Submission date: 17/12/2021
Camera ready submission date: 23/08/2022
1st round notification: 08/08/2022
Available online: 06/10/2022
Publication date: 30/12/2022

Section: extended version
Extended versions of selected papers from the X Brazilian Workshop on Social Network Analysis and Mining

Brazilian Reading Preferences in Goodreads: Cross-state and Cross-region Analyses

Mariana O. Silva[1], Clarisse Scofield[1], Luiza de Melo-Gomes[1], Juliana E. Botelho[1], Gabriel P. Oliveira[1], Danilo B. Seufitelli[1], Mirella M. Moro[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mariana.santos,clarissescofield,luizademelo}@dcc.ufmg.br
{juliana.botelho,gabrielpoliveira,daniloboechat,mirella}@dcc.ufmg.br

Abstract. *As a multicultural and ethnically diverse nation, Brazil has singular cultural identities in accents, gastronomy and traditions, also reflected in its literature. Here, we model a multipartite network to perform cross-state comparison analyses based on the cosine distance for Brazilian reading preferences. We also explore the impact of the relationships between geographic, socioeconomic, and demographic factors and both shared books and literary genres across Brazilian states. Finally, we extract the backbone of networks to identify cultural clusters in Brazil and each of its macro-regions. Such cross-state analyses highlight the country's rich cultural diversity, where each region shows its own identity. Our findings open opportunities to the book industry by enhancing current knowledge on social indicators related to reading preferences.*

Keywords. *Books; Goodreads; Reading Profiles; Cultural Identity; Brazilian Culture; Multipartite Networks; Social Network Analysis.*

1. Introduction

Brazil is a country of continental dimensions. It is also known as one of the most multicultural and ethnically diverse nations due to its immigration movements from various parts of the world over its five-century history. Such a diverse background highlights its cultural wealth: there is not one single type of art but rather a mosaic of different artistic sources that together, as a whole, form the Brazilian culture. Its inherent richness goes beyond, as each Brazilian state has its own micro-vocabulary, accent, fauna, flora, gastronomy, fashion, music, and literature preferences.

Especially in literature, reading books is a cultural behavior shaped by social, economic, and local backgrounds. People choose what to read according to their ideas and intellectual realizations, plus their surrounding influence from a social perspective. From an economic perspective, cultural access and consumption are directly affected by socioeconomic status. Then, from a location perspective, each state or macro-region within a country has characteristics that may interfere with reading preferences.

As a practical example, Figure 1 shows cross-state information regarding Brazil's top three most-read genres and books (dataset in Section 3). *Romance*, *fantasy* and *classic*

Cite as: Silva, M. O., Scofield, C., Melo-Gomes, L., Botelho, J. E., Oliveira, G. P., Seufitelli, D. B., Moro, M. M., (2022). Brazilian Reading Preferences in Goodreads: Cross-state and Cross-region Analyses. iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 15(1), 25:1-25:20. DOI: 10.5753/isys.2022.2411

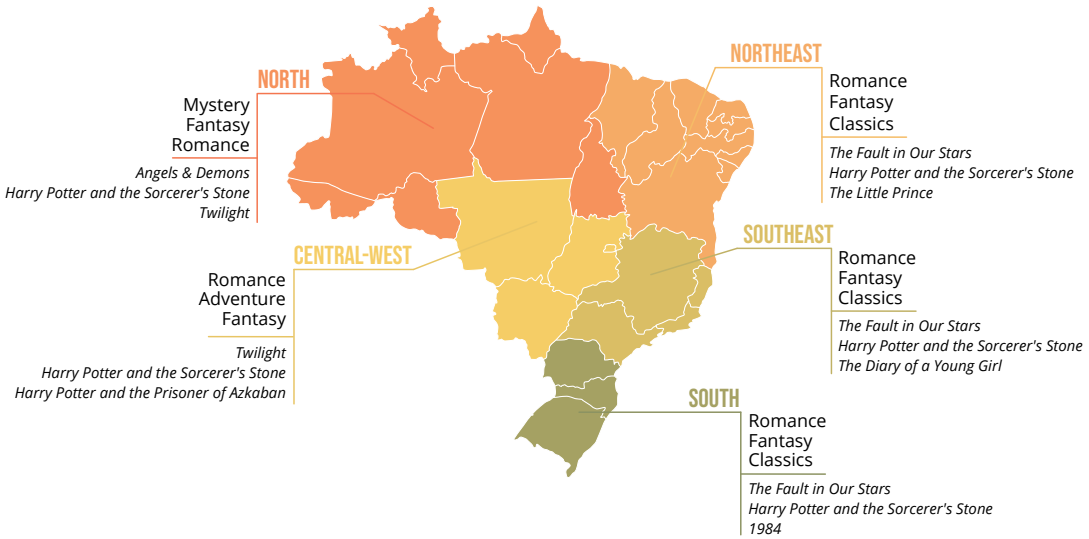

Figura 1. Brazilian Literary preferences map with the top-3 most read genres and their representative books separated by macro-region [Silva et al. 2021b].

are the preferred genres in the country for macro-regions Northeast, Southeast and South; whereas *mystery* and *adventure* are preferred for macro-regions North and Central-West. For the two main genres, the most read books are the same. However, the South elects "1984" as the most read of the *classic* genre.

Studying if and how cultural and socioeconomic factors influence reading preferences is a valuable tool for book-selling and understanding current cultural trends. Moreover, such insights may improve recommendation systems of books, sales projections and customer relationship management. With so many perspectives, the book market is bound to prosper. In this context, we analyze similarities and differences among Brazilian 26 states and Federal District based on reading preferences by using data from Goodreads.[1] We also use a bipartite configuration and a community detection model to explore cross-state relationships based on their book and genre preferences.

This article extends a paper published in the 10^{th} Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2021) [Silva et al. 2021b]. Such prior work finds relationships between cultural distances and geographical, demographic, and socioeconomic indicators. The results also highlight the cultural richness in Brazil by revealing that each state and region has its dynamics regarding what people choose to read.

As a new material, we complement our analyses by filtering out books popular in all Brazilian macro-regions, as such books hide the specific reading profiles in each state and region, i.e., the main goal of this research. Therefore, we run our evaluations in the filtered dataset and compare the results with our prior findings based on the complete, unfiltered dataset. Overall, the new evaluations enhance our findings by clearly showing a solid regional component within reading preferences in Brazil. This newly acquired knowledge has potential use in many applications, as it benefits the scientific community

[1]Goodreads: `https://www.goodreads.com/`

and the book industry. Our contributions are summarized as follows.

1. We build a dataset for genre and book preferences in Brazilian states by using data from the two largest Brazilian reading groups in Goodreads, one of the main online platforms for book readers. For all users in such groups and their friends, we collect their location and the list of books they have read. (Section 3);
2. We model a multipartite network for mapping users' reading preferences regarding books and their literary genres to their location. The proposed model contains four types of nodes (i.e., states, users, books, and genres), which provide interesting perspectives to our evaluations. We also propose two metrics to assess the cross-state cultural identity (Section 4); and
3. We analyze the networks from three perspectives regarding reading preferences (Section 5): differences between Brazilian states and macro-regions, demographic and socioeconomic relationships, and state identities. We also compare all results to our prior analyses.

2. Related Work

The Internet and other technological advances have shortened the distance between people and their contact with foreign information. In the last two decades, Online Social Networks (OSN) have become the main space in which people from all over the world interact. Following such evolution, the content provided by OSN is deeply studied by Social and Computer scientists within distinct tasks, for example, user recognition [Igawa et al. 2015], and topic detection [Barbon Jr. et al. 2017]. Research on this diverse environment also focuses on how cultural differences affect people interacting with new technologies, including social media per se [Nascimento et al. 2018], online music [Liu et al. 2018, Oliveira et al. 2020] and e-learning [Zhao et al. 2021].

