UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Stratification of Preterm Birth Risk in Brazil Through Unsupervised Learning Methods and Socioeconomic Data

Márcio Luiz Bezerra Lopes Júnior

Orientador: Prof. Dr. Marcelo Augusto Costa Fernandes

Co-orientadora: Profª Drª Raquel de Melo Barbosa

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Computação da UFRN (área de concentração: Engenharia de Computação) como parte dos requisitos para obtenção do título de Mestre em Ciências.

Natal, RN, 13 de junho de 2022

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Lopes Júnior, Márcio Luiz Bezerra.
Stratification of preterm birth risk in Brazil through unsupervised learning methods and socioeconomic data / Márcio Luiz Bezerra Lopes Júnior. - 2022.
84f.: il.

Dissertação (Mestrado) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPgEEC), Natal, 2022.
Orientador: Marcelo Augusto Costa Fernandes.
Coorientador: Raquel de Melo Barbosa.

1. Risco de PTB - Dissertação. 2. Clusterização - Dissertação. 3. Aprendizagem não-supervisionada - Dissertação. 4. k-Means - Dissertação. 5. Mapas auto-organizáveis - Dissertação. I. Fernandes, Marcelo Augusto Costa. II. Barbosa, Raquel de Melo. III. Título.

RN/UF/BCZM CDU 621.3

Elaborado por Raimundo Muniz de Oliveira - CRB-15/429

Acknowledgements

First and foremost, I'd like to thank my parents and sister for their lifelong love and support through every single step I take. Had I not had them, I would have never even gotten past the Introduction.

Next, I would like to thank my advisor, Prof. Marcelo, for all the assistance and guidance during my Master's studies. It wasn't the simplest period, with him being outside the country at first, then all of us being struck by the whole pandemic mess, having the research on hold for a while waiting for datasets that came a little too late, and, to top it all, having to adjust to my personal schedule while working full-time. I hope this work can do some justice to our initial expectations. I also thank my co-advisor, Prof. Raquel, who often attended our meetings to discuss the research proceedings and to give us constructive advice on what the research was missing and which direction to follow. Their work as advisors was very much appreciated.

I would also like to thank my sweetheart for bearing the overwhelming amount of complaints, grumbles and sighs she had to hear during the most stressful periods of these last years of work and study. I sincerely apologise, and I will assure to reduce it as much as possible – at least for a while.

Finally, I would like to extend these thanks to my entire family and friends. I must have done something incredible in a previous life to deserve having such phenomenal people caring so much about me. And at this very moment, in such a weird year, I would like to leave my most special thanks to my uncle Jackson, a man of love and smiles. Earth has gotten a little worse since February, tio, but I'm certain Heaven's never been better.

Resumo

Nascimento prematuro (PTB) é um fenômeno que traz riscos e desafios à sobrevivência de um recém-nascido. Apesar de muitos avanços na pesquisa, nem todas as causas do PTB estão bem definidas. Atualmente, entende-se que risco de PTB é multifatorial e que pode, também, estar associado a fatores socioeconômicos. Objetivando analisar essa possível relação, este trabalho busca estratificar o risco de PTB no Brasil utilizando-se apenas de dados socioeconômicos, extraindo e analisando clusters que apresentarem divergência relevante de PTB, todos os quais serão descobertos por processos de clusterização automáticos usando uma série de métodos de aprendizagem de máquina não-supervisionada. Através do uso de bancos de dados públicos disponibilizados pelo Governo Federal do Brasil, um novo banco de dados foi gerado com dados socioeconômicos a nível municipal e uma taxa de ocorrência de PTB. Esse banco de dados foi processado utilizando dois métodos de clusterização distintos, ambos construídos através da união de métodos de aprendizagem não-supervisionada, tais como *k*-médias, análise de componentes principais (PCA), clusterização espacial baseada em densidade de aplicações com ruído (DBSCAN), mapas auto-organizáveis (SOM) e clusterização hierárquica. Os clusters com alto PTB foram formados majoritariamente por municípios com baixos níveis educacionais, com pior qualidade de serviços públicos – como saneamento básico e coleta de lixo – e com populações mais brancas. A distribuição dos clusters também foi observada, com clusters com alto PTB concentrados nas regiões Norte e Nordeste. Os resultados indicam, uma influência positiva da qualidade de vida e da oferta de serviços públicos na redução do risco de PTB.

Palavras-chave: Nascimento prematuro, Risco de PTB, Clusterização, Aprendizagem Não-supervisionada, k-Means, Mapas Auto-organizáveis, Brasil.

Abstract

Preterm birth (PTB) is a phenomenon that brings risks and challenges to the survival of the newborn child. Despite many advances in research, not all the causes of PTB are yet clear. It is currently understood that PTB risk is multi-factorial and may also be associated with socioeconomic factors. In order to analyse this possible relationship, this work seeks to stratify PTB risk in Brazil using only socioeconomic data, extracting and analysing those clusters that present relevant PTB divergence, all of which will be found by automatic clustering processes using a series of unsupervised machine learning methods. Through the use of datasets made publicly available by the Federal Government of Brazil, a new dataset was generated with municipality-level socioeconomic data and a PTB occurrence rate. This dataset was processed using two separate clustering methods, both built by assembling unsupervised learning techniques, such as *k*-means, principal component analysis (PCA), density-based spatial clustering of applications with noise (DBSCAN), self-organising maps (SOM) and hierarchical clustering. The methods discovered clusters of municipalities with both high levels and low levels of PTB occurrence. The clusters with high PTB were comprised predominantly of municipalities with lower levels of education, worse quality of public services – such as basic sanitation and garbage collection – and a less white population. The regional distribution of the clusters was also observed, with clusters of high PTB located primarily in the North and Northeast regions of Brazil. The results indicate a positive influence of the quality of life and the offer of public services on the reduction of PTB risk.

Keywords: Preterm birth, Clustering, Unsupervised learning, PTB risk, k-Means, Self-Organising Maps, Brazil.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

CADU	Cadastro Único
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
IBGE	Instituto Brasileiro de Geografia e Estatística / *Brazilian Institute of Geography and Statistics*
MkM	Multiple k-Means
PCA	Principal Component Analysis
PMR	Preterm Birth Municipal Rate
PTB	Preterm Birth
SES	Socioeconomic Status
SINASC	Sistema de Informações sobre Nascidos Vivos / *Live Birth Information System*
SOM	Self-Organising Maps
t-SNE	t-Distributed Stochastic Neighbour Embedding

Chapter 1

Introduction

Preterm birth (PTB), defined as a birth happening prior to the 37th week of pregnancy, is the most common cause of mortality among children 5 years old or younger (França et al. 2017, Modell et al. 2012, Organization 2012). In addition, it was shown as a critical factor for the survival of newborns (Modell et al. 2012). Preterm born babies present a major challenge to medical assistance, which needs to supplement their yet not fully developed vital organs (Institute of Medicine 2007). Trying to understand and prevent the causes of PTB has become increasingly more common in scientific research, especially with the recent emergence of progressively more trustful and more complex government-owned datasets. Getting closer to this goal could mean finding ways of preventing or, at least, anticipating PTB, thus providing assistance to the mother in time, possibly reducing the number of lives lost.

The work presented by Adhikari et al. (2019) shows that PTB's etymology is multifactorial and that the risk of PTB could be associated with the socio-economical situation of a given region (*neighbourhood socioeconomic status* or *neighbourhood* SES). *Neighbourhood SES* is an area-level measurement that aggregates SES factors (e.g. income, education and employment) at a particular geographic level (Kawachi & Berkman 2003). Works in the literature show that PTB rates are higher in areas with low SES when compared to those with high SES (Metcalfe et al. 2011).

Numerous machine learning techniques have been previously applied to the problem of PTB prediction or stratification, including SVMs (Santoso & Wulandari 2018), neural networks (Włodarczyk et al. 2020, Catley et al. 2006, Kim 2019) and decision trees (Hill et al. 2008, Lee et al. 2021). However, the most commonly applied techniques are logistic regression and linear regression, employed in the analysis and prediction of PTB for various factors: poverty (DeFranco et al. 2008), pregnant mother's working conditions (Buen et al. 2020, Saurel-Cubizolles et al. 2004), general social factors (Kaufman et al. 2008, Beeckman et al. 2009) and, mainly, clinical and hereditary factors (Grjibovski et al. 2005, Oliveira et al. 2019, Chen et al. 2020, Alleman et al. 2013). There is a vast literature associating different elements with preterm birth using traditional statistical methods (Sun et al. 2019, Granese et al. 2019, Huang et al. 2020), including associations with social factors (Baker et al. 2018, Ruiz et al. 2015).

One way of comprehending the associated risk of the many SES factors to the occurrence of PTB is through data clustering. Clustering is a segment of unsupervised machine learning techniques seeking to associate and group elements together without any initial

comprehension of the data itself. In order to do so, clustering techniques use distancing algorithms to judge how close or similar two points are to each other, and whether or not they should be in the same cluster. Clustering techniques have been used for scientific analyses for decades in many areas, such as psychology (Borgen & Barnett 1987), genetics (Ben-Dor et al. 1999) or geophysics (Sun & Li 2015).

Clustering as the way to discover the groups which are more vulnerable to PTB risk is less common than traditional statistical methods. However, its application can already be seen in some recent studies. In Istvan et al. (2019), spacial clustering shows a possible relation between living closer to landfills and PTB occurrences. Studies presented by Passini et al. (2014) and Esplin et al. (2015) show the clustering of hereditary and behavioural factors, associating them with PTB risk. Also, a work by Deguen et al. (2018) investigates the geographical distribution of PTB risk in Paris by clustering at the level of "*census blocks*".

Thereby, the main goals of this work are to stratify the risk of PTB in Brazil from SES factors alone, to obtain a general and feature-level view of factors that may affect PTB, to uncover which areas of Brazil have a higher risk of experiencing PTB, and to do all that automatically – leaving the decision of feature relevance entirely to the machine – using linear and non-linear algorithms. The stratification process is done through clustering analysis based on unsupervised machine learning techniques. The analysis was done by combining three datasets collected by the Federal Government of Brazil: *Sistema de Informações sobre Nascidos Vivos* (SINASC) (Datasus n.d.), including data regarding gestation, birth, newborns and mothers; *Cadastro Único* (CADU) (Ministério da Cidadania n.d.), containing a wide range of socioeconomic data from Brazilian citizens on a personal and family level; and the population estimate as disclosed by IBGE (IBGE n.d.).

A new dataset was generated from these datasets, and a new metric – PTB Municipal Rate (PMR) – was created. These two were used together in analysis, at a municipal level, seeking to visualise the relation between SES factors and PTB risk.

The results show a regional disparity between richer and poorer parts of Brazil concerning PMR, with High PMR clusters placed primarily in the North/Northeast and Low PMR in the Centre-South. Based on the discovered clusters, it's possible to map municipalities according to PTB risk. Besides, this study presents the leading SES factors associated with High and Low PMR clusters. The results presented in this work might contribute to the elaboration of more efficient and specialised politics for the Brazilian public health service.

1.1 State of the Art

The relationship between SES factors and occurrence of PTB, as mentioned before, has been studied by many authors, mostly but not exclusively dealing with just one or two "dimensions" of SES (e.g., education and income) and traditional statistical comparison methods rather than machine learning. For instance, this is the case observed by Saurel-Cubizolles et al. (2004), where working conditions are observed together with preterm birth. Women with long work-hour schedules and those reportedly dissatisfied with their

current work are shown to have significant higher risk of PTB in European countries. Women working excessively long hours (over 43h/week) were found to have a preterm delivery odds ratio of 1.33 compared to the unity (30-39 h/week), and women who had to work in standing position for over 6 hours had an odds ratio of 1.26 compared to the unity (less than 2h). These findings put working conditions and stressful situations as some of the possible non-biological factors to influence PTB, a view also strengthened by the results observed by Stylianou-Riga et al. (2018), whose study observes the same relation of work, stress and PTB in Cypriot women.

The possibility of a certain region's sanitation and housing conditions affect birth delivery time is explored by recent related studies by Padhi et al. (2015), Baker et al. (2018), and Patel et al. (2019). These studies present access to proper sanitation facilities as possibly an important factor to help increase PTB occurrences among Indian women. All studies obtained a statistically significant difference on frequency of PTB outcomes when contrasting people with toilet access with people with no toilet access. Furthermore, the results of Baker et al. (2018) also suggest that the harassment of girls and women (stressful event) and excessive time fetching water (over 2h/day, manual labour) increase the risk of PTB, with odds ratios of 1.26 and 1.33, respectively. The results of Patel et al. (2019) and Padhi et al. (2015) also include analyses on education data, and, in both studies, women with higher levels of education appear to have significantly lower risk of PTB.

Education as a social factor that raises the risk of PTB is also defended by the meta-analysis presented by Ruiz et al. (2015). The analysis is done over 12 distinct countries' groups of mother, collected in different years and using different education indicators. Its results indicate that mothers with low levels of education are more likely to experience PTB, with an increased risk of 48% and 84% on two scoring methods used. This is a considerably large difference and is a strong indication that education is an important aspect when exploring PTB factors. The same idea is given by the study provided by Cantarutti et al. (2017), where the higher educated women in Lombardy are shown to have 19% less risk of experiencing PTB, a reduced risk also observed when analysing foreign-born and local-born mothers separately.

The notion of relating all or most of these social factors at once, studying and treating them all as factors of social deprivation or social inequality, is seen, with association to preterm in the study presented by Adhikari et al. (2019). The study merges these factors into an SES Neighbourhood feature and associates it with personal data from the patients. The results given by intra-cluster correlation indicate SES neighbourhood-level circumstances to be responsible for 5.72% of all variance in PTB. Although only a small portion of the total variance, this can have considerable impact on model fine-tuning if one aims to develop a preterm predictor, and it provides a strong case for the continued studies on socioeconomic factors and PTB.

Another study to tackle the relationship of SES Neighbourhood and preterm was presented by Ochoa et al. (2021) and also had results that advocate for the importance of the socioeconomic environment to PTB. The main difference of this study when compared to Adhikari et al. (2019), is that their study used income variation over time as a way of measuring the socioeconomic status of neighbourhoods, with this different method obtaining

final numbers that showed that women living in areas of low socioeconomic indices or in areas where socioeconomic levels are declining have higher risk of PTB occurrence. Stable Low-level areas (i.e., low-level areas that don't show progress in socioeconomic factors) had the highest odds ratio of 1.20 – compared to Stable High-level areas.

A recent study by Deguen et al. (2018) also uses SES Neighbourhood to investigate preterm birth across the city of Paris' block areas, using spatial clustering, and its results endorse the idea of SES factors as an influential factor of PTB. When using SES Neighbourhood as cluster detection variables, the clustering resulted in a final cluster division with a p-value of 0.06, but when adjusting for SES, removing it from the clustering, the p-value increased to 0.81, a much less significant number, indicating that SES Neighbourhood was responsible for a great part of the explainable PTB variance.

As it has already been put, most of these works presented above, as well as most of the non-cited related literature, make use of traditional statistical methods, associating a selected range of features and verifying possible correlations. The three latest mentioned works, Adhikari et al. (2019), Ochoa et al. (2021), and Deguen et al. (2018), go one step further and work with a merged value of many dimensions, but are still in need of a subjective human decision on how to unite these values into a significant feature. A few questions yet unanswered or only partially answered on PTB and SES are: (1) If such a relationship exists and is significant, can high and low PTB areas be discovered through the clustering of SES factors? (2) Is this relationship intrinsic enough that it can be found automatically by a machine without any significant feature selection? (3) Is it possible to uncover the socioeconomically deprived areas most likely to suffer from high PTB numbers? (4) Which SES factors are the most contrasting in regions with high and low PTB occurrences? Therefore, the current work contributes to the area by attempting to fully or partially answer these questions by using two distinct unsupervised learning methods, to explore large Brazilian datasets of SES and birth data.

By combining k-Means and DBSCAN, two very different clustering algorithms, another pivotal contribution of this work is a new method for targeted cluster analysis. The algorithm initially provides a target-blind clustering layer of k-Means, with results then filtered by the target variable. The filtered results are then passed to a final/decision DBSCAN layer, which generalises and removes the clusters found by k-Means to provide the final results. This method allows us to completely isolate PTB from the SES clustering, done through k-Means, while also finding significant clusters without having to rely on traditional optimal cluster techniques, which would ignore the external targeted variable and thus not meet the demands of some desired clustering cases.

We also contribute with a second method, which deviates from the first method by bringing a non-linear solution to the problem, as the data exploration is done through SOM, a special kind of neural network that is more naturally suitable for higher dimensionality in comparison to k-Means. The method combines SOM with hierarchical clustering to generalise the SOM output and provide humanly comprehensible results.

1.2 Organisation

This document is divided into six chapters: (1) Introduction, (2) Materials and Methods, (3) the *k*-Means Method, (4) the SOM Method, (5) Comparison and Overview and (6) Conclusion.

The second chapter introduces the datasets used for the purpose of this work, as well as the concepts and algorithms of machine learning applied.

The third chapter presents the first method used in the research, the *k*-Means Method, its architecture, its results, and some brief commentary on what it achieved.

The fourth chapter presents the second method used, the SOM Method, and similarly to the third chapter, it presents the architecture, results and brief commentary.

The fifth chapter presents a discussion of the results obtained in both methods, as well as a comparison between them. It contextualises our findings with comparisons to some related previous works.

The sixth and final chapter is the conclusion, with statements on what this research achieved, its implications and limitations, and some propositions of future research paths.

Chapter 2

Materials and Methods

In this chapter, we present in Section 2.1 the datasets used as inputs for our models. In 2.1.1 we first present the original datasets – all of which are publicly available government-maintained datasets – and their sources, giving a detailed explanation of their sizes and features. Then, we explain in 2.1.2 the preprocessing steps taken to turn the raw (as made available) data into a single municipal-level dataset ready to be processed by clustering algorithms, including how we calculated PMR (PTB Municipal Rate) for each municipality. Next, in 2.2, we define the key clustering elements employed in this research. Finally, in 2.3, we mention the implementations used.

