Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 239 (2024) 1839–1846

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2023

Unsupervised K-means Analysis of Tuberculosis Data in Brazil: Identifying High Prevalence States and Temporal Trends

Angelo Rossini[a]*, Domingos Alves[b], Vitor Cassão[b], Newton Shydeo Brandão Miyoshi[a]

[a]*Barão de Mauá University Center, Ribeirão Preto/SP, Brazil*
[b]*Ribeirao Preto Medical School, University of Sao Paulo, Ribeirão Preto, Brazil*

Abstract

This paper aims to demonstrate the findings obtained through the analysis and application of an unsupervised K-means algorithm on the SINAN database from 2001 to 2022 in Brazil, with the objective of understanding which states have the highest number of tuberculosis cases and identifying similarities among them that may contribute to a higher case rate relative to the local population. We will begin with a brief historical introduction, followed by an overview of the characteristics related to tuberculosis transmission. Subsequently, we will discuss the results obtained from the year-to-year analysis of the collected data.

Peer-review under responsibility of the scientific committee of the CENTERIS / ProjMAN / HCist 2023

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: angelorossini96@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the CENTERIS / ProjMAN / HCist 2023
10.1016/j.procs.2024.06.365

Keywords: K-means Clustering; Time Series Analysis; Unsupervised Analysis; Tuberculosis;

1. **Introduction**

1.1. **Contextualization**

Tuberculosis remains one of the deadliest infectious diseases in the world. Every day, over 4,000 people die from tuberculosis globally, and approximately 30,000 become ill with this preventable and curable disease. In the Americas, more than 70 people die and around 800 fall ill with tuberculosis every day. It is estimated that in 2020, there were 18.3 thousand children with TB in the Americas, half of them under the age of five. Globally, in the same year, 7.0 million bacteriologically confirmed pulmonary tuberculosis cases had 3.9 million (55%) contacts evaluated for both infection and disease, as stated in [1]. The years 2020 and 2021 were also impacted by COVID-19, which affected the delivery and access to essential tuberculosis (TB) services.

1.2. Objectives

The objective of this work is to identify temporal evolution patterns of tuberculosis cases through the analysis and development of a clustering algorithm using time series based on data provided by SINAN for tuberculosis cases in Brazil. The analysis was conducted at the state level, although the technique can be applied at different regional scales. The model can serve as an aid in the decision-making process for public health managers, allowing them to see how campaigns and initiatives impact not only over time but also by Federal Unit.

2. Materials and Methods

2.1 Knowledge Discovery in DatabaseResearch and Analysis Workflow

This study was developed based on the Knowledge Discovery in Databases (KDD) workflow applied to the analysis of tuberculosis data in Brazil. Following this methodology, the steps were divided in the following order: (i) defining the problem and searching for reliable data sources related to tuberculosis in Brazil; (ii) treating the collected data, such as data cleaning, normalization, handling missing values, and data integration; (iii) transforming the data into a suitable format for machine learning model input; (iv) interpreting the results obtained from machine learning through cluster analysis based on time series, allowing for future advancements to potentially project the model as a decision-making tool regarding the possible paths of tuberculosis evolution regionally.

2.2 Dataset

All the data used in this work were obtained from DATASUS through the PySUS library, which allows for more efficient data collection in CSV format (Comma-Separated Values). The collected data in question

originate from SINAN (Sistema de Informação de Agravos de Notificação), which aims to collect, transmit, and disseminate data routinely generated by the Epidemiological Surveillance System. The data of interest for this analysis consists of notifications received by SINAN between the years of 2001 and 2022. These notifications are filled out by healthcare units for each patient when there is suspicion of a reportable health problem of national, state, or municipal interest. Although rare, there is a possibility that the same patient may generate more than one notification.

There are multiple pieces of information available in the collected CSV files representing different types of datasets. As the objective is to analyze the tuberculosis progression among states over months and years, only the information containing data about Brazilian states was chosen, disregarding cities initially. This repository has been receiving daily updates since 2001. The filtered data for analysis includes the following indicators: date, state, total cases per year and month, and total cases per 1,000 inhabitants. In this study, the time series of the number of deaths per 1,000 inhabitants reported over 11 years for each Brazilian state was utilized.