Now, processing data from OSN is not free of challenges. The first one may be choosing what resource representation to process from such networks: text (e.g., Twitter posts), image (e.g., Instagram pictures), video (e.g., TikTok short videos), or any combination, as all of them accept comments, replies, likes and so on. The most explored media is probably text, which is also the focus of this article. Overall, text-based methods are mainly language-dependent, and the existing algorithms are primarily developed and tested in English written texts (e.g., [Belinkov and Glass 2019, Otter et al. 2021]).

Nonetheless, researchers who tackle language-oriented tasks (e.g., Natural Language Processing) have also started to explore, adapt and even propose multilingual methods or offer multilanguage support, e.g., [Cagliero and Quatra 2021, Guarasci et al. 2022, Krótkiewicz et al. 2016, Pessutto et al. 2020]. Regarding the Portuguese language, Morais et al. [2020] classify a set of current news from Brazilian news portals into fake, satirical, objective and legitimate news. Other studies focus on sentiment analysis techniques, considering both the methods themselves [Oliveira and Merschmann 2021] and their applications in real life, including the stock market [Carosia et al. 2021].

Language is also present in a country's culture through music lyrics, literature, scientific production, and online interaction, for example. Despite its solid artistic identity, Brazil has several cultural and behavioral differences between its states and regions,

primarily due to its continental dimensions, immigrant movements, and extensive borders with other South American countries. Digiampietri et al. assessed this regional factor by analyzing the relationships between Brazilian scientists using a cross-state social network [Digiampietri et al. 2014]. In addition, Borges de Souza et al. identified a significant correlation between cultural differences and how Brazilians from the five geographical regions interact with Web interfaces [Borges de Souza et al. 2015].

Cultural identity also affects reading preferences because what people choose to read is usually connected to their language, background, and personal beliefs. Indeed, there are several works on reading habits in specific countries and contexts, including Australian adults in prison [Garner 2020] and young people in Croatia [Müller 2021]. In addition, Ma [2021] analyzes reading attitudes among Chinese students by comparing regions with different economic development. All such findings highlight the importance of studying regional contexts individually, as they reveal specific reading profiles that may differ considerably from each other.

As an online data source, Goodreads is one of the leading social networks for book readers. Current research on such a platform includes using book reading behavior to predict best-sellers [Maity et al. 2017], measuring book impact by user reviews [Wang et al. 2019], and using such reviews to extract meaningful characters and their relationships [Shahsavari et al. 2020]. Data extracted from Goodreads is also used to understand cultural differences between countries. For example, Sabri et al. use a bipartite network model to explore cross-country relationships, grouping them in communities based on reading preferences [Sabri et al. 2020]. The results indicate that such preferences are highly associated with geographical distance, language, and individualism factors.

Within such a broad context, this article represents a relevant step towards understanding the relationship between cultural differences and reading habits. Brazil's continental nature is perfect for different cultural practices across its 26 states and Federal District. Hence, we start from an intuition that such cultural diversity may also influence Brazilians' reading habits. To the best of our knowledge, we are the first to study book preferences and cultural identity in Brazil through a network-based model. Overall, our cross-state and cross-region analyses may improve the existing knowledge of Brazilian reading preferences, benefiting both the academy and the book industry.

3. Brazilian Readers' Data at Goodreads

In this exploratory study, we use BraCID, an enhanced dataset comprising cultural, geographical, and socioeconomic information [Silva et al. 2021b]. For completeness, we summarize its main features and building process next in Section 3.1. Then, we present an initial exploratory data analysis in Section 3.2. Finally, we propose a filtering mechanism to make the differences among states and regions more prominent in Section 3.3.

3.1. Summary of Data Building Process

We chose the Goodreads website as our primary data source due to the sheer volume of data available, as well as its organized and easy-access API. We collect Brazilian readers' data through the *goodreads*[2] library. Goodreads is a social book-reading platform for book

[2]goodreads package: `https://github.com/sefakilic/goodreads/`

lovers to rate and review books, see what their friends and favorite authors are reading, participate in groups and discussion boards, and get/share suggestions for future reading choices. A distinct feature is the *bookshelf*, a list where one can add or remove books to facilitate reading, similar to a real-life bookshelf where one keeps books.

Specifically, we collect members of two of the largest Brazilian reading groups in February 2021: the "*Clube de Leitores em Português*" – "Club of Readers in Portuguese" (4,229 members) and "*Goodreads Brasil*" (3,222 members). We also collect data from their friends for all members of both groups. We extract user-based information from the Goodreads API, including age, gender, number of friends, the number of groups they are a member of, the number of reviews, and location. Then, we filter only those containing Brazil as the location information from the final users' set. Finally, we gather users' bookshelves to assess their reading preferences with the same API.

Regarding information on each book, its genre is determined by crowdsourcing users' bookshelves. For example, if multiple users shelve a book as *Science*, then that genre is assigned to the book. Such an approach may include fuzzy and/or noisy tags in the genre list for a particular book. Hence, we apply a data cleansing process to consider only meaningful tags for each book. To do so, we filter the book's tag list to contain only genres related to *Fiction* and *Nonfiction* (better explained later in Figure 3). Next, we manually detect and remove inaccurate records from the user's dataset's location field to keep only Brazilian readers.

To distinguish and investigate the Brazilian reading identity, we consider a medley of demographic and socioeconomic data from the Brazilian Institute of Geography and Statistics (IBGE):[3] territorial area, population estimate, demographic density, Human Development Index (HDI), Gross Domestic Product (GDP), and monthly household income per capita. All indicators refer to 2020, except the HDI, which refers to 2017.

The data collection period occurred from February 23 to March 04, 2021. The final dataset is publicly available and comprises 38,231 Brazilian Goodreads users, containing 75,093 distinct books belonging to 80 literary genres and six IBGE indicators regarding the 27 federative units of Brazil [Silva et al. 2021a, Silva et al. 2021b].

3.2. Exploratory Data Analysis

After collecting the Goodreads data, the next step is to characterize it from different perspectives. Here, we are interested in the reader's distribution by state and the most consumed literary genres. An exploratory data analysis (EDA) is the best approach, allowing us to understand Goodreads' social network ecosystem better. In other words, although Goodreads remains a valuable source of information on reading habits, we can determine the best means to manipulate and discover patterns in such data through EDA. Specifically, we start by analyzing the data related to the collected Brazilian readers, and next, we investigate the genres present in the dataset. We work with multivariate visualizations in both analyses for describing data and mapping interactions between different variables.

Figure 2A presents the distribution of the Brazilian population across its five macro-regions and the distribution of Goodreads users across them. For example, the