2.1 Datasets

2.1.1 Sources

As described in the introduction, three datasets were used to generate the training set: SINASC, CADU and IBGE. The entire analysis was performed using data from 2018 version of these datasets.

SINASC, here characterised by the variable $\mathbf{T}_{\mathrm{SN}}$, is a dataset with 61 features and almost 3 million samples, it stores data related to births that occurred in Brazilian territory and it can be found at the DATASUS website (Datasus n.d.). For the purpose of this work, two columns of $\mathbf{T}_{\mathrm{SN}}$ were used, one related to the gestational period length and the other to the mother's municipality of residence, as shown in Table 2.1.

Table 2.1: SINASC dataset variables, $\mathbf{T}_{\mathrm{SN}}$.

SINASC ($\mathbf{T}_{\mathrm{SN}}$)	
Indexer	Mother's Residential Municipality Code
Selected	Weeks of Pregnancy
Dropped	59 others

The CADU dataset was split into two distinct *datasets*: *CADU Individual*, characterised by the variable $\mathbf{T}_{\mathrm{P}}$, and *CADU Household*, characterised by $\mathbf{T}_{\mathrm{F}}$.

The $\mathbf{T}_{\mathrm{P}}$ dataset has over 12 million samples of Brazilian citizens, each with 26 features, storing basic individual data, such as gender, age, and race, but also more specific data about education and employment, as described in Table 2.2.

Table 2.2: CADU Individual dataset variables, $\mathbf{T}_{\mathrm{P}}$.

CADU Individual ($\mathbf{T_P}$)	
Indexer	ID Individual, ID Household
Selected	Gender, Age, Race, Residential Municipality Code, Place of Birth, Disability, Literacy, Type of School, Educational Level, Employment Situation, Type of Job, Total Income, Welfare Income
Dropped	Degree of Kinship to Head of Household, Regional Information

In the $\mathbf{T}_{\mathrm{F}}$ dataset, there are about 4 million household samples each with 23 features, including data about living conditions, household income, and type of family, as detailed in Table 2.3.

Table 2.3: Dataset variables, $\mathbf{T}_{\mathrm{F}}$.

CADU Household ($\mathbf{T_F}$)	
Indexer	ID Household
Selected	Type of Property, Amount of Rooms, Wall Material, Floor Material, Water Supply, Sanitary Drainage, Garbage Collection, Lighting, Pavement, Special Groups Classification, Average Household Income
Dropped	Register Date, Modification Date, Update Date, EAS/MS Code, CRAS/CREAS Code

The population dataset used here is the official 2018 population estimate dataset by IBGE, which includes estimates for 5,570 municipalities. Represented here by $\mathbf{T}_{\mathrm{IBGE}}$, this dataset has 5 columns, described in Table 2.4, of which only 2, referring to total population and municipality code, were used.

Table 2.4: IBGE dataset variables, $\mathbf{T}_{\mathrm{IBGE}}$.

IBGE Population Estimate ($\mathbf{T_{IBGE}}$)	
Indexer	Municipality Code
Selected	Population
Dropped	Municipality Name, Federal Unit, Federal Unit Name

The sources and dimensions of each dataset for the year 2018 can be seen in Table 2.5.

2.1.2 Preprocessing

To join the information available in all four datasets and create a combined dataset input for clustering, we designed a preprocessing stage, subdivided into Preprocessing

Table 2.5: Summary of datasets used.

Dataset	Dataset	Year	Samples	Features
$\mathbf{T}_{\text{SN}}$	SINASC	2018	2,944,932	61
$\mathbf{T}_{\text{IBGE}}$	IBGE	2018	5,570	5
$\mathbf{T}_{\text{P}}$	CADU Individual	2018	12,852,599	26
$\mathbf{T}_{\text{F}}$	CADU Household	2018	4,807,996	23

P_1, dealing with $\mathbf{T}_{\text{SN}}$ and $\mathbf{T}_{\text{IBGE}}$, and preprocessing P_2, dealing with $\mathbf{T}_{\text{P}}$ and $\mathbf{T}_{\text{F}}$. P_1 and P_2 generate intermediate datasets $\mathbf{I}_1$ and $\mathbf{I}_2$, respectively, as output. These intermediate datasets are composed of 5,570 samples each, one for each Brazilian municipality, and they can be joined together using the unique Municipality Code, generating $\mathbf{A}_0$. Figure 2.1 details the general preprocessing scheme.

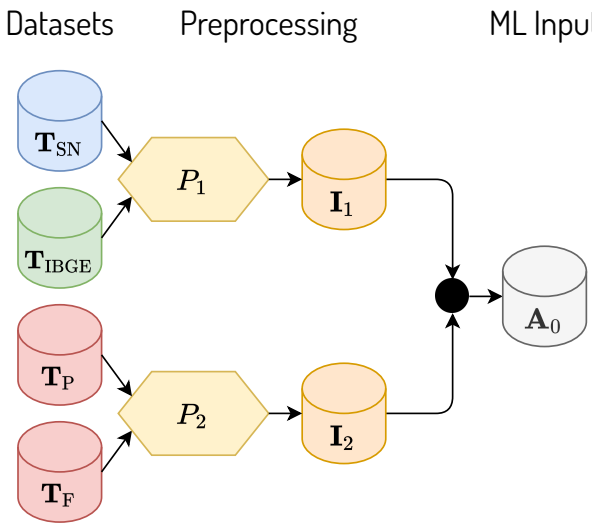


Figure 2.1: General preprocessing scheme to generate dataset $\mathbf{A}_0$ (machine learning process input), including preprocessing stages P_1 and P_2, their outputs $\mathbf{I}_1$ and $\mathbf{I}_2$, and the original datasets $\mathbf{T}_{\text{SN}}$, $\mathbf{T}_{\text{IBGE}}$, $\mathbf{T}_{\text{P}}$ and $\mathbf{T}_{\text{F}}$.

■ Preprocessing P_1

As described in Figure 2.1, the SINASC dataset, characterised by the variable $\mathbf{T}_{\text{SN}}$, was preprocessed to obtain the number of preterm births per municipality.

First, $\mathbf{T}_{\text{SN}}$ was filtered by weeks of pregnancy, keeping only the samples with less than 37 weeks. Then, $\mathbf{T}_{\text{SN}}$ was grouped by the mother's residential municipality code, counting the number of preterm births for each municipality.

Next, we joined the the grouped $\mathbf{T}_{\text{SN}}$ with the population estimate dataset ($\mathbf{T}_{\text{IBGE}}$) through the Municipality Code, adding to $\mathbf{T}_{\text{SN}}$ information on the municipalities' population sizes and thus generating the intermediate dataset $\mathbf{I}_1$

On $\mathbf{I}_1$, we calculated the PTB Municipal Rate (PMR). PMR is the metric proposed in this work to measure the frequency of PTB occurrences by municipality, expressed as the following:

$$\text{PMR} = \frac{N_{\text{NP}}}{N_{\text{P}}} \tag{2.1}$$

where N_{NP} is the total number of PTB occurrences in a given municipality and N_{P} is the population of that same municipality.

During the research, we decided to use the total population instead of the total number of births, which could be obtained from $\mathbf{T}_{\text{SN}}$, because we observed that some of the PTB over the total number of births fractions were resulting in very unrealistic percentages. For instance, one municipality had 70% of its registered births reported as PTB. Even though we remove the most extreme values later in process P_3, the total number of births was replaced by the population to reduce the possibility of these unbalanced values impacting municipalities of closer-to-the-median PMR, assuming that $\mathbf{T}_{\text{SN}}$ is missing data.

The output of preprocessing P_1 is the intermediate dataset $\mathbf{I}_1$, comprised of 2 columns: Municipality Code and PMR.

■ Preprocessing P_2

As described in Figure 2.1, datasets CADU Individual, expressed by variable $\mathbf{T}_{\text{P}}$, and CADU Household, characterised by variable $\mathbf{T}_{\text{F}}$, were preprocessed in order to turn their categorical features into numerical features, able to be used by the selected clustering algorithms.

First, the dataset $\mathbf{T}_{\text{P}}$ was filtered, removing all samples of people who are male, and also female under 14 or over 40 years of age, removing ages of recognisable less fertility (Group 2005). Next, one-hot encoding was applied to all the categorical features, generating 29 new binary features. Finally, an additional processing was to aggregate educational features representing the same educational level. As for the $\mathbf{T}_{\text{F}}$ dataset, that represents data on a family level, one-hot encoding was applied to all categorical features, generating 48 new binary features. The datasets were joined by the Household ID, present in both, adding household data present in $\mathbf{T}_{\text{F}}$ to each person sample of $\mathbf{T}_{\text{P}}$ that belongs to that household. At the end of the process, values were grouped by municipality of residence, calculating each feature's average by municipality. Thus, the output of P_2 was the intermediate dataset $\mathbf{I}_2$.

■ Output $\mathbf{A}_0$

In order to generate the general preprocessing output, indicated by the dataset $\mathbf{A}_0$, the outputs of P_1 and P_2 (datasets $\mathbf{I}_1$ and $\mathbf{I}_2$, respectively) were joined by Municipality Code. Dataset $\mathbf{A}_0$ is comprised of $5,529$ samples and 104 features, and each sample represents a Brazilian municipality. This dataset is the common starting point for the two clustering methods applied in this study, and all of its variables can be seen in the appendix Tables A.1, A.2 and A.3.

2.2 Theory

2.2.1 Unsupervised Learning

Machine learning algorithms are traditionally split into 3 main categories: (1) Supervised Learning, (2) Unsupervised Learning and (3) Reinforcement Learning. The first deals with previously labelled data, associating each element of the training data and its features with their respective given label, and finding patterns to link certain arrangement of features to each label found in the dataset. The latter category does not make use of labels, but uses a set of rules or scoring strategy and passes it to the training algorithm in order for it to train and discover which path leads to better scoring. The second, Unsupervised Learning, which makes use of no predetermined target (human-annotated label or rule), is the one applied here in this work. In unsupervised learning, the aim is to exploit similarities among the N dimensions of some input data, identifying data samples of similar behaviour automatically, and then classifying them into *clusters* (Marsland 2009).

A common problem of unsupervised learning is *finding the best cluster representation*, that is, finding the cluster division that best represents the input data. "Best" can be a subjective word, but in this sense, this means a representation that preserves a much information about the clusters while also maintaining a certain level of simplicity that allows it to be humanly comprehensible (Goodfellow et al. 2016).

The "best representation" problem just mentioned is unsupervised learning applied to learning individual samples behaviour for rows/samples classification. A different approach can be used and applied not to samples, but to features, to learn how features are related to each other. This is most commonly used to reduce the number of features in a dataset, which is referred to as "data compression" or "dimensionality reduction". The new features are, in a sense, clusters of the original features with different levels of belonging.

2.2.2 k-Means

The k-Means algorithm, presumably the most commonly used unsupervised learning algorithm for clustering, is an iterative and fixed-number of cluster type of algorithm. Its goal is to, starting from initially set group of k proposed clusters-centres, re-position these centres over and over again on the N-dimensional space, always associating, at each round, each data element to the proposed centre that is closer to such element, and moving the centres in the direction of the midpoints between the element that were associated with them. These actions will be repeated until such an arrangement is reached where all proposed cluster centres are located exactly in the midpoint of its associated samples.

The idea of the algorithm is very straight-forward and considerably simple to implement when compared to other clustering algorithms, but it relies on a few key parameters defined by the user: (1) starting position of proposed centres, (2) number of proposed centres, and (3) tolerance/limit of total iterations. In regards to the first parameter, it is common to either start the centres in random positions, or to use an optimisation algorithm, such as *k-means++*, to find better starting points. The second parameter, however, falls into one of the main difficulties of clustering, as there is no single truly right criterion

to measure how well a clustering was performed, and the number of centres for k-Means clustering can completely change the final result, therefore selecting the correct number of cluster isn't an ordinary task. Some "rules-of-thumb" do exist, and running k-Means over and over again while varying the number of centres and calculating cluster distances is a common strategy to select the best number of k, but it isn't applicable to every problem, as the "best" number might be too general that it extracts no relevant information for analysis, or too large that it becomes incomprehensible.

2.2.3 PCA

As mentioned in 2.2.1, some unsupervised learning algorithms are used with the purpose of dimensionality reduction, Principal Component Analysis, or simply **PCA**, is one of those algorithms. The main goal of PCA is to find a way of representing the dataset in lower dimensions, in a way where the newly PCA-created dimensions are all linearly uncorrelated (Goodfellow et al. 2016). The PCA process can be seen mathematically through the application of linear algebra concepts, such as Singular Value Decomposition (SVD) and Eigendecomposition.

Considering a matrix $\mathbf{X}$, the SVD decomposition of $\mathbf{X}$ can be expressed by

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{2.2}$$

where $\mathbf{U}$ is a unitary matrix, $\mathbf{S}$ is a diagonal matrix of singular values, and $\mathbf{V}^T$ is the transposed version of eigenvector $\mathbf{V}$. This eigenvector $\mathbf{V}$ is calculated by decomposing (Eigendecomposition) the covariance matrix of $\mathbf{X}$, here referred to as $\mathbf{A}$. $\mathbf{A}$, as covariance matrix, can be defined as

$$\mathbf{A} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \tag{2.3}$$

where n is the number of rows in $\mathbf{X}$.

And $\mathbf{A}$ can also be decomposed to obtain $\mathbf{V}$:

$$\mathbf{A} = \mathbf{V}\mathbf{L}\mathbf{V}^T \tag{2.4}$$

where $\mathbf{L}$ is a diagonal matrix of eigenvalues.

By replacing $\mathbf{X}$ in 2.3 with the decomposed form shown in 2.2, we obtain

$$\begin{aligned} \mathbf{A} &= \frac{\mathbf{V}\mathbf{S}\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T}{n-1} \\ &= \frac{\mathbf{V}\mathbf{S}^2\mathbf{V}^T}{n-1} \\ &= \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T \end{aligned} \tag{2.5}$$

which ends up in a similar form as the eigenvector decomposition shown in 2.4, to which

we can also formulate

$$\begin{aligned} \mathbf{VLV}^T &= \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T \\ \mathbf{L} &= \frac{\mathbf{S}^2}{n-1} \end{aligned} \tag{2.6}$$

meaning that the values of $\mathbf{S}$ can be obtained through $\mathbf{L}$.

The principal components of PCA, are then obtained through the multiplication of $\mathbf{V}$ – the principal directions – and $\mathbf{X}$, which results in

$$\begin{aligned} \mathbf{XV} &= \mathbf{USV}^T\mathbf{V} \\ &= \mathbf{US} \end{aligned} \tag{2.7}$$

so that now we can obtain the PCA components through the eigenvalues.

A limitation of PCA is that, since it aims at generating a linearly uncorrelated series of dimensions, this linear algebra algorithm ignores non-linear relations among the original dimensions, extracting only the linear behaviour through the eigenvalues. So, possible non-linear relations might actually be lost during the reduction process.

2.2.4 DBSCAN

Density-Based Spatial Clustering of Applications with Noise, or simply **DBSCAN**, is a density-based clustering algorithm, design to discover clusters of arbitrary shape (Ester et al. 1996). Unlike *k*-Means, the DBSCAN algorithm does not need as input an initial number of clusters, with the actual clusters being naturally discovered by the algorithm during the process.

DBSCAN initially treats every single data point as a cluster, and at each step, it "merges" the closest available two points into the same cluster, in a way turning two smaller clusters into a slightly bigger one. This is done iteratively throughout the entire process, until such a point is reached where no pair of points is close enough to each other, based on the only mandatory parameter of the DBSCAN algorithm: *maximum distance between two points*. Summarising the steps: (1) Find the two closest points where at least one hasn't been merged yet (2) If distance is less than the maximum allowed, merge (3) If not, finish process.

At the end of the process every point will belong to a final cluster, even if the cluster is formed by only that point, so in order to avoid these cases, one optional parameter can be used: *minimum number of samples*. Essentially, this means that every final cluster with less samples than the predetermined minimum will not be treated as a cluster, but as *noise* within the data.

2.2.5 Self-Organising Maps

Self-Organising Maps, also known as *Kohonen Network*, *Self Organising Feature Maps*, or simply ***SOM***, are a special unsupervised learning class of artificial neural net-

work designed specifically for the purpose of data analysis and visualisation (Kohonen 2014).

SOM networks are based on the principle of competitive learning, given a initial set of neurons, the network creates a competition among these neurons where only the neuron that best fits the data entered into the network can be activated at each time, every neuron adapts to a given data and only one is chosen, the "winner-takes-all" neuron. But one key point of the algorithm is that those neurons are not completely independent from each other, unlike most artificial neural models, SOM's neurons are given positions, and a neuron's result has impact on its neighbouring neurons. A SOM network organises its output neurons in a lattice of D dimensions, in predetermined size and topology, in such a way that the "winner-takes-all" neuron (given by the distance function) alters its neighbours' current state, through the neighbourhood function, by modifying their weight vectors to make them resemble more of the input data. This spatial-sensitive characteristic of the network allows it to conveniently process and translate complex feature-rich problems into fewer dimensions representations, and since it is usually done for data analysis and visualisation, SOM lattices are way more common in humanly comprehensible sizes $D = 2$ and $D = 3$. In a way, the output lattices of SOM network generate a brand new coordinates system for the input data, as an (X,Y) positioning within the lattice has actual meaning to the input data.

Mathematically, the network decides the winner neuron using

$$(u,v) = argmin||\mathbf{x}(t) - \mathbf{w}_{i,j}(t)|| \tag{2.8}$$

where (u,v) are the winner neuron's *index*, t is the step of the algorithm, $\mathbf{x}$ is the input data, and $\mathbf{w}_{i,j}$ is the weight vector of the neuron with coordinates (i,j) in the lattice.