2.3 Unsupervised Analysis

The choice of unsupervised algorithms was made due to their ability to discover patterns and form groups from a dataset without prior information. There are various types of algorithms, each with its specific application, which can perform better or worse depending on the use case. However, in general, they are often applied for pattern recognition [8], making them a useful tool in this context. Additionally, these algorithms allow for the dissemination of information through web visualization, such as dynamic maps or APIs.

Clustering algorithms refer to a broad range of techniques used to find subgroups or clusters within a dataset. Each identified cluster is compound by observations (or instances) that are highly similar based on some predefined metric (such as euclidean distance)r. The chosen technique in this study is K-means, which involves partitioning N elements into K groups or clusters, where N is the total number of elements being analyzed, and K is the total number of clusters. The value of "K" can be determined by user experience or by using available techniques to calculate a possible value, such as the elbow method or the silhouette method. Validating clustering structures is often the most challenging and frustrating part of cluster analysis, which, as emphasized by [9], can make cluster analysis appear like a black box.

K-means initializes an initial value for the cluster's central point, also known as a medoid, often by randomly selecting a value within the data range. The algorithm assigns each data point to the nearest medoid based on a distance metric, typically the Euclidean distance, and then recalculates the medoid considering the new associated data, until its convergence(where there’s no significant changes between the elements inside the clusters) or the algorithm iterations terminate.

2.4 Technologies

The entire code developed was created in the Python programming language. For graphical visualizations, the Matplotlib library was utilized. Scikit-learn was used for basic machine learning algorithms, such as K-means. Pandas and NumPy were employed for data cleaning and transformations.

2.5 Data preprocessing

The time series dataset was created using the records of cases per 1,000 inhabitants over an 11-year time period, with the data segmented by the total number of cases per month for each year, starting from the first SINAN record in January 2001.

2.6 Time Series Clustering

The clustering step had the value of K chosen as a way to map the disease spread based on the dataset grouped into 3 control groups (k = 3), aiming to cluster the response to the pandemic progression in Brazil into 3 relative levels: bad, moderate, and good.

3. Results

In Fig. 1 shows the number of notifications per year per thousand inhabitants (y-axis) and the respective years (x-axis), for each identified cluster, the results of the time series clustering with k=3 are presented, comparing the development outcomes of each state with the average of its respective cluster. This value identified Acre, Amazonas, Roraima, Pernambuco, and Rio de Janeiro as states with the poorest case development.