[3]IBGE: `https://www.ibge.gov.br/en/cities-and-states`

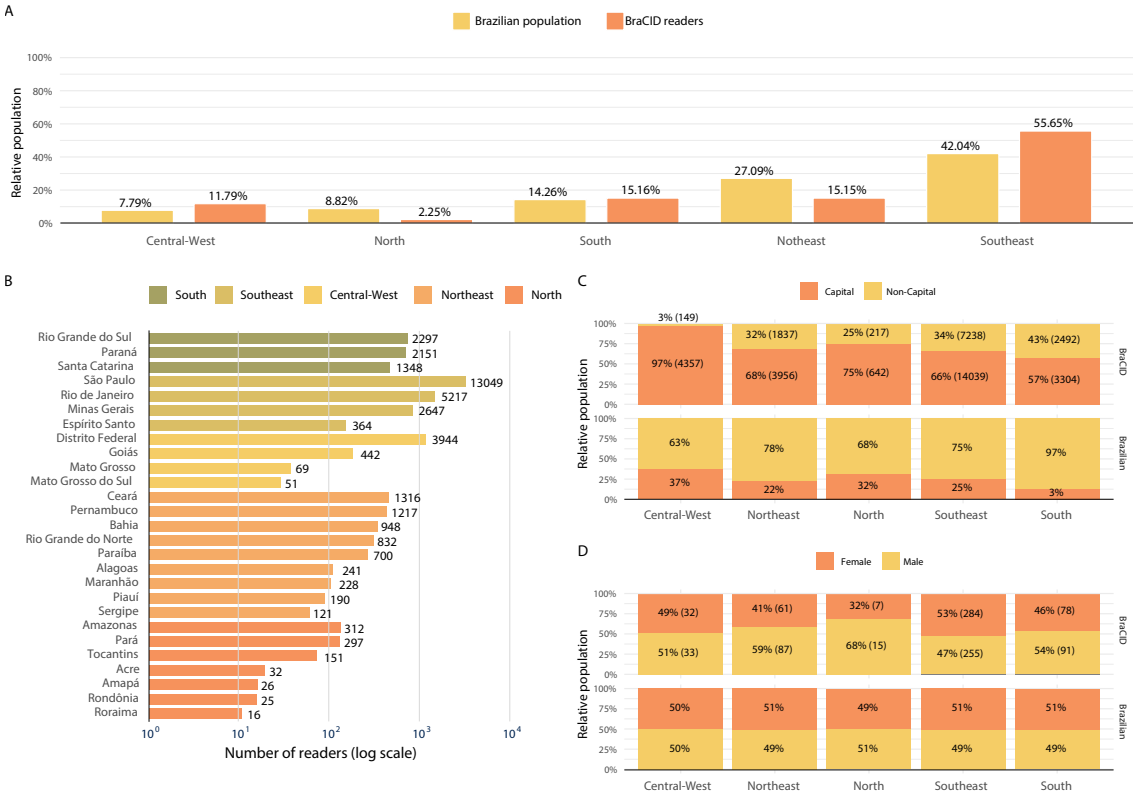

Figura 2. BraCID readers-related statistics grouped by Brazilian macro-region. (A) Relative Brazilian and BraCID populations in each macro-region. (B) Number of readers ranked by the federative units within each macro-region. Percentage of BraCID and Brazilian populations in (C) state capitals and (D) by gender.

Southeast region has 42% of the national population and 55.6% readers within Goodreads. Although the Northeast and South regions have the same percentage of users (15%), there is a much greater membership in the South region, which has half the population compared to the Northeast. The North region has low platform usage, with 8.8% of Brazilians and only 2.2% of readers. The Central West has only 7.7% of Brazilians and represents high adherence, with 11.7% Goodreads users. Overall, Central-West, South, and Southeast have good representations of users within the BraCID dataset, whereas North and Northeast have lower coverage of users. This is a limitation of the dataset that is hard to solve, as any data manipulation would imply losing information or adding synthetic data, which could jeopardize any meaningful analyses even more. Therefore, the dataset remains as it is, and we refer to this graphic as necessary during our analyses.

Figure 2B ranks a grouped bar chart where each group represents a Brazilian macro-region and bars within a group (i.e., macro-region). Figures 2C and 2D show the percentage of people who live in state capitals and by gender, respectively, also comparing the percentage of users in BraCID (top) and Brazilian populations (bottom). Such analyses reveal an evident dominance of the Southeast region, with the state of São Paulo leading the ranking. Such a macro-region compresses approximately 56% of the collected readers, followed by the South and Central-West regions, corresponding to 15% each.

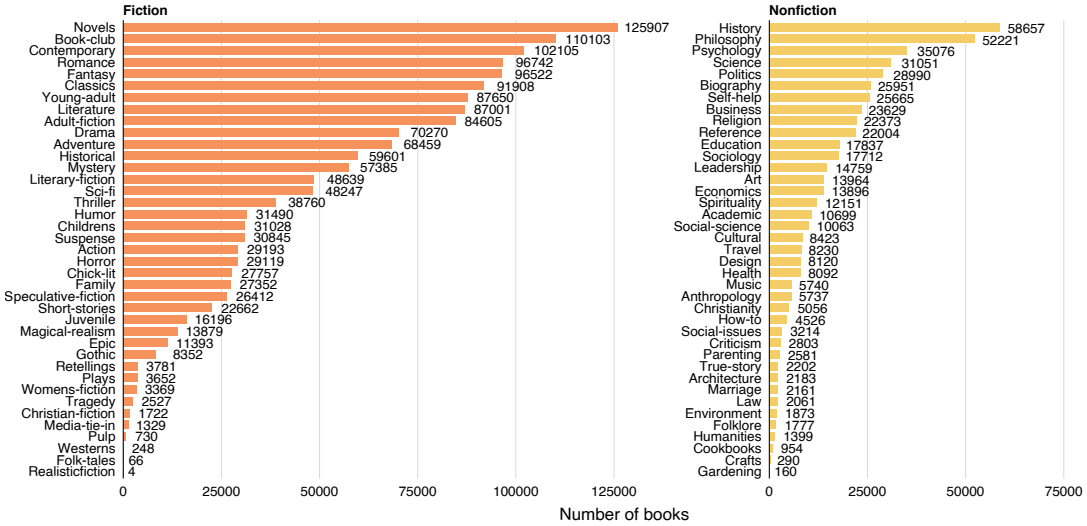

Figura 3. Breakdown of collected books into genre categories of (left) Fiction and (right) Nonfiction [Silva et al. 2021b].

Moreover, most readers reside in state capitals for all macro-regions, mainly in the Central-West, where the capitals are equivalent to 97% of the data. Such a result is inconsistent with Brazil's reality, where more than half its population resides in non-capital cities in each macro-region. The notable concentration of users in the state capitals could probably be due to lower Internet coverage in some areas of the countryside or a lack of interest in the Goodreads platform. Regarding gender information, there is a balanced distribution in general, similar to the country's reality. Finally, the readers' age is relevant and available data; however, most users (89.6%) concealed this information, being unavailable for any meaningful conclusions.

We now focus on the collected book-related information. With more than seven thousand distinct books, our dataset includes 80 different genres. Each genre is classified into *Fiction* or *Nonfiction* categories. Figure 3 breaks down all the collected books by category and genre. Within *Fiction*, most books fall into the *Novels* genre (125,907 books), making it the most popular genre. Indeed, in literature, fiction usually refers to a novel, one of the longest forms of literary prose. In the second place, *Book-club* has 110,103 books, which is more than twice the number of books in Sci-fi, the 15^{th} place with 48,247 books. Previous research shows the growing popularity of the *Fiction* category [Yucesoy et al. 2018]. Likewise, there is a clear predominance of such a category in our dataset, comprising about 76% of the books. To conclude, among nonfiction books, approximately 45% of the collected books belong to *History*, *Philosophy*, *Psychology*, *Science*, *Politics* and *Biography*.

3.3. Filtering out Common Books

All books within the collected Brazilian readers' shelves were considered in the cross-state analyses in the prior work [Silva et al. 2021b]. However, having trendy books across all states may skew the results and hide the actual regional reading profiles. For example, most Brazilian students must read Machado de Assis in high school, bestsellers such as

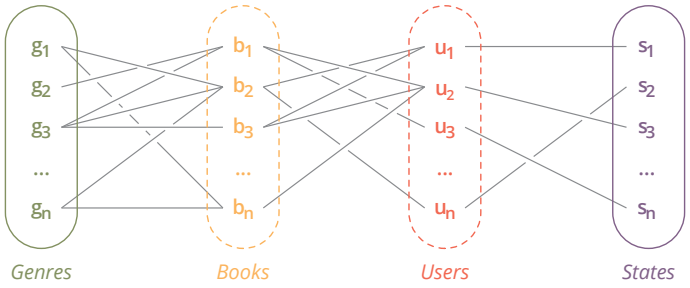

Figura 4. A generic example of a *state-user-book-genre* multipartite network.

Harry Potter's series are read by everyone, releasing movies from books tends to put such books on the most-read lists (e.g., *Angels & Demons*, and *Twilight*), and so on. Overall, considering such best-selling books within the current analyses does not help find specific reading profiles because many people read such stories.

To avoid potential bias, those books that are popular across *all* regions from Bra-CID were removed. Overall, there are 1,275 popular books present on the shelves of all five Brazilian macro-regions, corresponding to about 1.7% of the original book set. In the end, the new dataset has 73,818 books after filtering. Such filtered books will be referred to as *common books* from now on.