Once the winner is found, the changes on its weights and its neighbours is calculated by

$$\mathbf{w}(t) = \mathbf{w}(t-1) + \eta * neigh(u,v,\sigma) * [\mathbf{x}(t) - \mathbf{w}(t-1)] \tag{2.9}$$

where $\mathbf{w}$ is the weight vector, $neigh(u,v,\sigma)$ is the neighbourhood function calculation centred on the neuron with coordinates (u,v), σ is a decaying variable applied directly to the neighbourhood function during the process, and η is the learning rate, also a decaying variable. This weight update is done until some sort of criteria is matched, be it that it reached the maximum number of iterations, or that the lattice has become stable enough.

After mapping all the data entries to each winner neuron, each of these neurons are now linked to a number of indexes from the input data. The SOM algorithm, while aiming at creating a visualisation, also happens to naturally create new clusters, or "micro"-clusters, of the data. Since SOM lattices can be quite large in the number of neurons, one may want the generalise these "micro"-clusters given by the algorithm to make it easier for analyses. A method to produce this generalisation is by finding the most similar neurons and to merge them into larger and possibly more meaningful clusters, and because neighbourhood neurons are related to each other, these clusters will tend to incorporate adjacent and contiguous neurons.

2.2.6 Hierarchical Clustering

Hierarchical or agglomerative clustering is a type of clustering process that generates as output not a fixed group of clusters, as the aforementioned algorithms, but an entire hierarchy of all input elements, represented in a dendrogram (or tree) format. In this generated dendrogram, more similar inputs will be closer together, both belonging to one common branch, and the closer a common branch between two points is from the bottom of the dendrogram (the inputs), the more similar the two points are. The hierarchy goes from input level – every input is treated a cluster – up to a final single cluster – which every single input is included into. In-between these extremes, the algorithm generates several different "levels" of clustering, allowing many different choices and numbers of "final" clusters.

The algorithm works by calculating the cross-dimensional distance between all the existing data points, then it merges the ones that are closer together. Once the two clusters are merged, they are still going to be compared to others in the next steps, but only as part of a newly formed cluster, and therefore the other points will be compared to both points at the same time. The method used to compare these clustered points to other points is referred to as the *linkage method*.

A common linkage method is the *Ward*'s method, variance minimisation distance algorithm. It is given by

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \tag{2.10}$$

where u is a cluster of formed by the two points s and t, v is the point being compared to the cluster, T is the sum of absolute values $|s|+|t|+|v|$, and the function d is the distance between the points. This distance d is not inflexible, users can choose the most fitting distance for their problem.

Once all the hierarchy has been built through Ward's method, the next step in analyses is deciding where to limit the number of clusters. The number of clusters needs to be large enough that it actually separates interesting blocks of data, and small enough that it can be humanly interpreted. This decision is a typical problem of hierarchical clustering, and it is referred as *branch cutting*, *tree cutting* or *dendrogram pruning*. Tree cutting is usually done by choosing a specific level, how many branch splits there are between a cluster and the root of the dendrogram, or a weight distance, also from the cluster to the root.

2.2.7 Dynamic Tree Cut

In order to cut a generated hierarchical tree and select the "final" clusters to use, one common method is to use the *fixed height branch cut*, where a single distance value is chosen as a fixed cutting point. Given a simple dendrogram, and given that there is strong understanding of what means to cut the tree at such height or level, this technique can achieve satisfactory results. For working with many dimensions and inputs, though, an alternative to avoid problematic and biased results is to use a more complex cutting algorithm, such as *Dynamic Tree cut* (Langfelder et al. 2007).

Dynamic Tree Cut is an iterative algorithm that analyses the dendrogram in a top-down manner, starting from the root, and walking step-by-step down the tree. The algorithm can be summarised as follows: (1) it starts from the step's initial clusters, (2) then it analyses each of these initial clusters individually, searching for breakpoints – points that separate two neighbouring clusters, (3) eventually, it splits these clusters into subclusters. This process is repeated for each subcluster that is discovered until no new subclusters are found by the method.

This cutting method reduces a considerable part of human decisions on tree cutting, as it can adapt to many different types of dendrogram/tree architectures. Few relevant human decisions are requested, facilitating automation and reuses for different data sources.

2.3 Computational Tools

In order to achieve the goals of this research, we made use of a few implementations of the aforementioned algorithms, designed by these projects: Scipy (Virtanen et al. 2020), Scikit-Learn (Pedregosa et al. 2011), Minisom (Vettigli 2018) and Dynamic Cut (Langfelder et al. 2007). The exact way how each of them was used within our workflow can be thoroughly observed in an online repository (Lopes Jr 2022). All the main scripts developed for this research were made publicly available in the repository, as well as some of the obtained datasets, including the one representing dataset $\mathbf{A}_0$.

Chapter 3

The *k*-Means Method

In this chapter, we present the first method used for the clustering analysis, the *k-means Method*. In Section 3.1, the entire method is described, with a step-by-step description of its architecture. Next, in Section 3.2, the findings of the method are shown, with numerical and visual depictions of the discovered clusters and their characteristics. Finally, in Section 3.3, a brief discussion of the method's results is presented.

3.1 Methodology

The methodology used for generating the final clusters followed the sequence of stages seen in the diagram of Figure 3.1. In which $\mathbf{A}_0$ was turned into a intermediate dataset $\mathbf{A}_{\text{RN}}$ through the P_3 preprocessing. Then, $\mathbf{A}_{\text{RN}}$ was processed by a clustering block called multiple *k*-means, identified here as *MkM*. *MkM* generates a centroids matrix, here referenced by the variable $\mathbf{C}_0$. $\mathbf{C}_0$ then went through the last preprocessing stage P_4, generating a new matrix $\mathbf{C}_{\text{RN}}$. Finally, $\mathbf{C}_{\text{RN}}$ was processed by the density-based spatial clustering of applications with noise (DBSCAN) algorithm, where the final clusters were revealed. The DBSCAN model is referenced here as *DBS*.

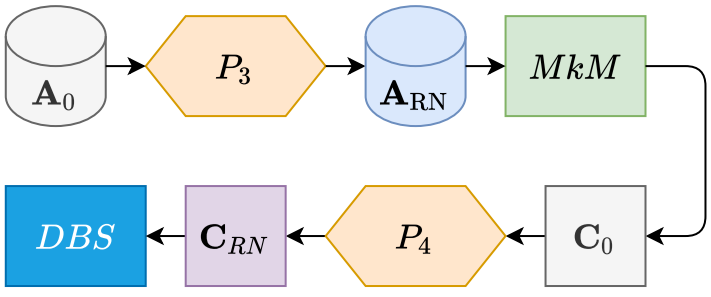


Figure 3.1: Clustering process sequence diagram, including both employed algorithms *MkM* and *DBS*, preprocessing stages P_3 and P_4, and intermediate datasets generated in each stage.

■ **Preprocessing P_3**

As shown in Figure 3.2, the P_3 preprocessing is characterised by removal of outliers, dimensionality reduction and normalisation. First, municipalities considered as PMR outliers, that is, those whose PMR values are three standard deviations above or below the national average, were removed from $\mathbf{A}_0$, generating a new dataset $\mathbf{A}_1$. The column that stores PMR values was removed from $\mathbf{A}_1$ and its data was stored separately for future uses. $\mathbf{A}_1$ is comprised of 103 features and considering the natural difficulty of optimising clusters in high dimensions (curse of dimensionality), principal component analysis (PCA) was used to reduce dimensions, generating a reduced dataset $\mathbf{A}_R$ that kept 95% of $\mathbf{A}_1$'s variance with 58 features. Next, A_R was normalised through the cascading application of 3 techniques: (1) Yeo-Johnson Transformation, making the original dimensions' distribution more normal distribution-like; (2) L2 sample normalisation, re-balancing samples individually to capture points of higher and lower impact in each; (3) and finally, 0 to 1 normalisation by feature. These normalisation techniques generate the k-means (MkM) input dataset $\mathbf{A}_{RN}$.

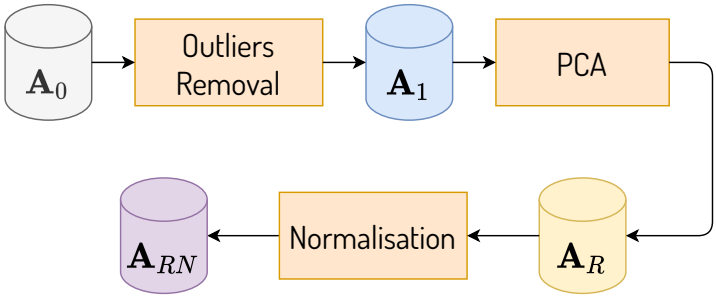


Figure 3.2: P_3 preprocessing diagram.

■ **Multiple k-means (MkM)**

In order to find representative centroids associated to our problem, a processing strategy here called MkM was proposed. MkM is characterised by a group of N k-means models that are all executed to the same input: $\mathbf{A}_{RN}$. Each ith k-means model, here called kM_i, is executed with a determined number of centres, Nc_i, randomly initialised. At the end, they generate a matrix expressed as

$$\mathbf{C}_0 = \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_N^T \end{bmatrix} \tag{3.1}$$

where each ith G_i is characterised by a set of Nc_i centres associated to each i-th kM_i model, and they are expressed as

$$\mathbf{G}_i = [\mathbf{c}_{i,1}, \ldots, \mathbf{c}_{i,Nc_i}] \tag{3.2}$$

where $\mathbf{c}_{i,j}$ is the j-th centre of the i-th kM_i model, and they are expressed as

$$\mathbf{c}_{i,j} = \left[c_{i,j,1}, \ldots, c_{i,j,H}\right] \tag{3.3}$$

where H represents the total number of features from the inputted dataset, which in this work equals $H = 58$, as mentioned in the past subsection. So, the matrix $\mathbf{C}_0$ can be rewritten as

$$\mathbf{C}_0 = \begin{bmatrix} \mathbf{c}_{1,1} \\ \vdots \\ \mathbf{c}_{1,Nc_1} \\ \vdots \\ \mathbf{c}_{N,1} \\ \vdots \\ \mathbf{c}_{N,Nc_N} \end{bmatrix} = \begin{bmatrix} c_{1,1,1} & \cdots & c_{1,1,H} \\ \vdots & \ddots & \vdots \\ c_{1,Nc_1,1} & \cdots & c_{1,Nc_1,H} \\ \vdots & \ddots & \vdots \\ c_{N,1,1} & \cdots & c_{N,1,H} \\ \vdots & \ddots & \vdots \\ c_{N,Nc_N,1} & \cdots & c_{N,Nc_N,H} \end{bmatrix}. \tag{3.4}$$

The number of rows in $\mathbf{C}_0$ can be determined as

$$L = \sum_{i=1}^{N} Nc_i. \tag{3.5}$$

For this work, $N = 290$ was used, that is, 290 k-means models were executed for $\mathbf{A}_{RN}$. The number of centres in k-means varied from 2 to 30, that is, $Nc_i \in \{2, \ldots, 30\}$ (29 different numbers), and for each number of centres 10 different instances of the model were run, adding up to $N = 290$ models. Each i-th model, kM_i, was optimised with the expectation-maximisation algorithm, the tolerance for convergence was set to 10^{-7} and the maximum number of iterations was $10,000$. The results from each of the $N = 290$ models were in a matrix $\mathbf{C}_0$ with $L = 4640$ samples, representing all the centres detected by the models.

Besides $\mathbf{C}_0$, a new intermediate dataset $\mathbf{A}_\mathrm{C}$ is created, comprised of $5,529$ samples and $N+1$ features. Each sample represents a municipality, and the features are the municipality's PMR and the clusters each sample is associated with for each n-th k-means model found in MkM. Thereby, each municipality, in each i-th row, belongs to a k-th cluster in each n-th column of $\mathbf{A}_\mathrm{C}$. Each n-th k-means model generated by MkM has N_c clusters grouping a set of B municipalities.

■ Preprocessing P_4

In preprocessing P_4, matrix $\mathbf{C}_0$ was filtered, keeping only the centres that represent clusters treated as "Clusters of Interest" (CoI), that is, those clusters whose mean PMR exceeds or is exceeded by the national average PMR in at least 10%. The data associated to each cluster is obtained through $\mathbf{A}_\mathrm{C}$ and the mean PMR can be calculated as:

$$TMP_\text{media} = \frac{1}{B}\sum_{i=1}^{B} TMP_i \tag{3.6}$$

where TMP_i is the PMR related to the i-th municipality and B is the total number of municipalities for a given cluster. For the calculation of the national PMR, the formula is used considering the cluster of municipalities to be one containing all the municipalities present in the datasets, and B to be the total number of municipalities, in this case $B = 5529$. After filtering $\mathbf{C}_0$, a new matrix is generated containing containing only the centres of the CoI, here referred as $\mathbf{C}_{ci}$.

In order to group the CoI, $\mathbf{C}_{ci}$, retrieved from the multiple k-means executions by the *MkM* block, a correlation matrix was generated from the samples of $\mathbf{C}_0$. The idea here is to work with the similarity between the centres - representing the CoI - to facilitate the clustering process in the following stages. The correlation matrix is characterised by the variable $\mathbf{C}_1$.

In order to reduce the dimensionality of the correlation matrix $\mathbf{C}_1$, the PCA algorithm was a applied to $\mathbf{C}_1$, generating a reduce matrix $\mathbf{C}_R$. $\mathbf{C}_R$ is composed of 4 columns, that maintain 99% of the original variance of $\mathbf{C}_1$. Next, $\mathbf{C}_R$ was normalised by column to keep values between 0 and 1, generating $\mathbf{C}_{RN}$. Figure 3.3 details the P_4 preprocessing.

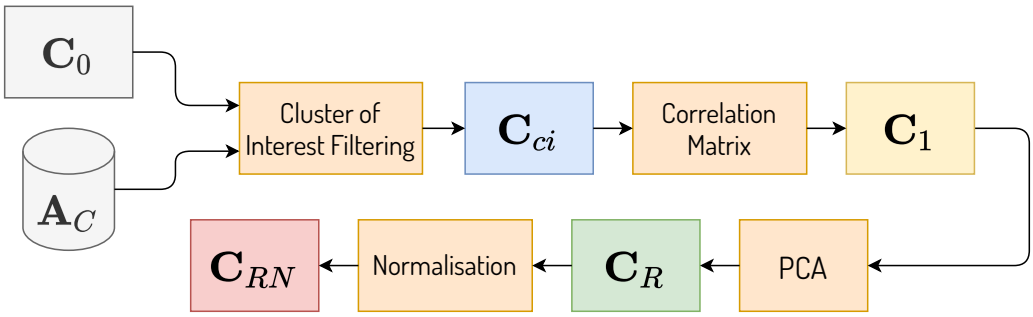


Figure 3.3: P_4 preprocessing diagram.

■ DBSCAN (*DBS*)

The DBSCAN receives $\mathbf{C}_{RN}$ as input, containing the sample correlation data. Unlike k-means, DBSCAN does not need a fixed target number of centres (clusters) to be set beforehand, its functionality is defined solely by the adjustments in two parameters: minimum euclidean distance between two points (ε) and minimum amount of points per cluster. As output, *DBS* generates a classification of the clusters discovered by the *MkM* from their respective centres, grouping together those so similar they can be treated as different instances of the same cluster, while also discarding centres without enough similar pairs and treating these as noise.

For *DBS*, the minimum number of occurrences was defined as 50, and $\varepsilon = 0{,}06$. As there isn't a well-defined or standard method to measure the quality of the validated clusters, data visualisation techniques were used to verify cluster consistency. Graphical representations were created to show the municipality-cluster association and also to visualise $\mathbf{C}_{RN}$ post-classification (2D plot with t-SNE). The parameters were adjusted in order to avoid too much regional fragmentation (clusters with no regional aspect, over-fitting) or empty regions (no classified municipalities in a large map area, under-fitting).

3.2 Results

After the application of P_1, P_2 and P_3, the *MkM* model generated a total of 1,337 CoI. The number of detected CoI grew according to the total numbers of clusters defined in the k-Means models, Nc_i, as it can be seen in Figure 3.4. It also shows that the first cases of CoI appear when the k-means input number of centres, Nc_i, equals 6, reaching about 90 CoIs for the highest number of Nc_i (27 to 30).

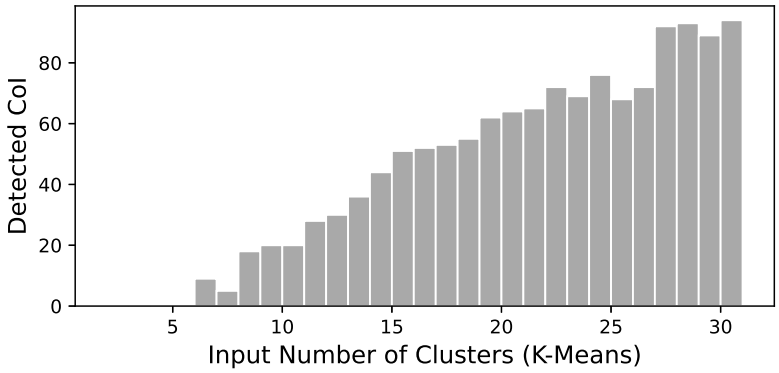


Figure 3.4: Clusters discovered by *MkM* by input number of clusters.

The correlation matrix, $\mathbf{C}_1$, and its reordered version can be seen, respectively, on items (a) and (b) of Figure 3.5. It's possible to observe some cluster patterns from the distance-based reordering alone. On item (c) of Figure 3.5, the final clusters for each sample are highlighted in different colours, allowing a visual comparison between the *DBS* output and the sample distance algorithm.

