoing stage involves using the 2022 demographic census for validating projections and updating the dataset.

One improvement opportunity for the disaggregation methodology lies in utilizing mortality data for all age groups obtained from *DataSus*, not just infant mortality. Thus, the population aging process will reflect the natural differences in the natural increase of populations with different ages, as each age has a particular mortality rate. This and other research avenues should undergo rigorous statistical and conceptual analysis in light of the literature on demographic projection and disaggregation to ensure that the adjusted methodology provides greater precision and rational consistency, making the most of the relevant data available to government agencies.

Finally, our aim is to make these datasets available in the connected open data format so that SPARQL queries can be used, and the updated datasets can be made continuously and openly available, along with their metadata whenever it is possible to update them.

Authors' Contributions

For the educational part, the author was mainly Abílio Barros, while for the population data, the part was developed by the author Aldéryck Albuquerque. The other authors worked on the text, and we all agree with the version published.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in

- Educational:`https://zenodo.org/record/6666613#.YrioP3bMLnY`
- Populational: `https://zenodo.org/record/6689160#.YrdBeHbMK5c`.

References

Albuquerque, A. (2022). A population projection engine. url-https://pypi.org/project/popro/.

[14] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

[15] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/saeb

Figure 7. Monitoring Indicator 1A Over the Years in the City of Recife

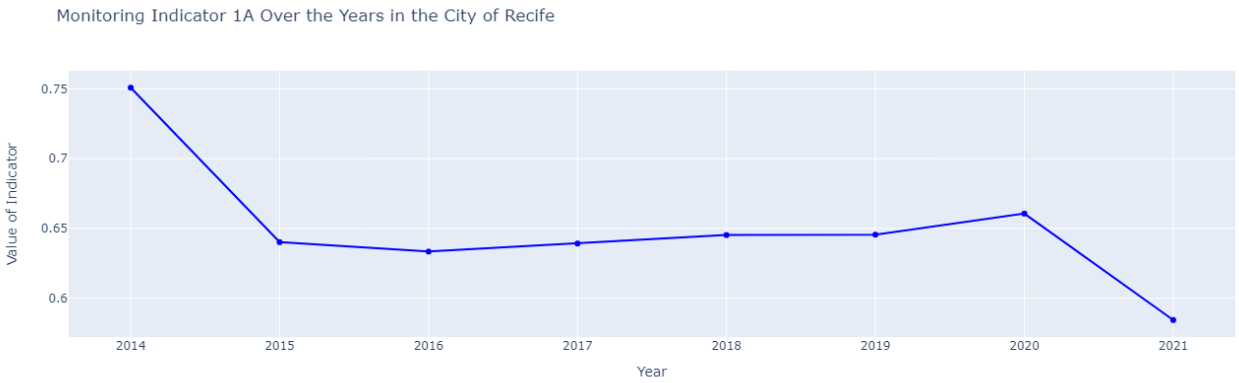


Figure 8. Monitoring Indicators of Goal 1 Over the Years in the City of Recife

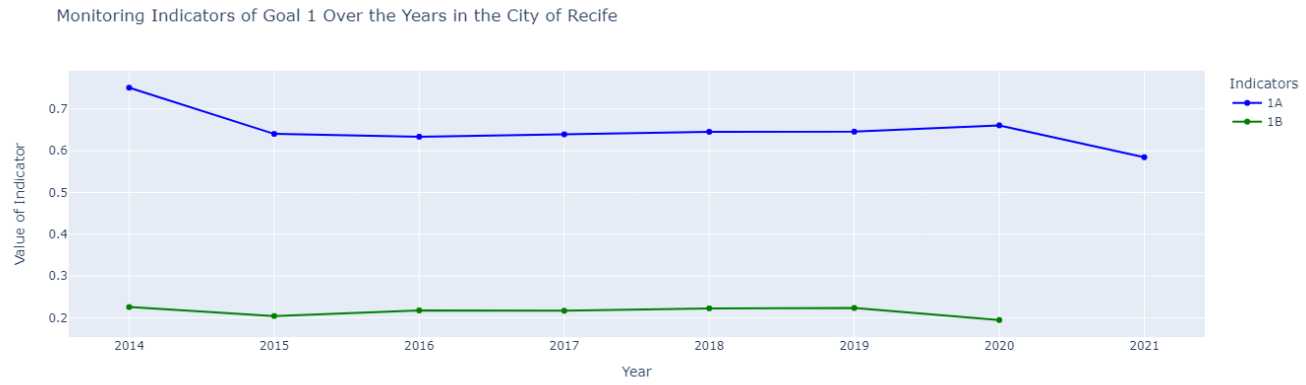