4. Methodology

Having collected the data, our methodology follows by modeling two bipartite networks to assess the cross-state reading identity (Section 4.1). We also apply a graph-level statistic test to explore the significance of demographic and socioeconomic indicators on reading preferences (Section 4.2). Next, we further explore identity through reading preferences by modeling two unimodal projections and applying a community detection to identify clusters of Brazillian states (Section 4.3). Finally, we list the research questions to be answered in the analyses results section (Section 4.4).

4.1. Multipartite Network Modeling

To investigate reading preferences between Brazilian states, we need to calculate cross-state cultural distances. Hence, we define a $u \times b$ matrix (denoted by UB) in which each row represents a Goodreads user, and each column represents a unique book. $UB[i][j]$ is set to 1 if book j is present in the bookshelf of user i. As aforementioned, each book has multiple tags assigned by its users (readers). Therefore, after the cleansing steps (Section 3.1), we connect each book in the *user-book* bipartite network to the list of its Top-3 genres.[4] In addition, we connect each user to its Brazilian state, finally creating a *state-user-book-genre* multipartite network, as illustrated by a generic example in Figure 4.

The multipartite network connects a Brazilian state to a Goodreads user, and each user is connected to one or more books. In turn, each book is connected to a list of genres. Such a complete composition can be projected into six different bipartite networks, each generated by omitting one node type from the multipartite projection. Here, we are

[4]The length of the genres list ranges from one to three items.

interested only in two bipartite networks: one connecting states and books, and another connecting states and genres. We then create the $s \times b$ adjacency matrix (denoted by SB), with rows and columns representing the states and books, respectively. $SB[i][j]$ is set to 1 if book j is present in the users' bookshelves residing in the Brazilian state i. Likewise, we also create the *state-genre* matrix SG.

Finally, after creating the adjacency matrices, we proceed to calculating the cross-cultural distances: book and genre distances. To do so, we compute the cosine similarity for each adjacency matrix created. The cosine similarity between two states i and i' is given by $1 - \frac{u \cdot v}{\|u\|\|v\|}$, where $u \cdot v$ is the dot product of vectors u and v, which represent two 1-D arrays $[i][:]$ and $[i'][:]$ of the respective adjacency matrix (SB or SG).

4.2. Cross-state distances

In order to study the significance of the relationship between the Brazilian socioeconomic and demographic landscape and the reading preferences (cultural distances), we define the following cross-state distance measures.

- **Geographical Distance**: considers the latitude and longitude of each state capital and measures the geodesic distance between pairs of states. The geodesic distance, or the great-circle distance, is calculated as the shortest path between two points on a surface.
- **Demographic and Socioeconomic Distances**: are individually calculated between state pairs for each IBGE indicator listed in Section 3 using the Canberra distance [Lance and Williams 1966, Lance and Williams 1967]. Such distance is a weighted version of the Manhattan distance, which calculates the distance between two points measured along axes at right angles. Formally, the Canberra distance d between vectors p and q in an *n*-dimensional real vector space is given by $\sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$.

Having calculated all the cross-state distances, we apply the Quadratic Assignment Procedure (QAP) test [Simpson 2001] to verify whether such measures are significantly associated with the reading preferences. The QAP test is commonly used in social network analysis and achieves high statistical power when performed on different types of complex networks [Choi et al. 2006, Fredrickson and Chen 2019, Krackardt 1987, Liu et al. 2018]. Moreover, it is a useful tool for analyzing dyadic datasets, i.e., datasets where pairs of entities are analyzed, which is our case. To perform the QAP test, we employ the *qaptest* function provided in R's SNA package [Butts 2020].

4.3. Finding Regional Identity

A central contribution of this article is to explore and distinguish potential regional identities within Brazil. Hence, to further explore identity through reading preferences, we also model two unimodal projections based on the state-book and state-genre bipartite networks (as explained in Section 4.1). In both projections, two states have a weighted edge when shared books/genres are in their population's bookshelves. Such projections result in highly dense networks, making it difficult to assess which relationships are relevant. Therefore, a method to identify significant associations in the unipartite projection is required; i.e., find a strong relation between states based on their reading similarity is non trivial. Then, with the unipartite networks modeled, we can apply a *community detection* algorithm to identify *clusters* of states with their own cultural identities.

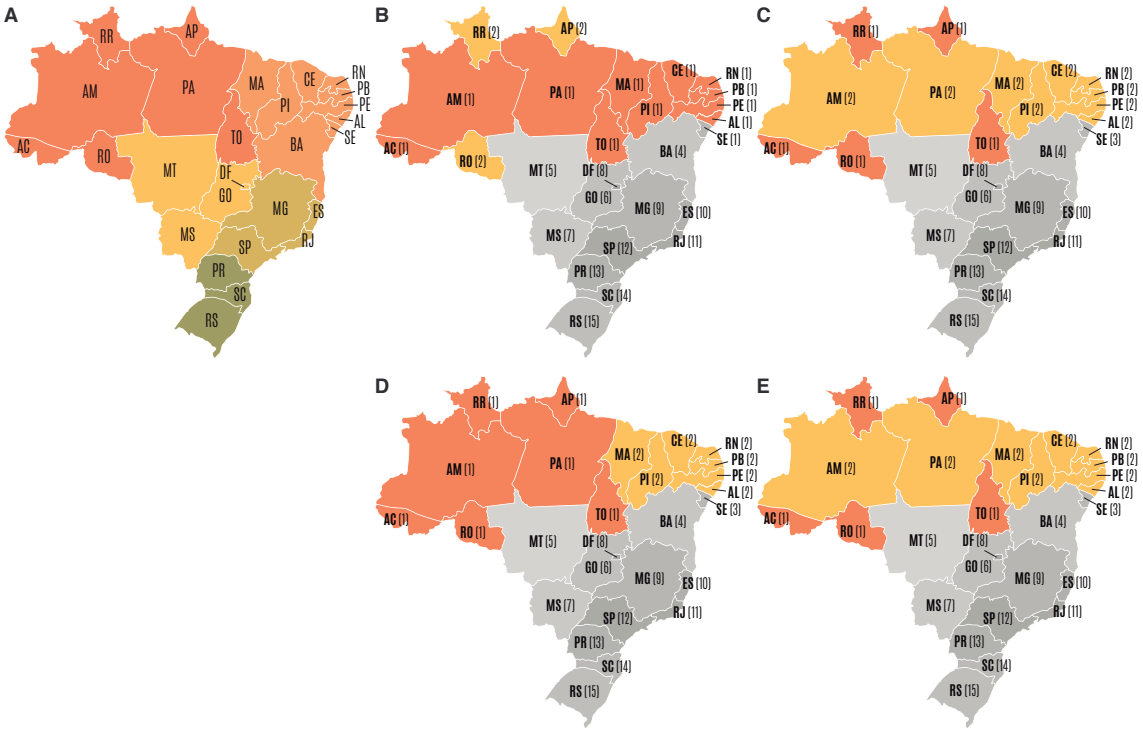

Figura 5. Comparison of the current (A) pre-filtering and (B) post-filtering book-based unipartite networks.

A simple approach to finding significant edges in unimodal projections is to preserve only links where the weight beats a specific threshold. However, such method systematically overlooks nodes with low strength (the sum of weights attached to ties belonging to a node). Hence, compelling features and structures below the cut-off scale are removed. To address such a problem, we use the disparity filter [Serrano et al. 2009]. The filtering method extracts the relevant backbone from a weighted network while retaining most of its nodes, total weight, global connectedness, small-world property, heterogeneous degree distribution and clustering. To exemplify, Figures 5A and B show the previous pre- and post-filtering of the book-based unipartite network, respectively.