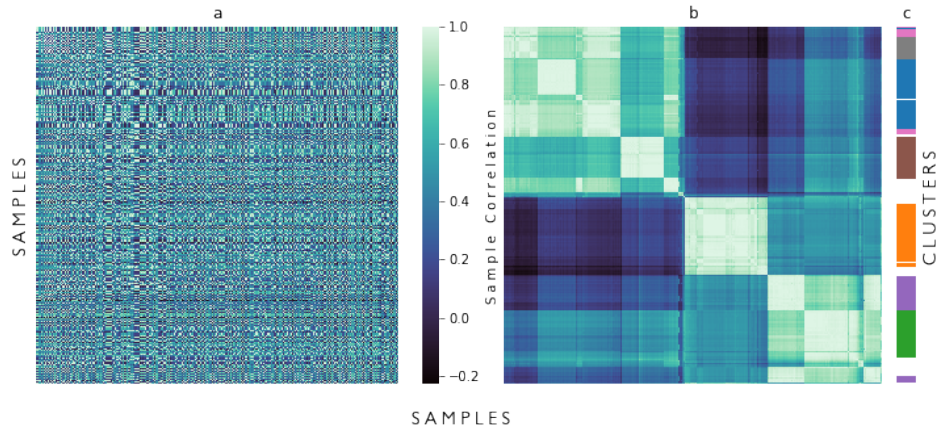


Figure 3.5: (a) Correlation matrix (b) Correlation Matrix reordered by distance between samples (c) Classification of reordered samples after DBSCAN

After applying the P_4 preprocessing steps, the final clustering was done by using *DBS*. *DBS* found 7 final clusters, divided into 4 clusters with high PMR (PTB Municipality

Rate) and 3 clusters with low PMR. On item (c) of Figure 3.5, it's shown how some of the rows of the correlation matrix were not selected to any final cluster. On item (a) of Figure 3.6, it's possible to observe a stagnation or even a reduction in the identification of valid clusters for the highest input number of centres in comparison with median values. On item (b), the same clusters are shown, but now separated not by high and low PMR, but for individual final clusters.

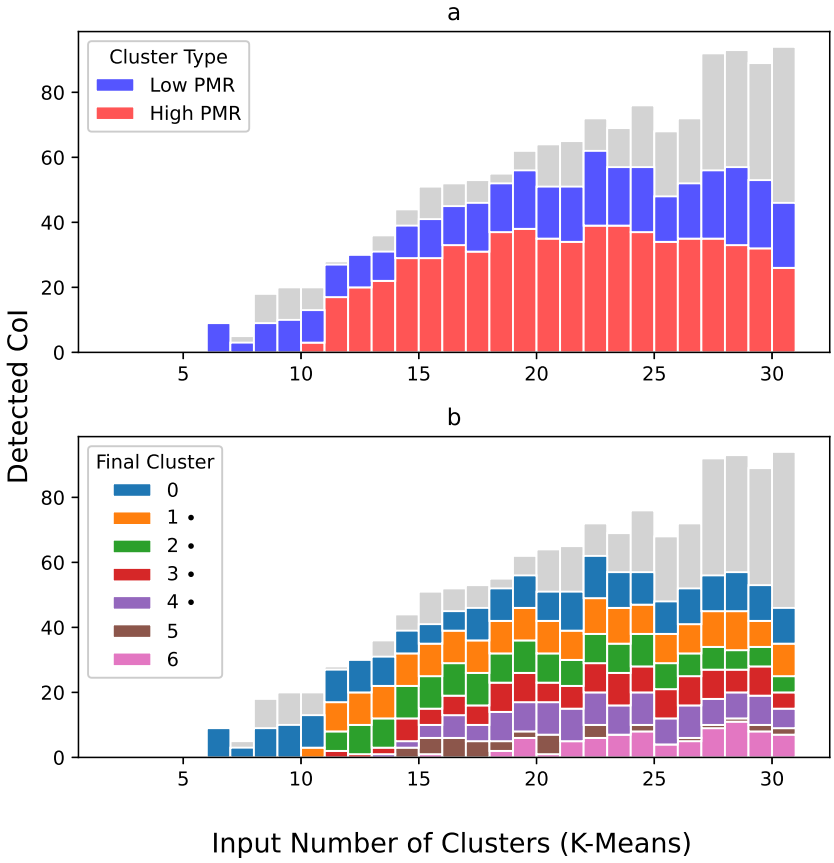


Figure 3.6: Clusters found per epoch in *MkM*. (a) by cluster type (b) by final cluster. Symbol (•) indicates high PMR cluster

To better visualise the action of *DBS*, we created a t-SNE 2D plot of the *MkM*-generated cluster centres, as seen in Figure 3.7. The t-SNE, similarly to PCA, reduces the dimension of the data, creating a 2D representation that tries to keep the 2D distances between points proportional to those of the hyperdimensional dataset. On item (a) of Figure 3.7, we observe how closer centres were indeed grouped in the same CoI. By comparing item (a) and item (b), we can also see how most points in the same grouping seem to present near values of PMR.

The CoI's PMR distribution for each final cluster was calculated. This distribution can be observed in Figure 3.8, where each cluster is represented on the x-axis, the distributions on the y-axis and the national average PMR is indicated visually (aprox. 1.4×10^{-4}). It's shown that almost all validated clusters have their centroid PMR varying from 1×10^{-4}

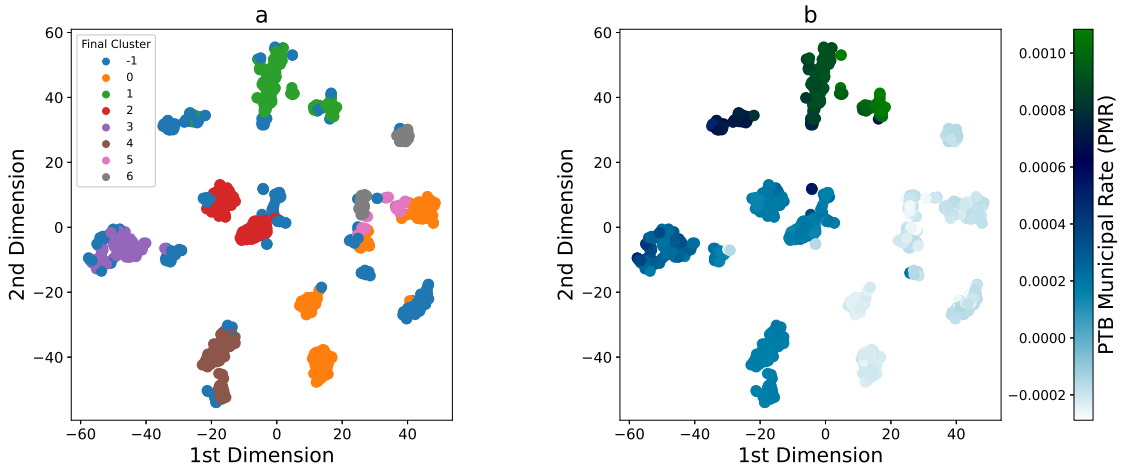


Figure 3.7: t-SNE visualisation of *DBS* Clustering effect. (a) shows how each *MkM* cluster centre was classified by *DBS*. (b) shows the same centres by PMR for comparison. Final Cluster (-1) indicates clusters not grouped in any final cluster.

to $2{,}5 \times 10^{-4}$ in comparison to the national average, with the exception of Cluster 1, with a centroid PMR almost 8×10^{-4} units above the average.

The regional distribution of these clusters was also observed, that is, which municipalities belong to which cluster. Through data visualisation it's possible to contextualise - as well as to validate - the discovered clusters. Since the input of the problem is social data, it was expected for at least some of the clusters to be located in socially similar concentrated areas. Three visualisations were generated to verify that.

The first visualisation is shown in Figure 3.9, it's a binary plot generated using the type of cluster (high or low PMR). The amount of times each municipality was classified into a validated high or low PMR cluster was counted, and each municipality was marked with the type it was mostly classified as. White-coloured municipalities were never classified in a CoI.

The second visualisation was generated from the subtraction of the total amount of times in which a municipality was classified as high PMR minus the total amount of times in which it was classified as low PMR, obtaining a type of degree of intensity or belonging of each municipality to the types of clustering, and it can be seen in Figure 3.10.

The third visualisation, seen in Figure 3.11, reveals in which of the 7 final clusters each municipality was mostly classified by the *MkM* models, making it possible to visualise the regional aspects of the clusters. Clusters 1, 2, 3 and 6 appear to be more concentrated in specific regions of the map, while 0, 4 and 5 have a more sparse distribution.

Looking at Figures 3.9 and 3.10, it's possible to see a clear regional aspect not only for the individual clusters, but also for the types of clusters, with High PMR clusters located mostly in the North and Northeast regions, and the Low PMR clusters in the Centre-South area. In the Northeast, High PMR clusters are concentrated in the state of Maranhão and across the São Francisco River valley. The most intense Low PMR clusters are seen in the state of São Paulo and in Southern Minas Gerais. The North region is almost entirely

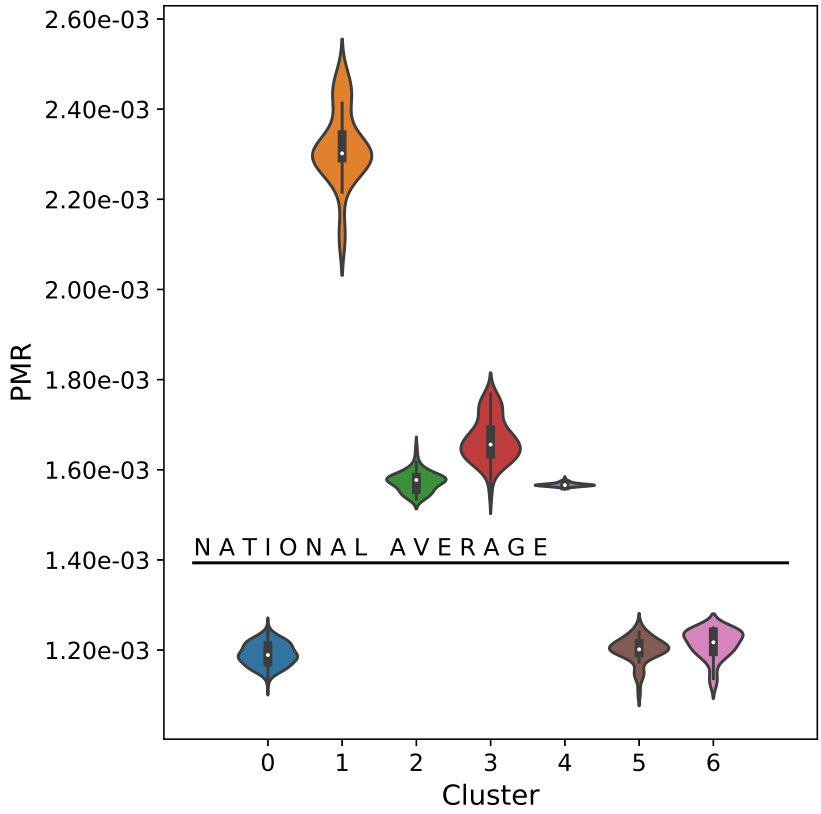


Figure 3.8: PMR distribution of final clusters.

classified in clusters of High PMR and, as it is shown in Figure 3.11, the most frequently observed cluster in the region is Cluster 1, notably the one with highest PMR.

In order to measure how the clusters are differing from one another, T-tests were performed to measure the p-value of each variable. Two additional sets, treated here as clusters, were created for comparison, N, containing all municipalities that weren't grouped in any of the final 7 clusters, and A, containing all municipalities, regardless of clustering.

The p-value was calculated for every variable and for every pair of clusters. The comparison of a cluster to itself was done by generating two random sub-samples of the cluster and testing them against each other. After every p-value was determined, the percentage of variables with a p-value above the 5% threshold for every pair of cluster was calculated and is shown in Figure 3.12. It is possible to observe how clusters are significantly distinct from each other through most variables. The closest similarity was observed between clusters 5 and 6, and between clusters 0 and 5 (only 37% and 50% of variables significantly different, respectively).

Also, in order to get a general view of how High PMR and Low PMR compare to each other on different aspects of SES, the features used for clustering were categorised in seven segments: Sanitation, Employment, Living Conditions, Education, Household Type, Race and Income. Then, a subset of the data was created for each segment, con-

taining all municipalities, their assigned clusters and only the features of the respective segment. Dimensionality was reduced using t-SNE for visualisation purposes to create 2D maps of the subsets, and a SVM-RBF classifier was applied to the t-SNE maps to find the boundaries in the generated space that best separates High PMR and Low PMR clusters. The t-SNE outcome and the boundaries can be seen in Figure 3.13. The first (upper) plot for each segment contains only the Low PMR cluster points, the second (lower) also shows the High PMR cluster points and the separation boundaries. It's possible to see, even without reaching the most easily understandable feature-level view, how High and Low PMR clusters follow distinct patterns SES-wise. Some segments, such as Sanitation and Living Conditions, show High PMR cluster points very well-grouped, and when comparing upper and lower plots, it's almost as if the High PMR points filled an empty space the lower plot. Others, such as Education and Income, show High and Low PMR cluster points more mixed up, but with High PMR points centred in a smaller area.

In addition, the core of each validated cluster was extracted, containing information about the mean and variance observed for each of the 7. With that information, it is possible to view the detailed features of each cluster.

It's possible, by checking the individual characteristics of each cluster, to see the relationship between SES factors used as input and the PMR. Figure 3.14 shows the percentage difference between each cluster and the national average for some of those characteristics: higher education, race, water supply, garbage destination, sewage access and number of rooms in residence. It's noticeable how there is a clear contrast between High PMR and Low PMR clusters among these characteristics.

Cluster 4 is notable for being the only cluster that does not strongly respect this contrast. In Figure 3.8, Cluster 4 is shown as the one with lowest PMR among those with PMR above average. And in Figure 3.11, it is visible how 4 is the most disperse among the High PMR type clusters, with a noticeable amount of coastal municipalities both in the Northeast and in the state of Rio de Janeiro. In contract, Clusters 1, 2 and 3, of higher PMR, are concentrated in the North region and in the Northeast region's countryside.

Finally, returning to the general perspective of High PMR vs Low PMR, we measured the mean values for the normalised (0 to 1) features of all clusters discovered by the *k*-Means algorithm to generate a general High PMR and Low PMR centre. That way, it becomes possible to verify which variables diverge the most. The results are shown in Table 3.1, ranked by the normalised distance. It is possible to see that most variables show a small difference in the space, which is normal considering we merged distinct clusters with different characteristics.

It is still possible, though, to see some clearly strong variables both for Low and for High PMR. Low PMR clusters seem especially superior in Sanitation and Living Conditions, and the highest normalised difference was obtained in the Sewage System feature. There is also a clear racial division, with High PMR clusters having a larger number of whites, and Low PMR having a larger number of pardos.

3.3 Discussion

In this work, unsupervised learning techniques, a two-level clustering, was used two discover clusters of High and Low preterm birth (PTB) rate among Brazilian municipalities while clustering only for SES factors. The clustering resulted in 7 final clusters, 4 with a High PTB Rate and 3 with a Low PTB Rate, and found significant socioeconomic difference between these High and Low PTB Rate clusters. Medical and health sciences extensively use data, specially biological data, to tackle daily problems. Preterm birth, despite much research, is still not totally comprehended, but studies suggest the influence of external factors, including SES factors. By finding SES neighbourhoods that are more suitable for the occurrence of PTB, the health system may be able adjust itself better, and earlier, in order to provide assistance to the maximum number of newborns.

The two-level clustering method described, *MkM* followed by *DBS*, allows k-Means clustering in contexts usually not covered by traditional "optimal number of clusters" techniques. By setting a initially designated cluster target rule, k-Means can be used to track down specific sorts of clusters, guaranteeing the significance of the found cluster(s) through recurrence and DBSCAN validation, while also maintaining the algorithm's explainability factor for posterior analyses. Using this method we were able to identify 7 distinct clusters of notably outlying PTB Rate (10% threshold) as well as how strongly each municipality is associated to those clusters, and how different these clusters are amongst each other at feature-level, segment-level and overall. The two-level clustering validation behaved as expected, selecting similar cluster centres and discarding the noise generated mostly from the "over-fitting" high number of cluster in some *MkM* units. The validated final clusters, even if chosen only by their PTB Rate, were shown to be significantly distinguishable in most of the SES factors used in the process. The clustering, working in a PCA-reduced hyperspace, was also able to find clusters that are shown to be distinct even in specific SES segments such as sanitation and living conditions. And since neighbouring municipalities tend to be more socially-alike to each other than to further away municipalities, the more regionally concentrated clusters found here are another previously expected outcome.

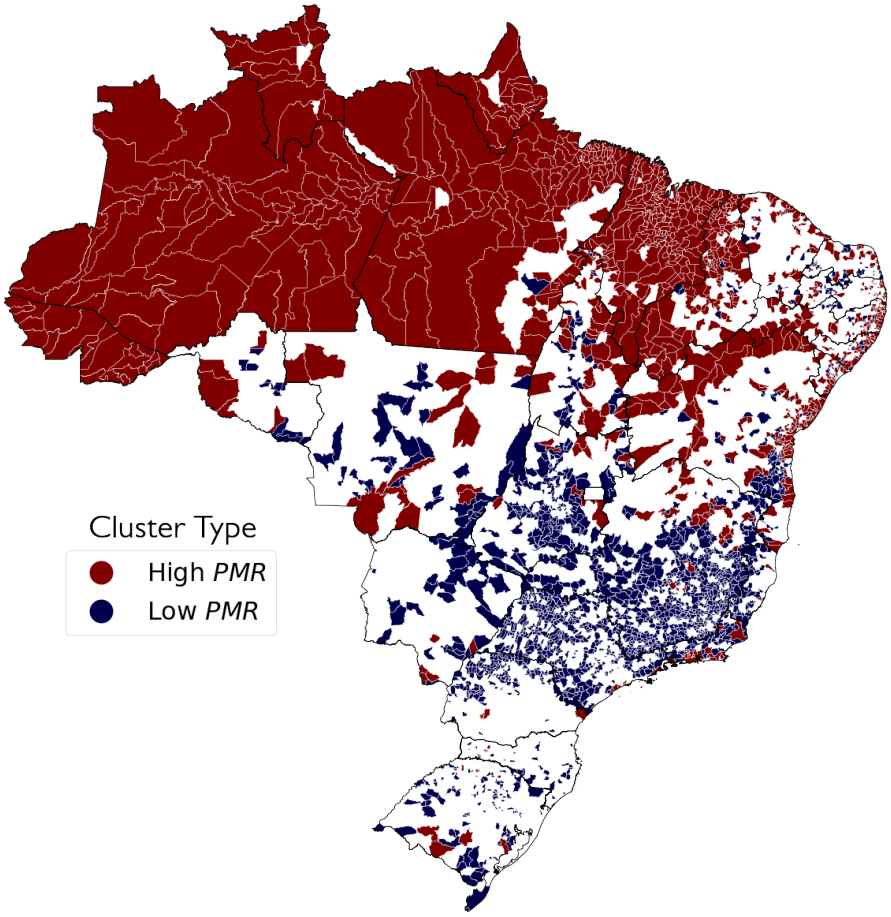


Figure 3.9: Municipalities by most common type of cluster (high or low PMR). White-coloured municipalities were not classified in a Cluster of Interest.