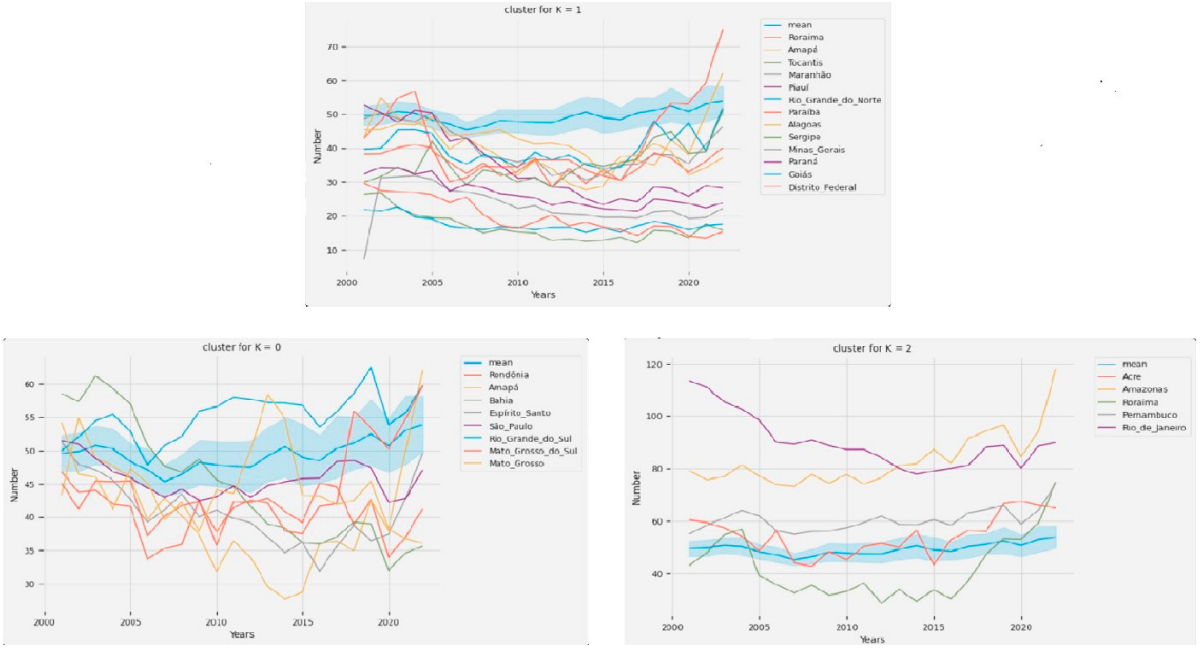

Figure 1. Clustering results with k = 3.

Upon analyzing the graph, we can observe that the average for k = 1 and k = 0 are similar, and most states had their developments below the average in k=1, approaching or surpassing it only at the beginning and end. On the other hand, states in k = 0 exhibited a slightly broader distribution, with three states surpassing the average, and the majority remaining close but below it, making them the best and moderate cases, respectively. Additionally, in k = 2, which grouped the worst cases, Amazonas and Rio de Janeiro stood considerably above the average, classifying them as the worst cases in Brazil.

4. Discussion

In this study, we aimed to identify how Brazilian states have been dealing with tuberculosis over the years based on clustering analysis using the number of cases per 1,000 inhabitants. We chose this metric as the most reliable among other available measures, which may suffer from a higher degree of underreporting. Using a metric that considers the population enables comparisons between different states. Thus, the use of the number of reported cases per 1,000 inhabitants was a judicious choice. If absolute notification reports had been used, the states of São Paulo and Rio de Janeiro would have formed an isolated cluster due to having the highest absolute numbers of cases. In the analysis conducted with k=3, we can observe that the main differences between the clusters lie in the total number of notifications per 1,000 inhabitants.

The states that presented the highest number of cases were Acre, Amazonas, Roraima, Pernambuco, and Rio de Janeiro. We also noticed a higher density between 20 and 60 with a slight decrease as both sides approached 40, suggesting that 30 to 50 cases per 1,000 inhabitants is a common average to be found.

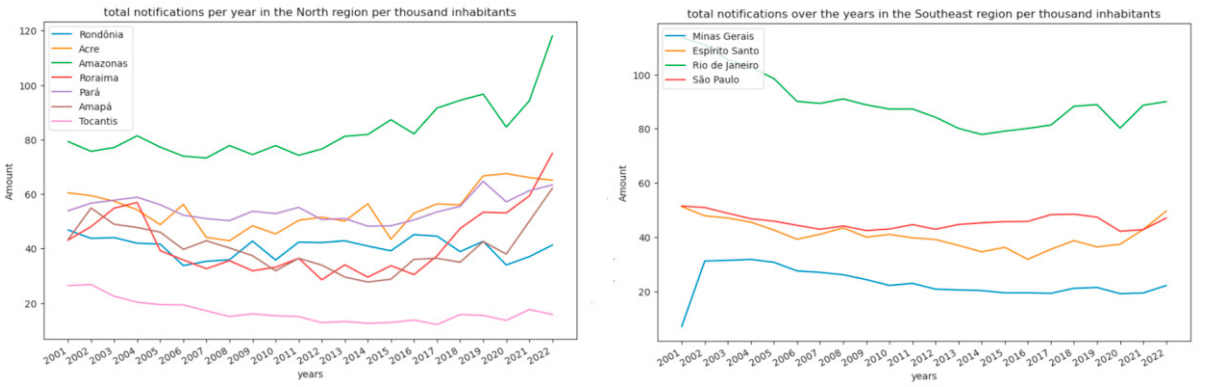

Figure 2.: Distribution of the number of tuberculosis notifications per thousand inhabitants in the North (left) and Southeast (right) regions

In Figure 2, we have the distribution of the number of notifications from the North and Southeast regions. This comparison was made because the technical note from the IEPS (Institute for Health Policy Studies) in [12] demonstrated that between the years 2010 and 2020, the states in the North and Northeast regions had the worst health indicators in 14 indicators, while the South and Southeast regions mostly had the best resource and mortality indicators. This may explain why the Northern region shows a higher number of reported cases, and

among the Federal Units, Amazonas demonstrated having a higher number than states with a higher population density such as Rio de Janeiro and São Paulo.