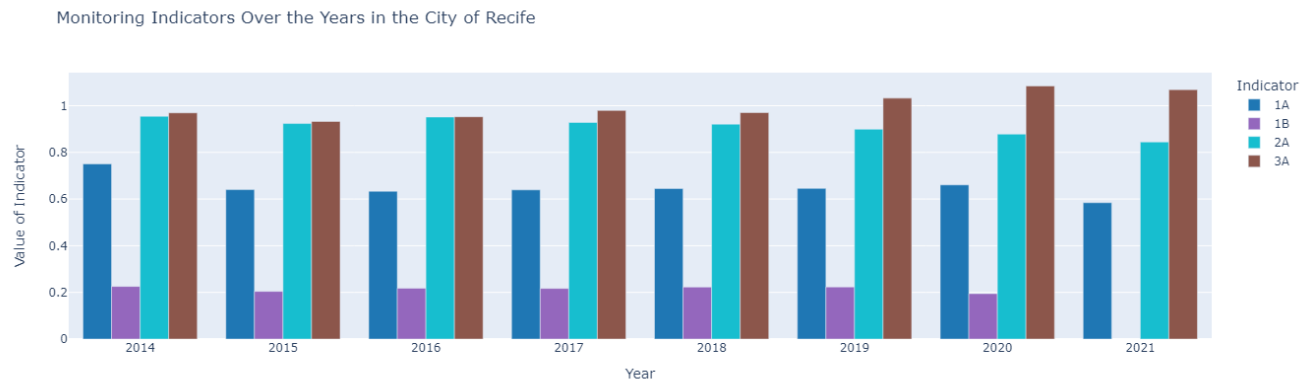


Figure 9. Monitoring Indicators Over the Years in the City of Recife



Albuquerque, A., Barros, A., Alencar, A., Nascimento, A., Bittencourt, I., and Mello, R. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.225525.

Balbinot, A. D. and Haubert, A. (2015). Análise temporal das matrículas em educação especial entre 2005 e 2013 no estado do paraná. *Revista Prâksis*, 2:121–132.

Balbinot, A. D. and Haubert, A. (2017). Análise de matrículas como indicadores da evolução da educação especial no estado do rio de janeiro. *REVISTA ELETRÔNICA PESQUISEDUCA*, 9(19):663–673.

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.

de Atividades Especiais TCE-SC, D. (2021). Metodologia estimação populacional. urlhttps://www.tcesc.tc.br/sites/default/files/2021-06/Metodologia

Ferreira, J., Miranda, M., Abelha, A., and Machado, J. (2010). O processo etl em sistemas data warehouse. In *INForum*, pages 757–765.

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.

Gonzaga, M. R. and Schmertmann, C. P. (2016). Estimativa de taxas de mortalidade por idade e sexo para pequenas áreas com regressão de topals: uma aplicação para o brasil em 2010. *Revista Brasileira de Estudos de População*, 33(3):629–652.

González, M., Fernández Vázquez, E., and Morollón, F. (2015). A methodological note for local demographic projections: A shift-share analysis to disaggregate official aggregated estimations. 16:43–50.

Ozkan, K. S., Khan, H., Deligonul, S., Yeniyurt, S., Gu, Q. C., Cavusgil, E., and Xu, S. (2022). Race for market share gains: How emerging market and advanced economy mnes perform in each other's turf. *Journal of Business Research*, 150:208–222. DOI: https://doi.org/10.1016/j.jbusres.2022.04.040.

Vasconcelos, F. F., Tavares, J. V., Ribeiro, M. U., Coutinho, F. J., and Clarindo, J. P. (2021). Candidata: um dataset para análise das eleições no brasil. In *Anais do III Dataset Showcase Workshop*, pages 160–168. SBC.