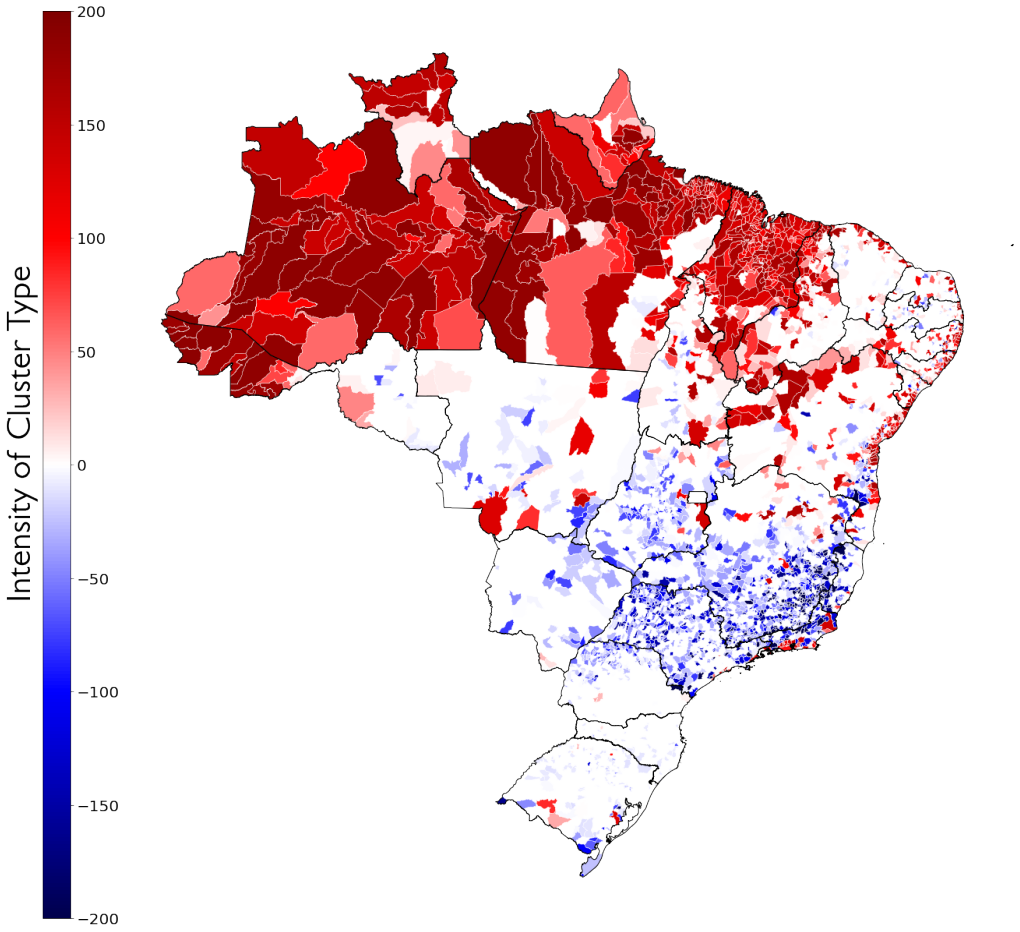


Figure 3.10: Municipalities by difference of number of times they were classified as each cluster type. Negative values (■ blue) indicate that a municipality was mostly classified in low PMR clusters, positive values (■ red) indicate it was mostly classified in high PMR clusters.

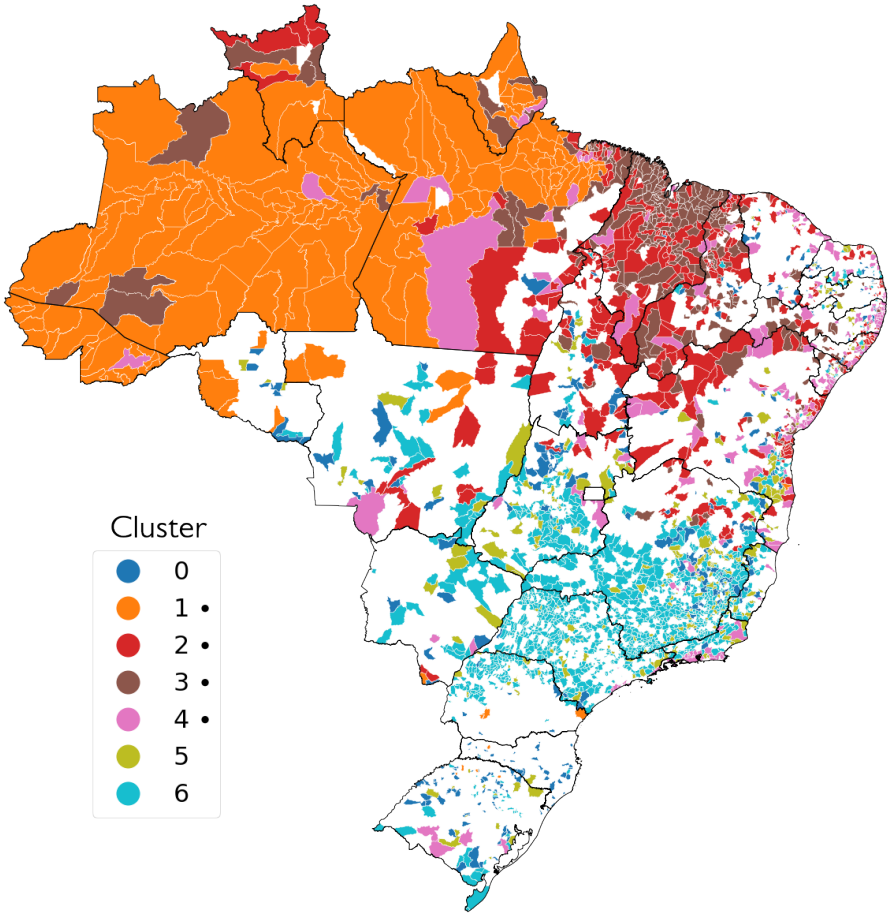


Figure 3.11: Municipalities by most common cluster. Symbol (•) means high PMR cluster. White-coloured municipalities were not classified in a Cluster of Interest.

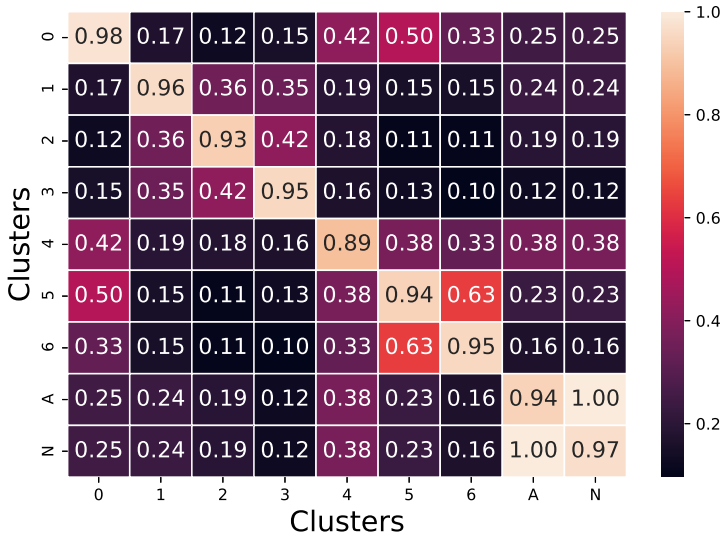


Figure 3.12: Percentage of variables with a p-value under 1%, when comparing the pairs of clusters via T-test.

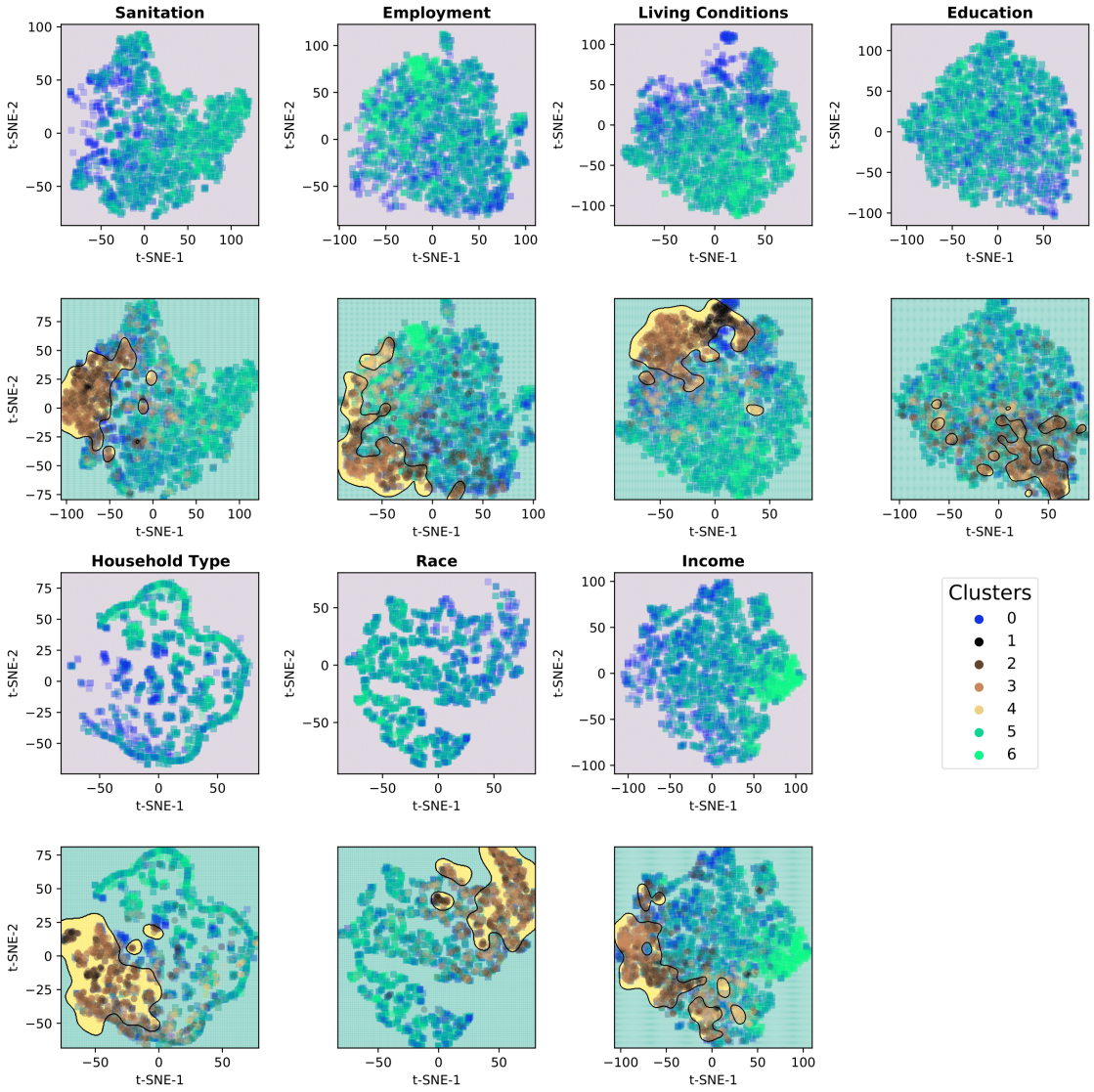
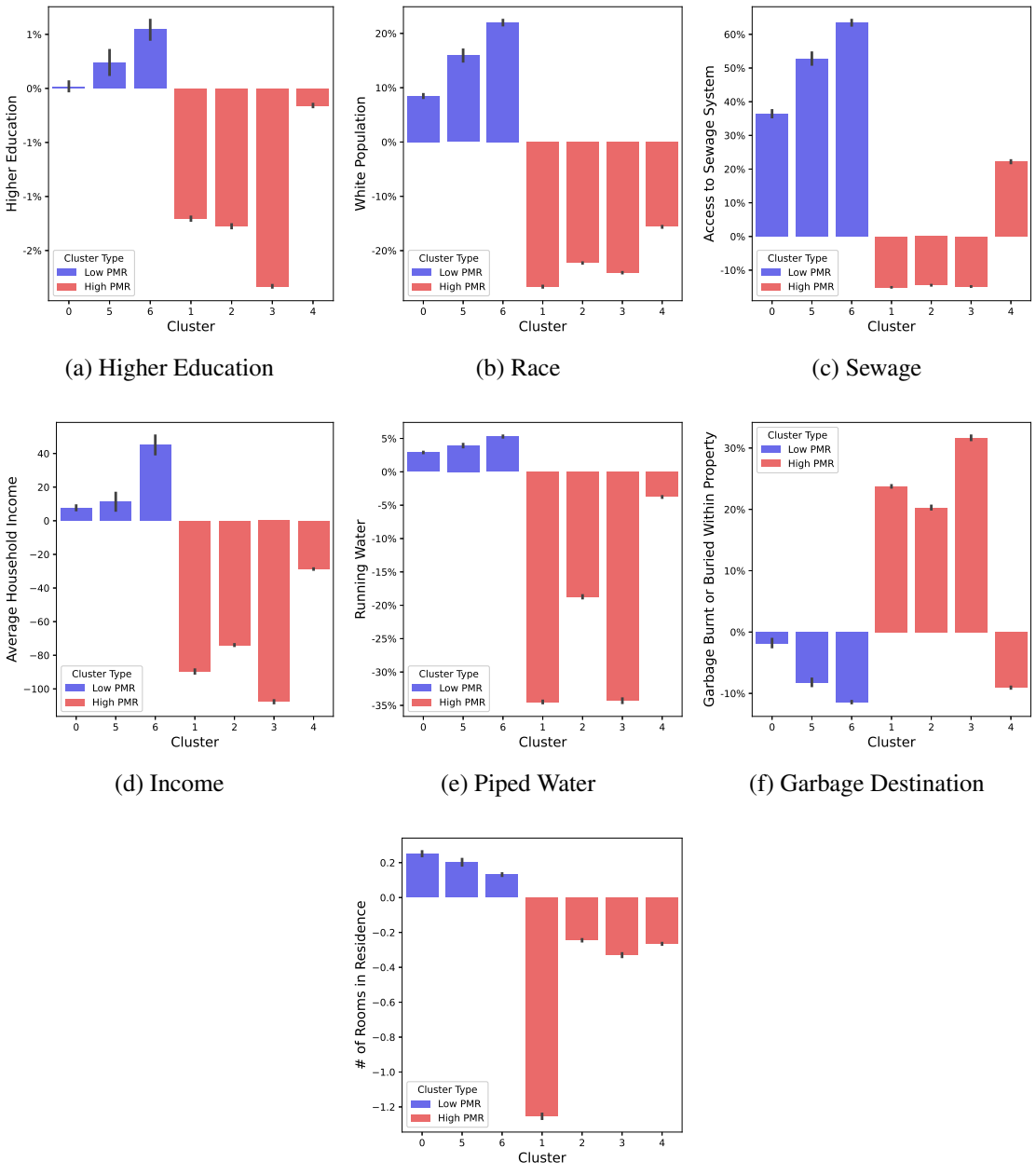


Figure 3.13: t-SNE 2D representation of final clusters by SES variable type, including SVM-RBF boundaries for type of cluster.



(a) Higher Education (b) Race (c) Sewage

(d) Income (e) Piped Water (f) Garbage Destination

(g) Rooms

Figure 3.14: Final clusters' characteristics.