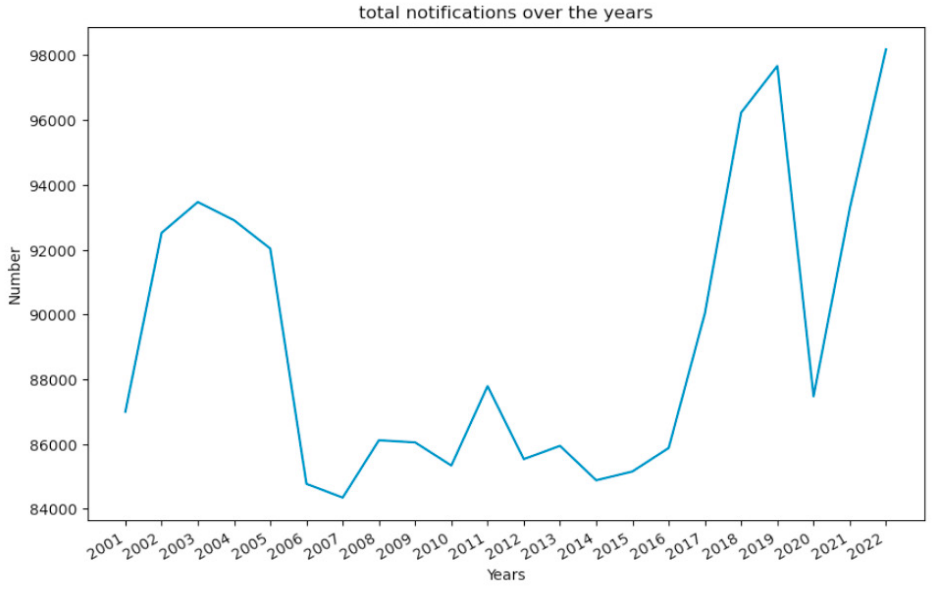

Figure 3. Total number of TB notifications in Brazil over the years in SINAN

In Figure 3 above, it is possible to observe the distribution of the total number of tuberculosis cases over the years, where we can see a clear improvement in the number of cases from 2005 to 2016, which can be attributed to awareness campaigns, the implementation of a rapid disease diagnosis network in 2014, and the decentralization of treatment to Primary Care during this time. The slight increase in notifications from 2018 to 2019 may be due to a reduction in tuberculosis vaccine coverage. Until 2018, the vaccination rate remained above 95%, while in 2019, coverage did not exceed 88%. The increase from 2020 onwards could potentially be due to misdiagnosis or consequences of the pandemic.

Another important factor to consider is the data quality of SINAN notifications data. As a national wide system, SINAN-TB can be used as a bureaucratically-driven system and some limitations are highlighted by Rocha et. al. [13] such as: incomplete and inconsistent reporting of key variables, such as demographic details (schooling, race/skin color), associated conditions (AIDS, alcoholism, diabetes), and treatment follow-up information. Additionally, essential data related to special populations and beneficiaries of government programs are often inadequately captured. The presence of unclosed or inaccurately closed records, coupled with the lack of timely updates for ongoing laboratory tests, further hinders accurate representation and evaluation of the interventions conducted. SINAN-TB does not have a unique person identifier and its hard to promote seamless integration with other systems. An effort is needed to change its architecture and embrace modern technologies to streamline data transfer and analysis in the context of TB control in Brazil.

5. Conclusion

Time series clustering has proven to be an important tool for analyzing data and finding behavioral similarities in the progression of tuberculosis among Brazilian states. Initially, the number of notifications per 1,000 inhabitants was chosen as the parameter for analysis. The number of clusters was set to 3, chosen as an attempt to categorize the disease progression into 3 control groups.