We use the *backbone* function provided in R's *disparityfilter* package[5] to apply the disparity filter algorithm in both unipartite networks. Specifically, the *backbone* function identifies the "backbone structure" of a weighted graph using the disparity filter algorithm. In the algorithm, the preserved edges (i.e., relevant edges) have weights statistically different from what is expected by chance at the significance level $alpha$. Thus, we select the critical gap ($alpha$) that yields the maximum congruency with topological communities in the backbone, i.e., that maximizes the number of remaining nodes in the network and minimizes the average node degree [García-Pérez et al. 2016]. Finally, we perform the Louvain community detection algorithm [Blondel et al. 2008] on the resulting networks. Note that such cultural identity analysis is also performed on the filtered reading preference data, i.e., after removing the commonly read books.

[5]disparityfilter: `https://CRAN.R-project.org/package=disparityfilter`

4.4. Research Questions

Finally, as this article is an extension of a published paper [Silva et al. 2021b], we also compare our results with those from such prior work.[6] In particular, we organize our analyses on four research questions (RQ). Each answer is based on the aforementioned methodology with some differences as explained next.

RQ1. *Are there significant differences between Brazilian states/macro-regions concerning reading preference?* We perform a cross-state comparison by using cosine distance for books and genres preferences. We then compare the distribution of such distances for each macro-region and test such comparison with the Kruskal-Wallis test [Kruskal and Wallis 1952], a non-parametric method for testing whether samples originate from the same distribution. Finally, for an intra-region view, we use the Wilcoxon rank-sum test [Wilcoxon 1945] to confirm whether there is a significant difference between each macro-region's prior and current results. Such a method is a non-parametric approach to compare two independent groups of samples.

RQ2. *Is there a geographic, demographic or socioeconomic relationship in the reading preference of Brazilian readers on Goodreads?* We perform the aforementioned QAP tests to investigate such relationships. Then, we use Cohen's convention [Cohen 2013] to evaluate the correlation results. This convention regards the relative strength of the differences between the means of two populations based on sample data.

RQ3. *Is it possible to distinguish cultural clusters in Brazil based on the states' reading preferences?* Here, we run the *backbone* function to identify the central structure of the unipartite networks, followed by the Louvain community detection algorithm (as explained in Section 4.3).

RQ4. *Are Brazil and each of its macro-regions homogeneous or heterogeneous regarding reading preferences?* The previous RQ is at state level; now, this last RQ focuses on the macro-regions levels, then allowing a broader analysis of the Brazilian identities. Hence, it uses the same community detection methodology (as explained in Section 4.3).

5. Results

In this section, our primary goal is to provide meaningful insights into the Brazilian reading identity and associations between socioeconomic and demographic indicators. It is organized by answering the four research questions as follows: Section 5.1 answers RQ1 on differences among Brazilian states/macro-regions; Section 5.2 answers RQ2 on geographic, demographic or socioeconomic relationships; and Section 5.3 answers RQ3 on identifying cultural aspects and RQ4 on homogeneous or heterogeneous preferences.

5.1. Cross-state Cultural Analysis

We start our analyses with a cross-state comparison to explore the similarities and differences based on people's favorite books and genres. To avoid biased results generated by less populated states, we grouped the Brazilian states into macro-regions. Figures 6 and 7

[6] We refer to the results in [Silva et al. 2021b] as *prior results* from now on.

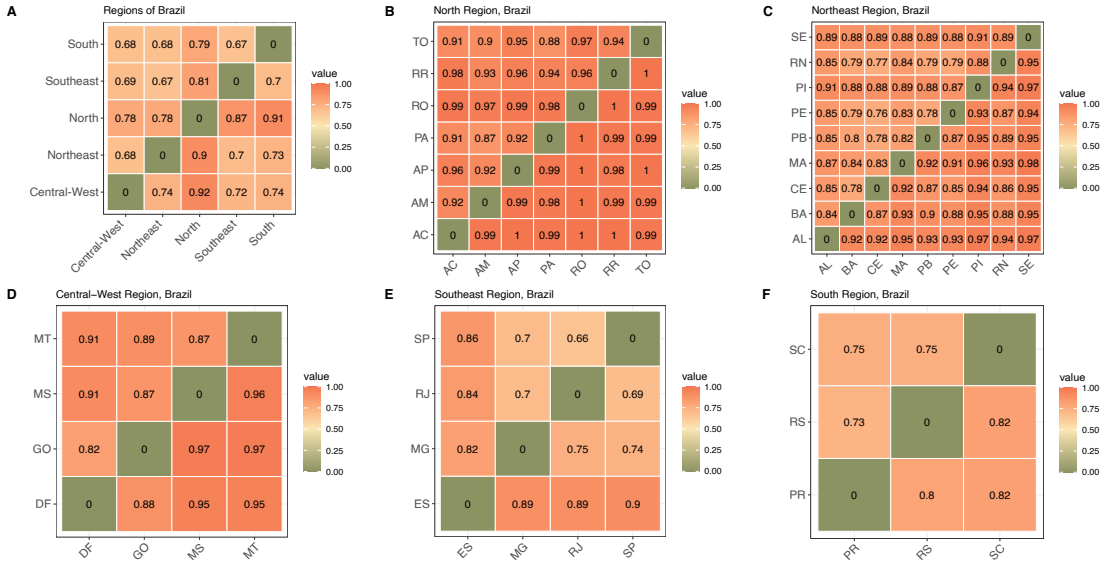

Figura 6. Cross-state Cultural Distance based on shared books. Prior and current results are presented in upper and lower triangles in the correlation matrix, respectively.

present the cross-cultural (cosine) distance of books and genre preferences, respectively. Prior and current results are presented in upper and lower triangles for each matrix. The cosine distance is bounded within $[0, 1]$, where a low value indicates a smaller distance between the respective pair of states. In both figures, shades of green suggest cultural proximity, whereas orange tones indicate a cross-cultural distance.

In prior evaluations, book-based distances were considerably more substantial than those based on genres. Such a result remained after removing the common books in the five macro-regions of Brazil. Indeed, the preference for books is naturally more specific than the genre level, as the available book search space is more extensive (75,093 books versus 80 genres). Besides book and genre perspectives, we plan to include the books' authors as an indicator of reading preference in future work. Thus, we believe that cross-state cultural distances from this additional view could bring intermediate results, i.e., not as specific as those based on books, but not as generic as in the case of genres.

From an inter-cluster perspective, North remains the most distant compared to the others regarding personal bookshelves (Figure 6A). Being the most extensive macro-region in Brazil, covering approximately 45%[7] of the entire national territory, its demographic indicators may be directly related to such a large area. On the intra-cluster scale, North (0.94/0.99 on average), Central-West (0.88/0.95 on average) and Northeast (0.84/0.92 on average) were and continue to be the most diverse macro-regions. In comparison, South and Southeast have a slightly lower level of diversity, with an average cosine distance equal to 0.76/0.81 and 0.74/0.81, respectively. Even so, the state of Espírito Santo (ES) notably remains the most distant within the Southeast, when compared to the other states in such a region.

[7]With a territorial area of 3,853,676.948 km^2.