Table 3.1: Comparison between Low PMR and High PMR clusters, using the Normalised Distance (ND) between the mean values across all clusters. ■ - Living Conditions, ■ - Race, ■ - Sanitation, ■ - Education, ■ - Working Conditions, ■ - Income, ■ - Household Type

		Higher On Low PMR			**Higher On High PMR**	
P	T	Feature	ND	T	Feature	ND
1	■	Sewage System	0.513	■	Pardo	0.386
2	■	Coated Bricks House	0.467	■	Employed in Agriculture	0.292
3	■	Ceramic Tile/Stone Floor	0.409	■	No Pavement	0.248
4	■	White	0.350	■	Bolsa Família-assisted Family	0.240
5	■	Full Paving	0.339	■	Rubbish Burnt or Buried	0.223
6	■	Direct Rubbish Collection	0.303	■	Cesspit	0.219
7	■	Piped Water	0.276	■	Wooden House - Proper Wood	0.179
8	■	Household with Bathroom	0.211	■	Dirt Floor	0.128
9	■	Water Supply Network	0.190	■	Concrete House	0.121
10	■	Rooms per Household	0.156	■	Months Employed	0.099
11	■	Urban	0.147	■	Coated Rammed Earth House	0.089
12	■	Individual Electricity Meter	0.131	■	Wooden Floor - Proper Wood	0.089
13	■	Avg Household Income	0.100	■	Employed	0.085
14	■	Community Electricity Meter	0.081	■	No Electricity Meter	0.081
15	■	Literate	0.065	■	Self-Employed	0.078
16	■	Attended School	0.065	■	Uncoated Brick House	0.077
17	■	# of Rooms for Sleeping	0.055	■	Any Job Last 12 Months	0.072
18		Age	0.052	■	Septic Tank	0.072
19	■	Personal Income	0.049	■	Well as Water Source	0.068
20	■	Preschool	0.046	■	Uncoated Rammed Earth House	0.065
21	■	Literacy-level at School	0.046	■	Other Water Sources	0.060
22	■	Primary/Middle School	0.042	■	Rubbish Dumped on Wasteland	0.058
23	■	Black	0.040	■	Never Attended School	0.057
24	■	Regular Employment	0.039	■	Farmer (*Type*)	0.050
25	■	Wastewater to River/Lake/Sea	0.038	■	People in Dwelling	0.050
26		Disability	0.037	■	Attended Public School	0.048
27	■	Salary	0.031	■	Oil/Gas/Kerosene Illumination	0.039
28	■	High School	0.029	■	Rural	0.038
29	■	Private Residence	0.026	■	Adult Literacy Program	0.033
30	■	Higher Education	0.026	■	Incomplete Pavement	0.029

Chapter 4

The SOM Method

In this chapter, we present the second method used for the clustering analysis, the *SOM Method*. In Section 4.1, the method is described, with a detailed description of its architecture and stages. Next, in Section 4.2, the results of the method are depicted, with numerous visualisations of the final clusters, their characteristics and locations, as well as of some of the aspects of the SOM that are relevant to the problem. Finally, in Section 4.3, a brief discussion of the method's results is presented.

4.1 Methodology

The methodology used for generating the final clusters followed the sequence of stages seen in the diagram of Figure 4.1. In which $\mathbf{A}_0$ was processed by Preprocessing P_3, generating two outputs: $\mathbf{A}_N$ and $\mathbf{A}_{RN}$, both were then fed to multiple Self Organising Maps with varying parameters, generating the final clusters for analysis.

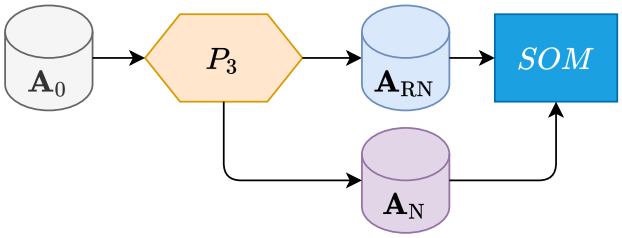


Figure 4.1: Clustering process sequence diagram, including both employed algorithms *MkM* and *DBS*, preprocessing stages P_3 and P_4, and intermediate *datasets* generated in each stage.

■ Preprocessing P_3

As shown in Figure 4.2, the P_3 preprocessing is characterised by removal of outliers and normalisation. First, municipalities considered as PMR outliers, that is, those whose PMR values are three standard deviations above or below the national average, were removed from $\mathbf{A}_0$, generating a new dataset $\mathbf{A}_1$. The column that stores PMR values was

removed from $\mathbf{A}_1$ and its data was stored separately for future uses. Next, A_R was normalised through the cascading application of 3 techniques: (1) Yeo-Johnson Transformation, making the original dimensions' distribution more normal distribution-like; (2) L2 sample normalisation, re-balancing samples individually to capture points of higher and lower impact in each; (3) and finally, 0 to 1 normalisation by feature. These normalisation techniques generate the *k*-means (*MkM*) input *dataset* $\mathbf{A}_N$.

Besides, a copy of $\mathbf{A}_1$ was reduced using PCA, keeping 95% of the original variance. This copy was also normalised using the same procedures as the original dataset, generating $\mathbf{A}_{RN}$. $\mathbf{A}_N$, the complete dataset, and $\mathbf{A}_{RN}$, the PCA reduced dataset will both be used in the clustering process.

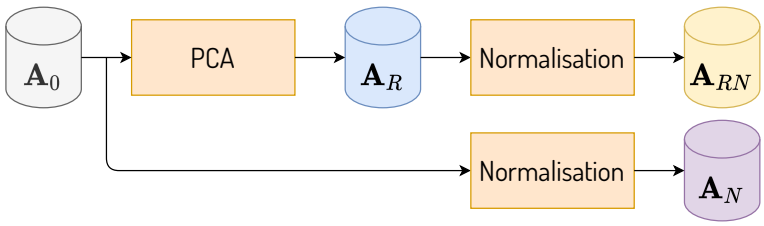


Figure 4.2: P_3 preprocessing diagram.

■ SOM

SOMs' results are dependent on a series of parameters to be chosen by the user. These include (1) Activation distance, (2) Neighbourhood function, (3) Grid topology, (4) Constant σ, (5) Learning rate, and (6) Number of maximum epochs.

Three of these parameters remained fixed: learning rate (0.1), number of maximum epochs (100,000) and grid topology (*rectangular*). We kept the topology fixed in order to facilitate visual comparison and reduce overall work, as some internal aspects of the network, neighbourhood functions, in particular, may need distinct calculations for different topologies. The learning rate, on the other hand, remained fixed because tests done with a rate of 0.01 resulted in overwhelmingly long training, commonly not filling the entire grid during the process, and tests done with a rate of 1 seemed rather close to those of 0.1, but slightly more irregular. We ran tests for more than one value for each remaining parameter.

Once each combination's training finished and each input municipality was assigned, by the algorithm, to one of the 400 SOM neurons, we then used hierarchical clustering to unite those neuron clusters into major clusters. SOM neurons were linked through Ward distance, generating a hierarchical tree of cells. Then, we used Dynamic Tree Cut algorithm to automatically and iteratively find the optimal cutting level at the generated tree. The number of major clusters differed depending on the SOM parameters.

We trained each possible combination of parameters using our SOM network, and Quantization Error (at SOM neuron-level) and Davies-Bouldin Score (at final cluster-level) were used to compare and evaluate each combination.

4.2 Results

Once the data was assembled, and both the complete dataset and the PCA reduced dataset were ready, they were both clustered in 36 different configurations, as shown in Table 4.1. For each configuration, we calculated the Quantization Error and the Davies-Bouldin Score. As Table 4.1 also shows, configurations using Manhattan distance had some difficulty at clustering the data, generating a much smaller number of clusters in comparison with Euclidean and Cosine configurations.

Each of these configurations was used to train the SOM network and create SOM maps of Brazilian municipalities' socioeconomic aspects. SOM generates maps such as the ones observed in Figures 4.3, 4.4, 4.5, and 4.6. There we can see both the neurons as they were classified after the Dynamic Cut algorithm was applied, and the SOM distance map (in black and white), where neurons are painted according to the overall distance between them and their neighbouring neurons. By comparing the two views in each figure, the final clusters and the distance map, side by side, it's perceptible how high distance regions are located in boundary regions between two final clusters, which is an expected outcome, as high distance means neurons are considerably different feature-wise and therefore are not good candidates for being in the same cluster. But since the overall distance map deals simultaneously with all dimensions (dataset features), some considerable differences seen across a more limited number of features, enough to classify neighbouring neurons into separate clusters, will not appear large in the distance map. Still, by paying close attention to the cluster borders in the distance map, those regions are generally darker than the cluster centres.

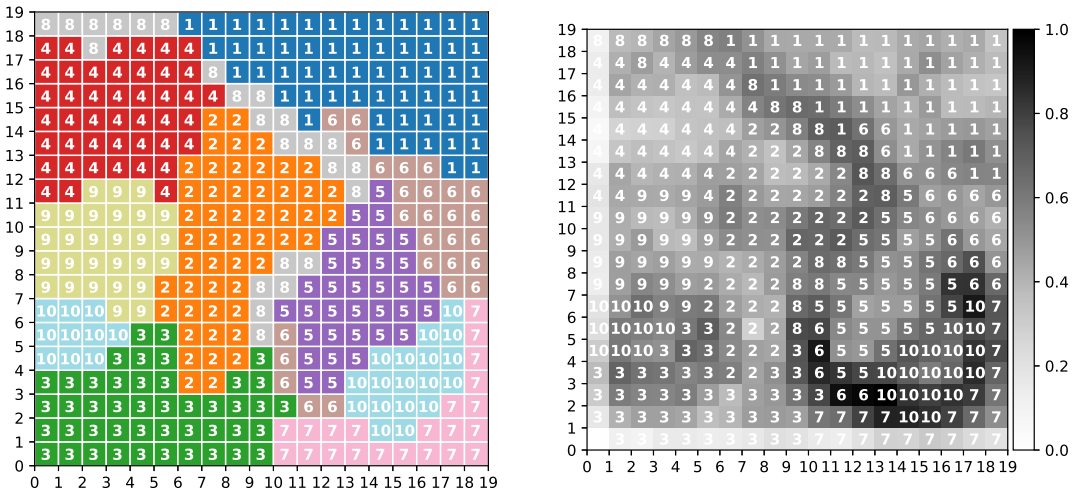


Figure 4.3: Clusters and Distance Map of SOM neurons. Cosine distance, Triangle neighbourhood function, $\sigma = 4$, Complete dataset.

A characteristic of the Dynamic Cut that was mentioned is how it can deal with non-uniformly spread hierarchies, without having to select a fixed cut-height or cut-level. In Figure 4.7, one of the dendrograms of the posterior hierarchical clustering is shown, cut

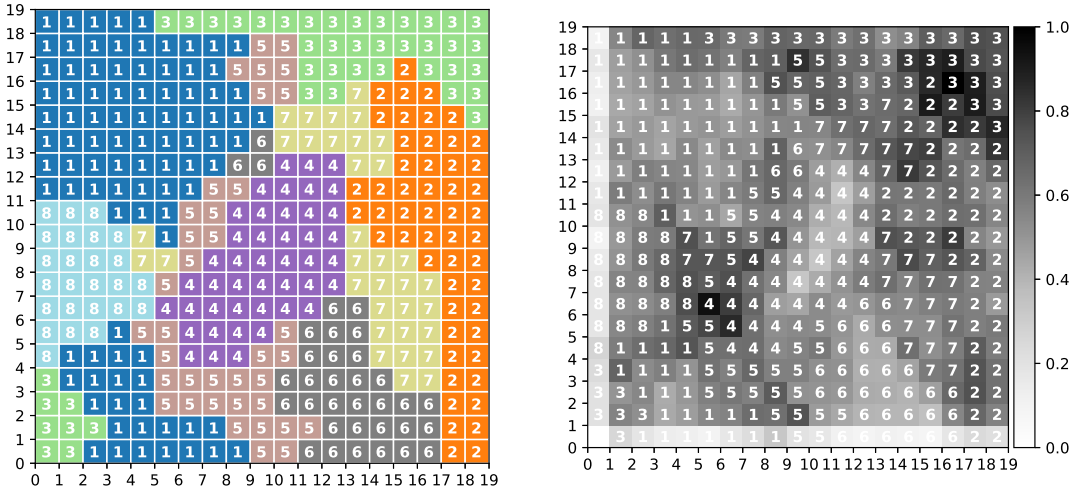


Figure 4.4: Clusters and Distance Map of SOM neurons. Euclidean distance, Triangle neighbourhood function, $\sigma = 3$, PCA reduced dataset.

at level 5. It's possible to observe how branches are uneven level-wise. This diversity of arrangements among the branches is seen throughout all the SOM instances and signals the complexity of selecting a single fixed cut value, especially one for all SOM maps generated.

In order to examine the regional aspect of the final clusters, we combined the results of High PMR and Low PMR clusters and created a map representation of high and low PMR, as shown in Figure 4.8. In these maps, two thresholds were used to define a cluster as High PMR or Low PMR, 5% and 10% PMR difference in comparison to the national average, and clusters were also split according to the dataset used. It's clear in 4.8 how High PMR clusters tend to be located in the North and Northeast regions, while Low PMR clusters are concentrated in the Southeast and South. The Centre-West is the only region without a clearly observable tendency towards high or low values of PMR. These points stand valid on both the complete and the PCA reduced datasets.

Another interesting view is created by segmenting the clusters by neighbourhood function and σ value, as shows in Figure 4.9. There, we can see how different configurations generate different outcomes, but also how the regional characteristics are kept regardless of the changes, with High PMR clusters located in the North-Northeast, and Low PMR in the South-Southeast regions. It's also shown how High PMR clusters are more easily recognised by the algorithm, with some configurations, in particular [*Complete, Gaussian,* $\sigma = 3$], [*PCA Reduced, Gaussian,* $\sigma = 3$] and [*PCA Reduced, Bubble,* $\sigma = 4$] being barely able to recognise any Low PMR cluster, while High PMR clusters, specially in the North region, were easily discovered by every single configuration shown. When comparing results of the complete dataset with the results of the PCA-reduced dataset, the latter seems to more regularly find Low PMR clusters around the Centre-West region and High PMR clusters located in Northeastern coastal municipalities.

Another form of observing the regional disparity over cluster types is by checking how

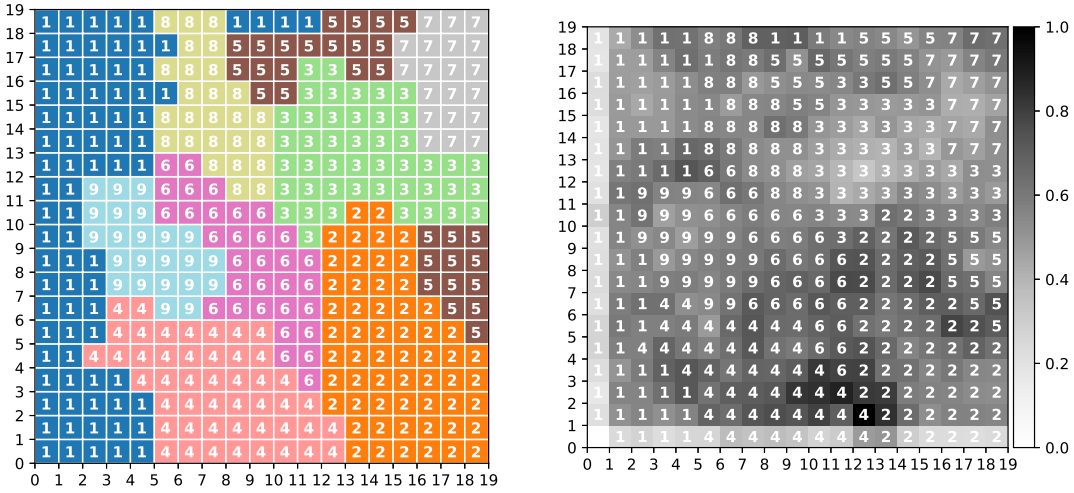


Figure 4.5: Clusters and Distance Map of SOM neurons. Euclidean distance, Triangle neighbourhood function, $\sigma = 4$, Complete dataset.

much of each region is included in clusters of High and Low PMR. For each configuration, the percentage of a region's municipalities that belonged to these two cluster types was calculated. The mean over the 36 configurations for both the complete and PCA reduced datasets can be seen in Figure 4.10 (High PMR clusters) and in Figure 4.11 (Low PMR clusters). In Figure 4.10, it's clear how high PMR is exceptionally common in the North, with Northeast and Centre-West regions having a significant number of occurrences of high PMR, but nowhere near as much as the North. In Figure 4.11, it's noticeable how the Centre-West and the South municipalities have a higher presence in Low PMR clusters, with the Centre-West having a considerably high presence of around 25% for the PCA reduced dataset. The Southeast region has a noticeable presence of around 10%, much higher than it has on High PMR clusters, and the Northeast region has the lowest presence in Low PMR clusters.

It's also noticeable in Figures 4.10 and 4.11, by comparing the results of the complete dataset and the PCA reduced one, how even if exact values may differ, the same tendencies are observed for both datasets.

The final clusters found by each SOM configuration also show a tendency of being concentrated in specific geographical regions, which is an expected outcome, since nearby municipalities tend to be socially close to each other. Two examples of this behaviour can be observed in Figure 4.12: in the lower plot, it's noticeable that cluster 2 is mainly concentrated in the South region, clusters 4, 6 and 8 in the Southeast region, 1 in the Northeast, and 7 in the Centre-West. Cluster 5 is concentrated in the Northeast and Southeast, and cluster 3, even if spread across the whole territory, clearly tends towards North municipalities. In the upper plot, even with a different configuration and a different number of final clusters, the general regional behaviour is kept. This is true for every configuration tested, even configurations that resulted in only 2 final clusters, which basically divided the country into North-Northeast and Centre-South.

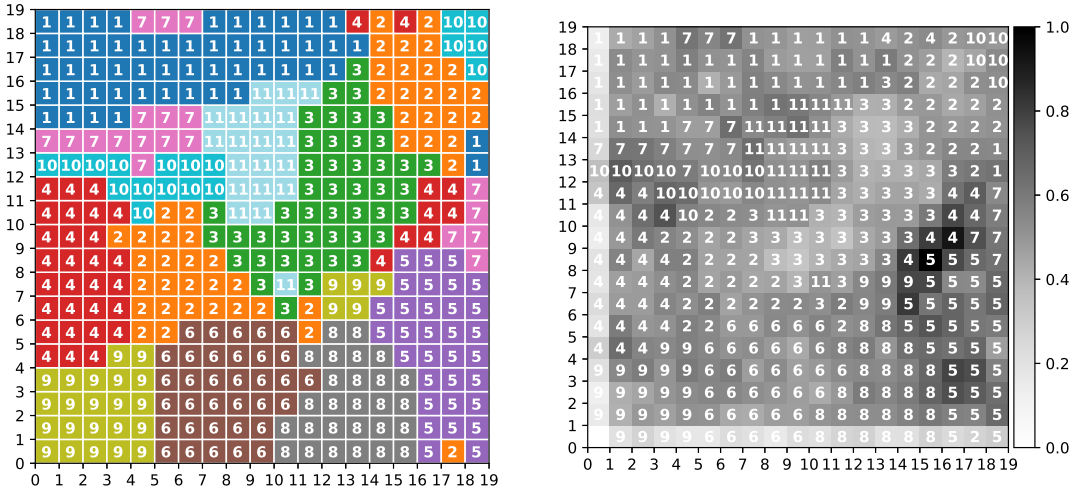


Figure 4.6: Clusters and Distance Map of SOM neurons. Cosine distance, Bubble neighbourhood function, $\sigma = 2$, PCA Reduced dataset.