Despite these findings, considering a country of continental dimensions like Brazil, further in-depth analysis is still needed. It is necessary to identify the reasons behind the differences in disease progression among Brazilian states. As a future work, additional parameters will be included in the clustering process to consider other information about the states, as done in [6]. Some states have specific characteristics that may not be fully captured by a parameter considering only the total resident population. States like MG and AM have vast green areas, requiring more careful analysis. Another improvement that can be made in future work is to automate the process of finding the optimal number of clusters through a hierarchical clustering approach [7]. It would be interesting to compare the hierarchy among states and extract meaningful characteristics from it. Another point of improvement is the utilization of metrics that consider the entire time series during clustering, highlighting variations in disease spread velocity and other factors, such as the Dynamic Time Warping (DTW) clustering algorithm. The integration with other datasets from states holds immense potential for enhancing the depth and efficacy of TB analysis. By merging clinical, demographic, socioeconomic, and geographic data, we can achieve a holistic view of TB incidence, prevalence, and associated risk factors. This comprehensive approach aids in identifying intricate patterns and potential determinants of TB, enabling tailored interventions and policies that address the unique challenges faced by different regions.

The obtained results can assist in identifying patterns and evaluating the success of actions and strategies adopted by Brazilian states in combating tuberculosis. This analysis can provide insights to guide future actions and determine the best responses for similar cases.

6. Acknowledgements

We would like to thank Barão de Mauá University Center for the financial support.

7. References

[1] Global Tuberculosis Report 2021.Available in: <https://www.who.int/publications/digital/global-tuberculosis-report-2021>.

[2] BARBOSA, Isabelle Ribeiro et al . Análise da distribuição espacial da tuberculose na região Nordeste do Brasil, 2005-2010. Epidemiol. Serv. Saúde, Brasília , v. 22, n. 4, p. 687-695, dez. 2013 . Disponível em <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742013000400015&lng=pt&nrm=iso>. acessos em 27 mar. 2023. http://dx.doi.org/10.5123/S1679-49742013000400015.

[3] BIERRENBACH, A. L. et al. Incidência de tuberculose e taxa de cura, Brasil, 2000 a 2004. Revista de Saúde Pública, v. 41, p. 24–33, 1 set. 2007.

[4] CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás (UFC), v. 1, n. 1, p. 1-29, 2009..

[5] BATISTA, Gustavo Enrique de Almeida Prado Alves. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2003. doi:10.11606/T.55.2003.tde-06102003-160219. Acesso em: 2023-03-27

[6] CASSÃO, Victor et al. Unsupervised analysis of COVID-19 pandemic evolution in brazilian states. Procedia computer science, v. 196, p. 655-662, 2022.

[7] SPOLAÔR, Newton et al. Um estudo da aplicação de clustering de séries temporais em dados médicos. In: III Congresso da Academia Trinacional de Ciências. 2008. p. 1-10.

[8] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

[9] Anil K. Jain and Richard C. Dubes. 1988. Algorithms for clustering data. Prentice-Hall, Inc., USA.

[10] Ministério da Saúde lança campanha de combate à tuberculose e reforça ações para eliminação da doença no Brasil. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2023/marco/ministerio-da-saude-lanca-campanha-de-combate-a-tuberculose-e-reforca-acoes-para-eliminacao-da-doenca-no-brasil#:~:text=Uma%20das%20principais%20formas%20de>. Acesso em: 14 set. 2023.

[11] Incidência da tuberculose cai 20,2% no Brasil em uma década. Disponível em: <https://www.unasus.gov.br/noticia/incidencia-da-tuberculose-cai-202-no-brasil-em-uma-decada>. Acesso em: 14 set. 2023.

[12] RACHE, Beatriz et al. A Saúde dos Estados em Perspectiva Comparada: Uma Análise dos Indicadores Estaduais do Portal IEPS Data. Instituto de Estudos para Políticas de Saúde, 2022.

[13] ROCHA, M. S. et al. Sistema de Informação de Agravos de Notificação (Sinan): principais características da notificação e da análise de dados relacionada à tuberculose. Epidemiologia e Serviços de Saúde, v. 29, n. 1, mar. 2020.