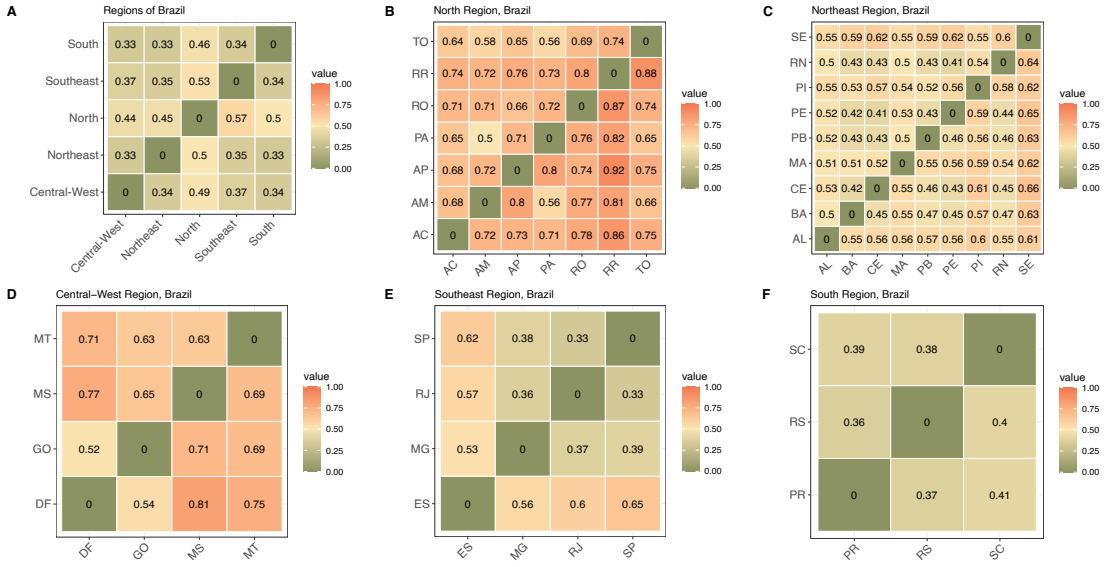

Figura 7. Cross-state Cultural Distance based on shared genres. Prior and current results are presented in upper and lower triangles in the matrix, respectively.

Regarding distances based on shared genres, Figure 7 reports similar results, although not as pronounced as in the book preferences. North remains the most distant among all, especially after filtering the commonly read books. Also, on the intra-cluster scale, North (0.68/0.77 on average), Central-West (0.65/0.70 on average) and Northeast (0.51/0.55 on average) were and continue to be the most diverse macro-regions. It is not surprising that such macro-regions correspond to the largest macro-regions in Brazil in terms of territorial area, which may indicate that factors such as geographic location and area size may be related to such cross-state cultural differences.

To answer **RQ1**, we compare the distribution of cross-cultural distances for each macro-region, grouped by prior and current results. Figures 8A and B show the comparison for distances based on shared books and genres, respectively. We use the nonparametric Kruskal-Wallis test to compare the mean distances between macro-regions, both for prior and current results. The results show a significant difference between the average cross-cultural distances within the five Brazilian macro-regions from the inter-cluster perspective. Such results support the presence of differences (and similarities) between Brazilian states/macro-regions and regarding reading preference.

For the intra-cluster view, we use the Wilcoxon rank-sum test to confirm whether there is a significant difference between each macro-region's prior and current results. Regarding the book-based distances (Figure 8A), only the Southeast does not show significant differences when removing the common books, showing greater cultural differences on average than those previously reported. Such results indicate a possible intensification of cultural diversity and Brazilian regionalism in reading preferences. On the other hand, considering distances based on shared genres, the only macro-regions that show significant differences comparing to the prior results are North and Northeast, which is to be expected, given the generic nature of the genres.

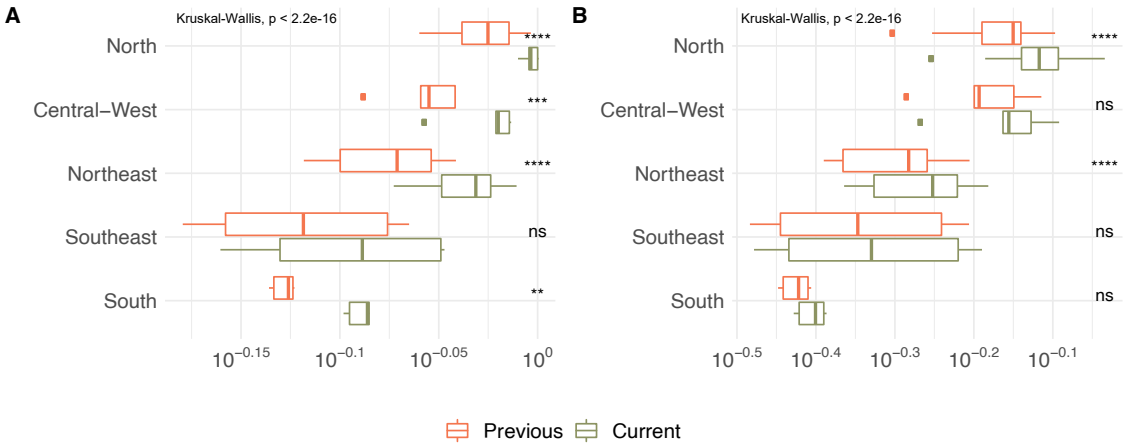

Figura 8. Cross-cultural (cosine) distances distributions for each macro-region, grouped by prior and current results: (A) based on shared books and (B) based on shared genres. Significance levels of Wilcoxon test: '*' for $p <$ 0.001, '**' for $p <$ 0.01, 'ns' otherwise.**

Tabela 1. QAP correlation results. Cross-marked values indicate the correlation is not statistically significant ($p \geq 0.01$).

Factor	$Book_p$	$Book_c$	$Genre_p$	$Genre_c$
Cultural — Genre/Book	0.914	0.831	0.914	0.831
Geographical location	0.313	0.331	0.344	0.353
Territorial area	×	×	×	×
Population estimate	×	×	×	×
Demographic density	×	×	×	×
Monthly household income	×	×	×	×
Gross Domestic Product (GDP)	×	×	×	×
Human Development Index (HDI)	×	×	×	×

5.2. Cross-state Demographic and Socioeconomic Analysis

To answer **RQ2**, we perform QAP tests to investigate the significance of socioeconomic and demographic factors on reading preferences in different states. Here, the dependent variables are matrices of cross-state cultural distances (as represented by both book and genre preferences) and the independent variables are matrices of cross-state geographical, demographic and socioeconomic distances. Table 1 reports the QAP correlation coefficients among each adjacency matrix for both prior ($Book_p$ and $Genre_p$) and current ($Book_c$ and $Genre_c$) results. Note the prior results were generated again for this analysis, as we identified an error in the data collected from the IBGE website.

To evaluate the correlation results, we use Cohen's convention [Cohen 2013] to interpret the effect size: small (0.1 to 0.3), medium (0.3 to 0.5), and large (0.5 to 1.0). Although current results' correlation coefficient between cultural distances decreased, as expected, we confirm a strong and significant association between $Book_p$ ($Book_c$) and $Genre_p$ ($Genre_c$), even after filtering the most read books. Consequently, the book- and genre-based cultural analyses resulted in similar relationships between all indicators

analyzed for both prior and current results. All associations are statistically insignificant ($p \geq 0.01$), except for geographical distance. That is, considering only demographic and socioeconomic distances between states based on the pure IBGE indicators does not reflect their difference in reading preferences.

Comparing prior with current results, the relationship between cross-state cultural and geographic distances result in slightly more significant positive and moderate associations, indicating a more solid regional component within reading preferences in Brazil. Such findings partially support that there are factors directly impacting the Brazilian population's literary choices (**RQ2**), as given by Goodread users. However, considering some indicators individually (as a single factor) generates a lack of solid and meaningful correlations, indicating the need to explore additional, more related social factors; e.g., indicators related to education, including illiteracy and schooling rates. In addition, studying the combination of certain external factors might reflect an improved association.

We also performed the same statistical analysis for each Brazilian macro-region separately.[8] However, the outcomes reveal almost no significant association at the regional level. A possible explanation is that as the level of analysis gets lower, the number of observations also decreases, affecting the correlation test. Moreover, there may be distinct social contexts within each macro-region and, therefore, the cross-region socioeconomic distances cannot explain the diversity in reading preference.

5.3. Brazilian Reading Identities

This section goes over RQ3 and RQ4. Both have the Louvain community detection algorithm in common, but each at a specific level: RQ3 for states, and RQ4 for macro-regions.

Figure 9 shows the 27 Brazilian states grouped by (**A**) macro-regions and communities detected in the unipartite projections (**B** and **D**) based on books and (**C** and **E**) genres without statistically insignificant edges. Colors distinguish communities of states. Louvain returned 15 communities for the unipartite networks based on books and genres, both prior and current results. Overall, all grouping results reported a union of most of the north and northeast regions (except for Bahia and Sergipe), which differs from what was reported in Section 5.1. Furthermore, while most northern states have gathered into a separate community, the other Brazilian states appear in unique communities. The disparity filter and, as a result, the resulting statistical connections explain such divergent findings.