It's also perceptible, looking at Figure 4.12, how the largest metro areas in the North and Northeast are an exception to the regional grouping, metro areas, such as Manaus, Recife, Fortaleza, Teresina, Natal, Juazeiro-Petrolina and others, are commonly classified in clusters centred in the Southeast region, as it was the case for both examples shown in Figure 4.12.

Finally, to properly understand what are the main characteristics of High PMR and Low PMR municipalities, we recovered the original SOM neurons of all configurations, and separated those with most extreme values of mean PMR. Only neurons with at least 5 data points (municipalities) classified in it were considered. For High PMR, neurons with PMR with 3 standard deviations away from the mean were used, while for Low PMR, because values are generally closer to the mean, this same rule would result in a too small sample of neurons, so 2 standard deviations of difference (negative) were used. This resulted in 229 total SOM neurons of high PMR, and 219 SOM neurons of low PMR. The features' means over these two sets of neurons was calculated and compared to each other. Normalised values were used, so all the features can be compared in the same value range (0 to 1). The results can be observed in Table 4.2, where the left half of the table shows the characteristics that were larger, in value, on high PMR clusters, while the right half shows the characteristics larger on low PMR clusters. From the results, it's clear that the highest PMR neurons are composed of municipalities with lower income, lack of access to proper public services and sanitation, and less educated. In contrast, lowest PMR neurons are composed of municipalities with higher income, better education, better sanitation, and also with more regularly employed people or people employed by the government.

It's important to notice that, since the values were normalised, a high value of one feature should not be mistaken by a high presence of that feature in the neuron/cluster, if *Oil/Gas/Kerosene Illumination* has a difference higher than 0.9, that means higher PMR neurons have values close to the country's highest values, regardless of how much that is,

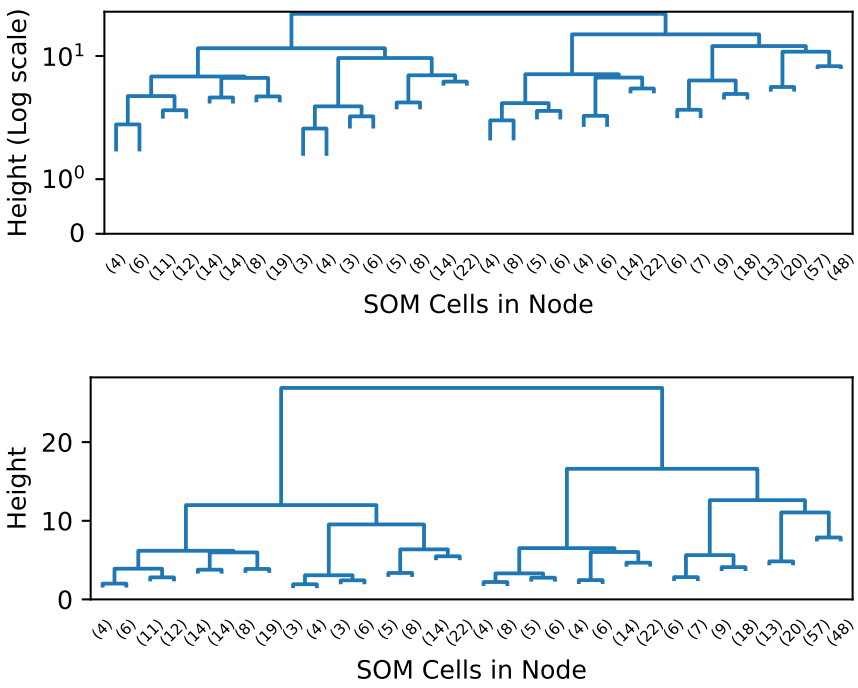


Figure 4.7: Dendrogram of SOM neuron

and lower PMR neurons have values close to the country's lowest.

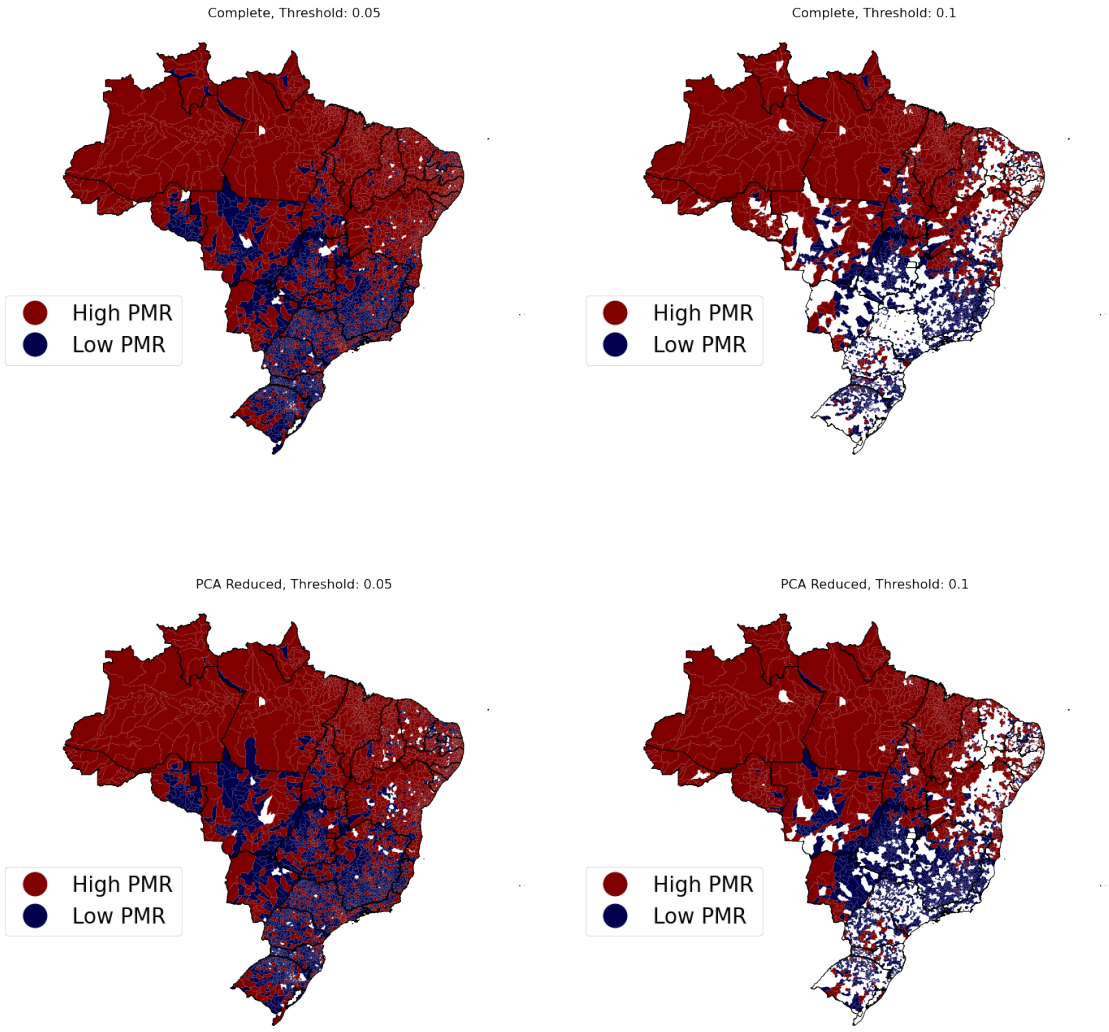


Figure 4.8: Map of municipalities that were classified in clusters of either High PMR or Low PMR, considering two different PMR thresholds and two datasets.

4.3 Discussion

In this work, self-organising maps (SOMs) were used to arrange Brazilian municipalities into socioeconomic clusters, using only SES factors as input. Then, the clusters were rearranged into larger and easier to comprehend clusters using hierarchical clustering and the Dynamic Cut algorithm. Different parameters were used in the SOMs, resulting in different clusters, which still tended to follow a similar or close regional pattern. Using a 10% threshold, the same strategy used in the *k*-Means method, we were able to visually compare High and Low PMR cluster regions using map visualisations. The results found using SOM corroborate with those found with *k*-Means, which worked with the same data using a different – and linear – clustering strategy, and the SOM-originated High and Low PMR clusters were pretty close regionally to the k-Means-originated ones. The main difference was that using the current work's model, South and Centre-West regions' clusters

were more easily discovered. This work also corroborates with the previously mentioned work on what the characteristics of these clusters are, with High PMR clusters' municipalities having below-average levels of education, sanitation and income, and Low PMR clusters' municipalities having above-average level of those.

The Self-Organising Maps model allows for a totally non-linear arrangement of municipalities by SES factors, and the hierarchical clustering – followed by branch cutting – allows these results from the many SOM configurations to be turned into more understandable groups for posterior analysis. The more specific clusters represented by each SOM neuron are merged when significantly similar, using distance metrics and *ward* distance, producing a final set of clusters. Using this method we were able to segment Brazilian municipalities into multiple socioeconomic clusters in such a way that every SOM configuration's final clusters included some with significantly high or low PMR levels (10% threshold used) – even without using preterm birth data as input to the algorithms. The method's outputs also leaned towards regionally centred clusters, with most clusters' municipalities being centred at some geographic region, which points to a good grouping by the hierarchical procedure used, as closer regions are likely to be more similar. We also applied the method to both a complete dataset and a PCA-reduced dataset (95% variance kept), to explore if the reduction in size or the loss of variance would have any significant impact on the result, which was not observed, as results by configurations on the complete and reduced dataset were different, but still pointed to very similar tendencies on the aggregated results.

Table 4.1: Quantization Error and Davies-Bouldin Index by configuration. **NF**: Neighbourhood Function. **DM**: Distance Metric. **QE**: Quantization Error. **DBS**: Davies-Bouldin Score

NF	σ	**DM**	**Complete**			**PCA Reduced**		
			Clusters	**QE**	**DBS**	**Clusters**	**QE**	**DBS**
Triangle	2	Manhattan	2	2.12	2.53	3	1.81	2.72
Triangle	3	Manhattan	2	2.08	2.69	3	1.83	2.62
Triangle	4	Manhattan	2	1.98	2.62	3	1.90	2.65
Triangle	5	Manhattan	2	1.98	2.67	4	1.96	3.34
Triangle	2	Euclidean	5	1.91	3.32	11	1.78	3.55
Triangle	3	Euclidean	9	1.90	3.53	8	1.80	3.32
Triangle	4	Euclidean	9	1.97	3.36	8	1.88	3.69
Triangle	5	Euclidean	8	2.02	3.46	8	1.93	3.41
Triangle	2	Cosine	10	1.88	3.53	11	1.79	3.61
Triangle	3	Cosine	12	1.90	3.44	11	1.81	3.40
Triangle	4	Cosine	10	1.98	3.44	12	1.88	3.68
Triangle	5	Cosine	12	2.03	3.28	10	1.94	3.52
Bubble	2	Manhattan	2	1.99	2.58	3	1.83	2.93
Bubble	3	Manhattan	4	2.15	2.66	5	1.99	3.35
Bubble	4	Manhattan	3	2.14	3.40	5	1.99	3.19
Bubble	5	Manhattan	4	2.14	3.09	4	1.99	3.31
Bubble	2	Euclidean	9	1.90	3.41	8	1.80	3.50
Bubble	3	Euclidean	9	2.07	3.01	8	1.97	3.44
Bubble	4	Euclidean	10	2.07	3.65	8	1.98	3.83
Bubble	5	Euclidean	9	2.07	3.20	9	1.97	3.32
Bubble	2	Cosine	12	1.90	3.57	11	1.80	3.44
Bubble	3	Cosine	12	2.07	3.80	11	1.97	3.73
Bubble	4	Cosine	10	2.07	3.61	9	1.98	3.62
Bubble	5	Cosine	12	2.07	3.70	10	1.98	3.62
Gaussian	2	Manhattan	3	2.07	2.74	3	1.92	2.79
Gaussian	3	Manhattan	3	2.14	2.64	6	2.00	3.17
Gaussian	4	Manhattan	3	2.20	2.83	7	2.06	3.62
Gaussian	5	Manhattan	4	2.27	3.25	6	2.12	3.27
Gaussian	2	Euclidean	10	2.00	3.63	10	1.90	3.58
Gaussian	3	Euclidean	10	2.07	3.50	9	1.98	3.09
Gaussian	4	Euclidean	11	2.14	3.40	9	2.04	3.54
Gaussian	5	Euclidean	10	2.20	3.65	11	2.10	3.44
Gaussian	2	Cosine	12	2.00	3.62	11	1.91	3.63
Gaussian	3	Cosine	13	2.07	3.66	11	1.98	3.47
Gaussian	4	Cosine	10	2.13	3.42	9	2.05	3.33
Gaussian	5	Cosine	11	2.19	3.76	12	2.11	3.64

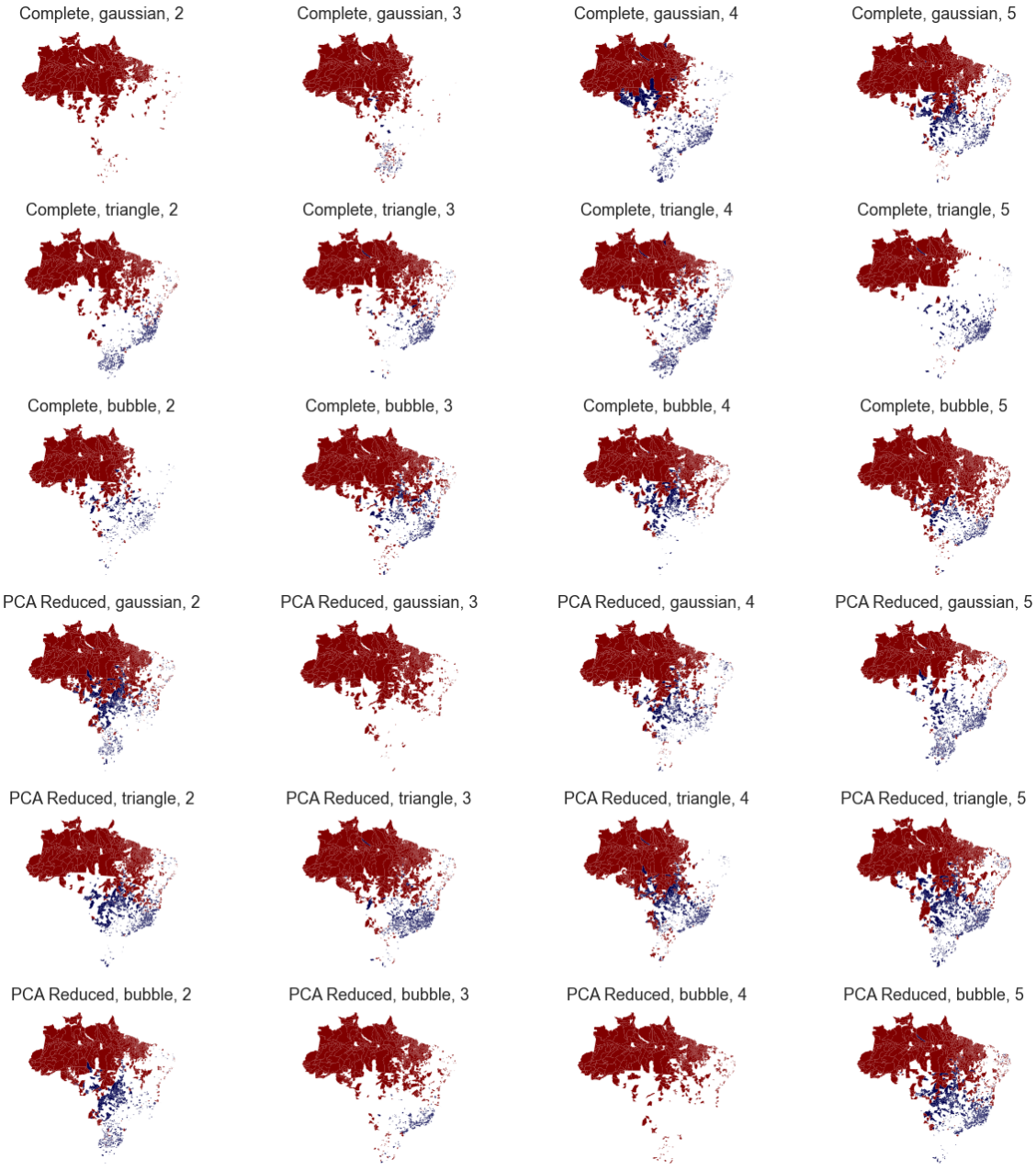


Figure 4.9: Map of municipalities that were classified in clusters of either High PMR or Low PMR using 10% threshold and split by dataset, neighbourhood functions, and value of σ.