When comparing prior and current results, there are no significant differences between the unipartite networks based on literary genres after filtering the *common books*. However, regarding book-based unipartite networks, the exclusion of *common books* highlights the cultural differences between North and Northeast, which were merged in the previous results. By comparing maps **A** and **D** (Figure 9), the first communities detected by Louvain form exactly Brazil's North and Northeast macro-regions, except for the states of Bahia and Sergipe, which continue to be separated into isolated communities.

Answering our **RQ3**, it is possible to classify distinct reading identities in Brazil based on the states' preferences. However, the state groupings depend directly on how

[8]Full results are available in Project Bàde homepage: `https://bit.ly/proj-bade`

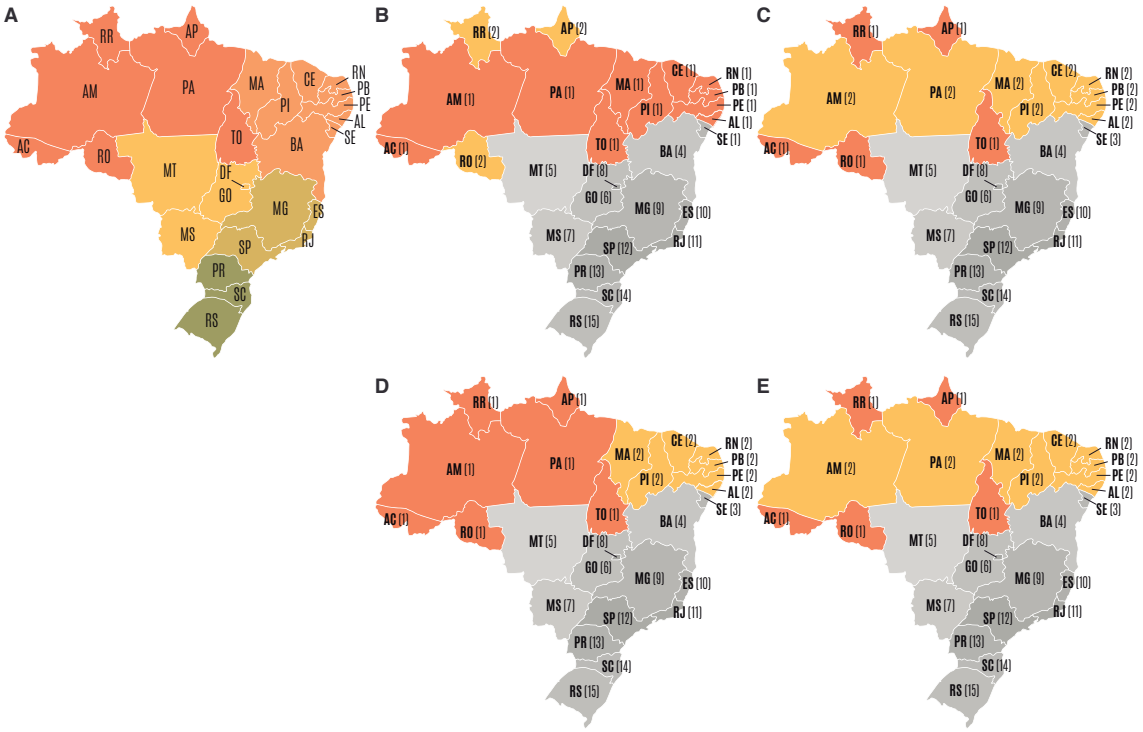

Figura 9. Groupings of Brazilian states: (A) macro-regions, (B) and (D) book-based, and (C) and (E) genre-based unipartite network communities. First and second rows represent prior and current results, respectively. Shades of gray indicate isolated communities.

cross-state connections are explored. By relying solely on cultural distances, the North stands out culturally, whereas the others are more similar to each other. Nevertheless, by evaluating the statistically significant connections, North and Northeast form two distinct communities, with the other states forming unique communities. Overall, the findings indicate that most Brazilian states hold their own reading preference characteristics.

Finally, to answer our fourth and last research question (**RQ4**), we perform the same community detection analysis in each Brazilian macro-region networks. Figure 10 shows the detected communities for each macro-region, both for the book-based (top) and genre-based (bottom) projections. As in the networks of Brazilian states, there is an evident heterogeneity in the subnetworks of macro-regions regarding reading preferences. Although there are also some merged clusters between North, Northeast and Central-West, as shown in Figures 10A(top), 10C(top), 10A(bottom), 10B(bottom) and 10C(bottom), most states form unique, isolated communities within each macro-region. Therefore, such results may indicate that Brazil and its macro-regions are indeed culturally heterogeneous, regarding reading preferences.

6. Conclusion

This paper explored the diverse Brazilian cultural identities through reading preferences. We shed light on cross-state differences and similarities based on their book and genre preferences, mainly analyzing the effect of removing books read in all five macro-regions

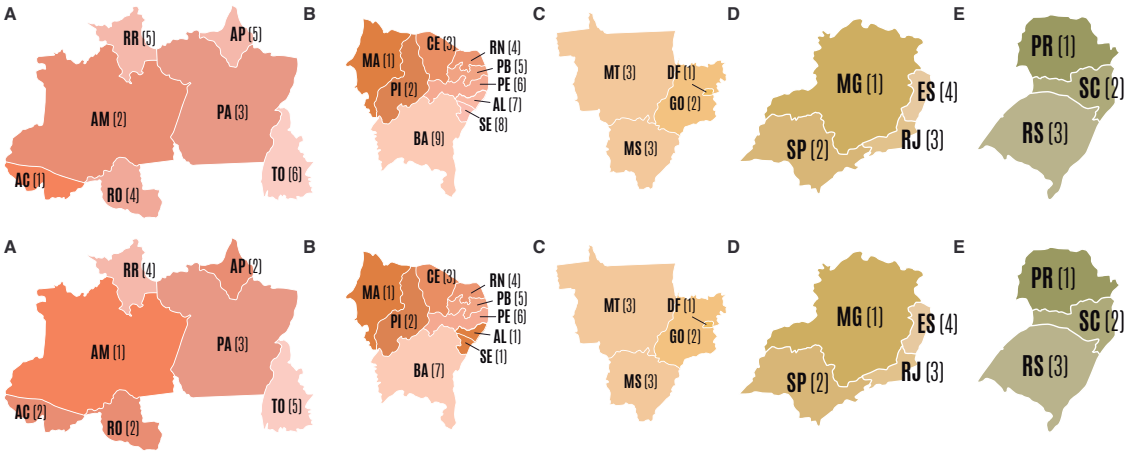

Figura 10. Detected communities for each macro-region subnetwork, both for the book-based (top) and genre-based (bottom) projections, respectively. (A) North, (B) Northeast, (C) Central-West, (D) Southeast and (E) South.

of Brazil. Additionally, through QAP correlation tests, we revealed significant relationships between cultural distances and geographical indicators. Finally, we also used a bipartite projection and a community detection method to assess statistical significance relationships between states based on their reading preferences. Overall, our findings reinforce the cultural wealth of Brazil, emphasizing that each state and region holds its own culture when it comes to what people choose to read. Such results provide insights into the driving forces behind Brazilian reading choices, indicating great opportunities for both the academy and the book industry.

From an academic view, this work represents a step further in interdisciplinary research fields involving Computer Science, Anthropology, and Sociology. For instance, knowledge of cultural behavior patterns may improve recommendation systems' accuracy and prediction models' accuracy. Besides understanding regional reading preferences, the Brazilian editorial market could properly direct its efforts to promote specific book releases for each region and state. Similarly, the government could benefit from such knowledge by investing in tax incentive policies to encourage reading habits since Brazil has lost more than 4.6 million readers from 2015 to 2019.[9] Overall, such investments directly impact the population by guaranteeing easy access to culture as well as raising literacy and other educational indicators.

Limitations and Future Work. Although Goodreads is a valuable source of data on reading habits, our findings can be biased since the final dataset does not reflect the preferences of the whole Brazilian population. As depicted in Figure 2, there are a minimal number of users in some Brazilian states, and this unbalancing may impact the quality of our results. Hence, as future work, we plan to expand our data collection to better cover readers from all Brazilian regions, enhancing future analysis. We also plan to explore additional social indicators, aiming at stronger and more significant associations between cultural distances, and consider the books' authors as a reading preference perspective.

[9]Agência Brasil, (April 15, 2021). `https://bit.ly/34EijlY`

Finally, we shall consider the weighted networks between states and books/genres to calculate cross-state cultural distances.

Acknowledgments. The work is supported by CNPq and CAPES, Brazil.

Referências

[Barbon Jr. et al. 2017] Barbon Jr., S., Tavares, G. M., and Kido, G. S. (2017). Artificial and natural topic detection in online social networks. *iSys - Brazilian Journal of Information Systems*, 10(1):80–98.

[Belinkov and Glass 2019] Belinkov, Y. and Glass, J. R. (2019). Analysis methods in neural language processing: A survey. *Trans. Assoc. Comput. Linguistics*, 7:49–72.

[Blondel et al. 2008] Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Borges de Souza et al. 2015] Borges de Souza, T. R. C. et al. (2015). Brazilian cultural differences and their effects on the web interfaces user experience. In *Int'l Conf Cross-Cultural Design Methods, Practice and Impact*, pages 209–220.

[Butts 2020] Butts, C. T. (2020). *sna: Tools for Social Network Analysis*. R package version 2.6.

[Cagliero and Quatra 2021] Cagliero, L. and Quatra, M. L. (2021). Inferring multilingual domain-specific word embeddings from large document corpora. *IEEE Access*, 9:137309–137321.

[Carosia et al. 2021] Carosia, A. E. O., Coelho, G. P., and Silva, A. E. A. (2021). Investment strategies applied to the brazilian stock market: A methodology based on sentiment analysis with deep learning. *Expert Syst. Appl.*, 184:115470.

[Choi et al. 2006] Choi, J. H. et al. (2006). Comparing world city networks: a network analysis of internet backbone and air transport intercity linkages. *Global Networks*, 6(1):81–99.

[Cohen 2013] Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

[Digiampietri et al. 2014] Digiampietri, L. et al. (2014). Análise da rede dos doutores que atuam em computação no brasil. In *BraSNAM*, pages 33–44.

[Fredrickson and Chen 2019] Fredrickson, M. M. and Chen, Y. (2019). Permutation and randomization tests for network analysis. *Social Networks*, 59:171–183.

[García-Pérez et al. 2016] García-Pérez, G., Boguñá, M., Allard, A., and Serrano, M. Á. (2016). The hidden hyperbolic geometry of international trade: World trade atlas 1870–2013. *Scientific Reports*, 6(1):33441.

[Garner 2020] Garner, J. (2020). Experiencing time in prison: the influence of books, libraries and reading. *J. Documentation*, 76(5):1033–1050.

[Guarasci et al. 2022] Guarasci, R., Silvestri, S., Pietro, G. D., Fujita, H., and Esposito, M. (2022). BERT syntactic transfer: A computational experiment on italian, french and english languages. *Comput. Speech Lang.*, 71:101261.

[Igawa et al. 2015] Igawa, R. A., Almeida, A., Zarpelão, B., and Barbon Jr., S. (2015). Recognition on online social network by user's writing style. *iSys - Brazilian Journal of Information Systems*, 8(3):64–85.

[Krackardt 1987] Krackardt, D. (1987). Qap partialling as a test of spuriousness. *Social networks*, 9(2):171–186.

[Krótkiewicz et al. 2016] Krótkiewicz, M., Jodlowiec, M., and Wojtkiewicz, K. (2016). Introduction to semantic knowledge base: Multilanguage support of linguistic module. In *European Network Intelligence Conference, ENIC*, pages 188–194.

[Kruskal and Wallis 1952] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

[Lance and Williams 1966] Lance, G. N. and Williams, W. T. (1966). Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"). *The Computer Journal*, 9(1):60–64.

[Lance and Williams 1967] Lance, G. N. and Williams, W. T. (1967). Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, 1(1):15–20.

[Liu et al. 2018] Liu, M. et al. (2018). The relation of culture, socio-economics, and friendship to music preferences: A large-scale, cross-country study. *PloS one*, 13(12):e0208186.

[Ma 2021] Ma, Q. (2021). Reading attitude among elementary school students in china: A comparison between regions of different economic development. In *ICIMTECH*, pages 55:1–55:4. ACM.

[Maity et al. 2017] Maity, S. K. et al. (2017). Book reading behavior on goodreads can predict the amazon best sellers. In *ASONAM*, pages 451–454.

[Morais et al. 2020] Morais, J. I., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Barbon Jr, S. (2020). A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in brazilian portuguese. *iSys - Brazilian Journal of Information Systems*, 13(4):126–149.

[Müller 2021] Müller, M. (2021). Reading habits of young people in the context of digital progress: An example of research of republic croatia. In *ICEIT*, pages 219–225. IEEE.

[Nascimento et al. 2018] Nascimento, M. L. et al. (2018). Uma análise do fator cultural em tecnologias persuasivas: um estudo de caso da rede social facebook. In *BraSNAM*. SBC.

[Oliveira and Merschmann 2021] Oliveira, D. N. O. and Merschmann, L. H. C. (2021). Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. *Multim. Tools Appl.*, 80(10):15391–15412.

[Oliveira et al. 2020] Oliveira, G. P., Santos, M., Seufitelli, D. B., Lacerda, A., and Moro, M. M. (2020). Detecting collaboration profiles in success-based music genre networks. In *ISMIR*, pages 726–732.

[Otter et al. 2021] Otter, D. W., Medina, J. R., and Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Networks Learn. Syst.*, 32(2):604–624.

[Pessutto et al. 2020] Pessutto, L. R. C., Vargas, D. S., and Moreira, V. P. (2020). Multilingual aspect clustering for sentiment analysis. *Knowl. Based Syst.*, 192:105339.

[Sabri et al. 2020] Sabri, N. et al. (2020). A cross-country study on cultural similarities based on book preferences. *Soc. Netw. Anal. Min.*, 10(1):86.

[Serrano et al. 2009] Serrano, M. Á. et al. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.

[Shahsavari et al. 2020] Shahsavari, S. et al. (2020). An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *WebSci*, pages 277–286.

[Silva et al. 2021a] Silva, M. O., Scofield, C., Oliveira, G. P., Seufitelli, D. B., and Moro, M. M. (2021a). BraCID: Brazilian Cultural Identity Information Through Reading Preferences. DOI: 10.5281/zenodo.4890048.

[Silva et al. 2021b] Silva, M. O., Scofield, C., Oliveira, G. P., Seufitelli, D. B., and Moro, M. M. (2021b). Exploring brazilian cultural identity through reading preferences. In *Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126, Porto Alegre, RS, Brasil. SBC.

[Simpson 2001] Simpson, W. (2001). Qap: The quadratic assignment procedure. Technical Report 1.2, North American Stata Users' Group Meetings 2001.

[Wang et al. 2019] Wang, K. et al. (2019). Exploring goodreads reviews for book impact assessment. *J. Informetrics*, 13(3):874–886.

[Wilcoxon 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. biom. bull., 1, 80–83.

[Yucesoy et al. 2018] Yucesoy, B., Wang, X., Huang, J., and Barabási, A. (2018). Success in books: a big data approach to bestsellers. *EPJ Data Sci.*, 7(1):7.

[Zhao et al. 2021] Zhao, Y. et al. (2021). Do cultural differences affect users' e-learning adoption? A meta-analysis. *Br. J. Educ. Technol.*, 52(1):20–41.