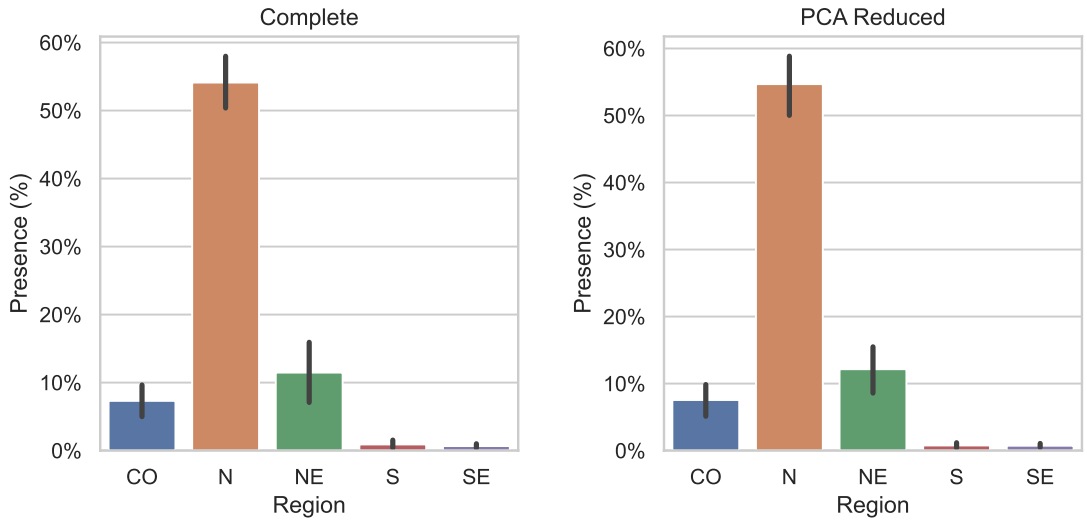


Figure 4.10: Presence (in %) of municipalities by region in High PMR clusters. **CO**: Centre-West, **N**: North, **NE**: Northeast, **S**: South, **SE**: Southeast

.

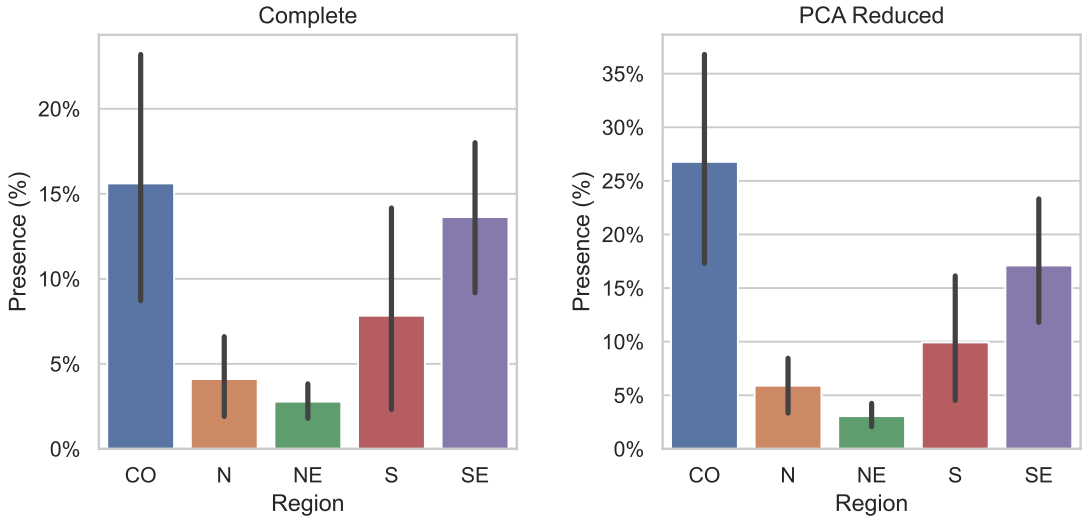


Figure 4.11: Presence (in %) of municipalities by region in Low PMR clusters. **CO**: Centre-West, **N**: North, **NE**: Northeast, **S**: South, **SE**: Southeast

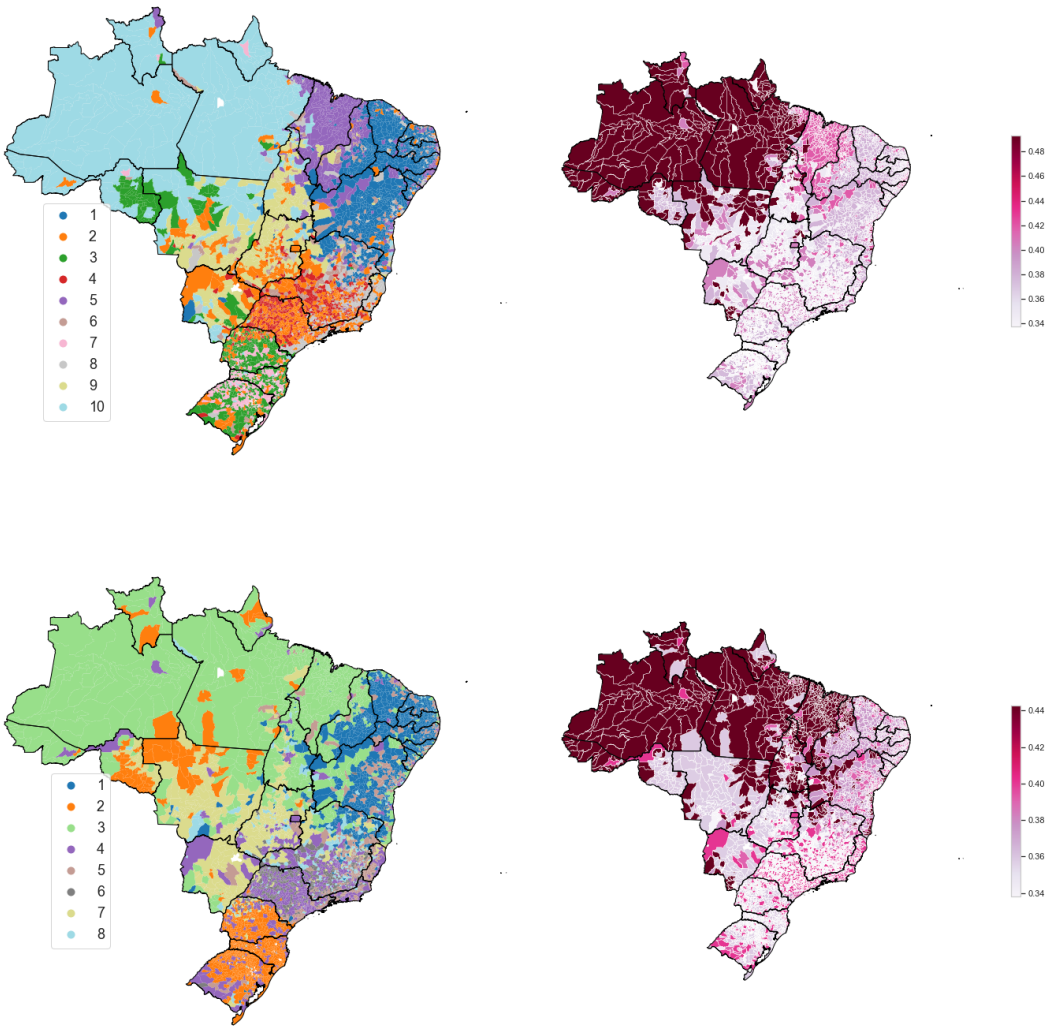


Figure 4.12: Municipal clusters shown on map. Left plot shows the final clusters given by the Dynamic Cut algorithm, right plot also shows these clusters, but colours them according the each cluster mean PMR.

Table 4.2: Comparison between Low PMR and High PMR cells, using the Normalised Distance (ND) between the mean values across all cells. ■ - Living Conditions, ■ - Race, ■ - Sanitation, ■ - Education, ■ - Working Conditions, ■ - Income, ■ - Household Type

		Higher On High PMR			**Higher On Low PMR**	
P	T	Feature	ND	T	Feature	ND
1	■	Oil/Gas/Kerosene Illumination	0.931	■	Piped Water	0.752
2	■	Wooden House - Proper Wood	0.912	■	Household with Bathroom	0.715
3	■	Wooden Floor - Proper Wood	0.909	■	Preschool	0.655
4	■	Ribeirinha (*Type*)	0.832	■	Literacy-level at School	0.651
5	■	Candle Illumination	0.813	■	Primary/Middle School	0.571
6	■	Wooden Floor - Improper Wood	0.806	■	Coated Bricks House	0.569
7	■	Wooden House - Improper Wood	0.789	■	Individual Electricity Meter	0.491
8	■	Open-Air Sewage Ditch	0.684	■	White	0.440
9	■	Never Attended School	0.654	■	Ceramic Tile/Stone Floor	0.419
10	■	Other Water Sources	0.647	■	Direct Rubbish Collection	0.377
11	■	Indigenous Family	0.589	■	Water Supply Network	0.373
12	■	No Electricity Meter	0.555	■	Literate	0.346
13	■	Straw House	0.488	■	Higher Education	0.335
14	■	Rubbish Dumped on Wasteland	0.475	■	Precollege Prep Course	0.329
15	■	Fishermen (*Type*)	0.454	■	Military or Public Employee	0.295
16	■	Adult Literacy Program	0.450	■	Rooms per Household	0.291
17	■	Employed in Agriculture	0.411	■	Regular Employment	0.269
18	■	Farmer (*Type*)	0.404	■	Private Dwelling	0.264
19	■	Bolsa Família-assisted Family	0.382	■	Complete Pavement	0.255
20	■	Rubbish Burnt or Buried	0.362	■	Avg Household Income	0.229
21	■	Pardo	0.337	■	Attended School	0.185
22	■	Rubbish Dumped on Sea/River	0.337	■	Black	0.169
23	■	Collective Dwelling	0.291	■	Urban	0.163
24	■	Extractivist (*Type*)	0.287	■	Salary	0.158
25	■	Well as Water Source	0.266	■	Concrete House	0.143
26	■	Incomplete Pavement	0.266	■	High School	0.128
27	■	Improvised Private Dwelling	0.256	■	Cistern Water	0.125
28	■	Wastewater to River/Lake/Sea	0.248	■	Income From Alimony	0.122
29	■	No Pavement	0.228		Age	0.114
30	■	Unpaid Job	0.225	■	Sewage System	0.107

Chapter 5

Comparison and Overview

In this chapter, we present a couple of comparisons regarding the methods applied in this work. First, in Section 5.1, we make a limited comparison between the results of the two methods, denoting their similarities and differences. Then, in Section 5.2, we confront the results of the methods to the findings of some of the key studies previously mentioned in Section 1.1. Finally, we present a general overview of the research in 5.3.

5.1 Comparison of Methods

At it was already mentioned in these last two sections, both algorithms seemed to be successful at clustering the SES data and finding clusters that are interesting in a PTB point of view. Working with over 100 dimensions, it's not quite possible to comprehend the relationship between every single pair (or group) of dimensions. For that reason, a comparison between the findings of the linear (k-Means) and non-linear (SOM) methods must restrain itself to analysing the most prominent features among the High PMR and Low PMR clusters found by them, as well as the regional/municipal differences between them.

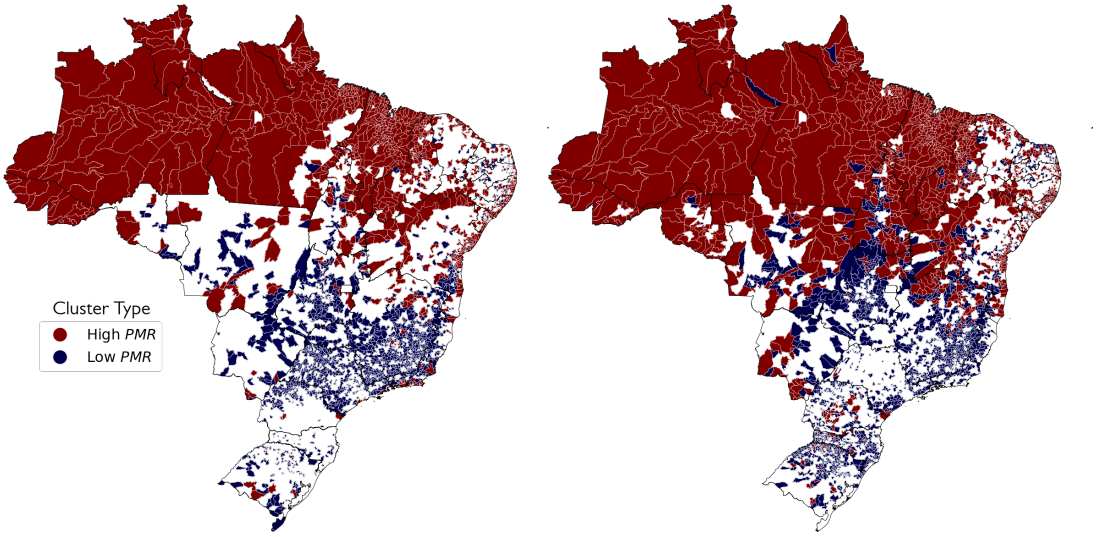
First, we can make a comparison by looking at the relevant features present in Figure 3.14 and 3.1 and contrasting them with relevant features from Table 4.2. Figure 3.14 shows us that Higher Income, Piped Water, # of Rooms per Household, Sewage System, Higher Education and Proportion of White Individuals are characteristic of Low PMR municipalities. All of these are also shown in Table 4.2 to be some of the most diverging factors of PMR, being considerably higher on Low PMR SOM cells. Piped Water stands out as it had the largest normalised difference observed in favour of Low PMR, with a value of 0.752, which converges with the k-Means method as Low PMR clusters presented significantly larger levels of residences with Piped Water. Both methods found High and Low PMR clusters that diverged mostly on the Sanitation and Living Conditions features, the SOM method also showed to give a larger importance to educational features.

For the regional comparison, we can look at Figure 5.1. There, in item (a) we can see the final High PMR and Low PMR cluster regions as discovered by the k-Means method, and in item (b) the aggregate of High and Low PMR findings by the SOM method configurations. In a direct comparison, the maps show a very similar pattern, both methods found the High PMR clusters to be in the northern parts of the country,and the Low PMR

clusters in the southern parts of the country.

One visible difference, though, is how the Low PMR clusters on the linear model are strongly present in the State of São Paulo, and considerably absent in both Santa Catarina and in the North of Rio Grande do Sul. The non-linear Low PMR clusters had that reversed, with a big presence in Santa Catarina, in the North of Rio Grande do Sul, and considerably absent in São Paulo. Another difference, as already mentioned in the last section, is how the non-linear model found more clusters including cities located on the Centre-West of Brazil.

Generally speaking, both models saw a considerably similar geographical behaviour for cluster with high and low levels of PTB.



(a) k-Means (b) SOM

Figure 5.1: Side by side view of the 10% threshold maps as shown in 3.2 and 4.2, for comparison.

5.2 Comparison to State of the Art

The results found in this clustering process corroborate and add to the discoveries made in Adhikari et al. (2019). Their study uses a considerably smaller group for analysis (5,297 pregnant women), does prediction - logistic regression - instead of clustering, and uses SES factors of income, education and employment. Their results suggest that SES factors can help improve accuracy when predicting PTB, thus implying the existence of a relationship between SES factors and PTB. The fact that we were able to observe a similar relationship, with significant difference in PTB among manifold SES clusters, even when working with data from a different country, with a much larger population, with a larger number of features and a different learning algorithm, strengthens the idea that such relationship is indeed meaningful. Their work also notes how such relationship is restricted, how having such SES information – combined with some individual-level

data that they used – is still insufficient for a real-world clinical application of predicting PTB (their work's goal), which matches with our pre-work thoughts on PTB, that it is a multi-factorial problem, and also helps to explain our difficulty and our need to develop an alternative method to achieve our clustering goals using *k*-Means. There are many aspects of PTB invisible to both our and their works, and small differences when dealing with different levels of SES are observable, but are naturally limited.

How we were able to cluster SES factors in a non-personal level and observe that the regions found had considerable PTB difference draws comparison to and supports the findings in Deguen et al. (2018). Their study finds socioeconomic clusters around the city of Paris' blocks using a spatial clustering technique. They clustered the city into areas and found out the ones where mothers are most likely of experience PTB, and then adjusted the clustering using as control variable an SES index created from 41 original SES variables available. After the adjustment, their results suggest a considerable influence of SES factors on PTB occurrences, as the non-adjusted model showed a much more significant (smaller) *p*-value. This interpretation is supported by the results of our work by both the *k*-Means method and the SOM method, as it's perceptible in the Brazilian map of clusters shown for both methods how some regions differ considerably in terms of PTB rate, and similarly to their work, those clusters predominantly assume a regional-centred aspect, creating contiguous areas of similar SES characteristics, of which some have significantly higher or lower PTB rate. Unlike in Deguen et al. (2018), which uses the geographical location as part of its spatial algorithm, this regional-centred aspect was not intended nor influenced by any location feature, which were removed from the original dataset. For that reason, we obtained an outcome that reemphasises this strong regional aspect of SES and PTB, and consequently how neighbouring regions affect the SES and PTB outcomes of a given municipality or city block.

In the feature-wise view presented for both the *k*-Means and the SOM method, a few types of variables stood out. Sanitation variables were among the most outstanding in both models, with proper sewage system, water system, and garbage collecting system, being considerably more present in High PMR regions, this can be linked to and reinforces the findings by Padhi et al. (2015), Baker et al. (2018), and Patel et al. (2019). Although their works don't assess all of these sanitation points, many key variables found are strongly linked to their work, and the "Household with Bathroom" variable, a major point of their analyses showed a much higher presence on higher PMR clusters, being one of the most diverging features in the *k*-Means method. As shown in the SOM method results, in the areas with highest PMR, education was a major factor, literate populations and populations with at least some standard education (Primary/Middle) were mostly among Low PMR clusters. This was also viewed in a general comparison for the *k*-Means method and corroborates with the findings by Ruiz et al. (2015), Cantarutti et al. (2017) and Taha et al. (2020), all of which found statistically significant differences in PTB when stratified by mother's education. Another clear disparity observed was related to race/skin, with white skin being one of the features most strongly related to Low PMR, this corroborates the findings of the meta-analysis presented by de Oliveira et al. (2018), which aggregates several studies related to race and preterm birth occurrences in the United States between 2010 and 2015 and finds a higher risk (1.51 OR) of PTB among historically disfavoured

racial groups. Although the exact groups can't be compared properly, as the Brazilian and U.S. populations have significant racial differences, the high presence of whites in Low PMR clusters and the high presence of *pardos* and indigenous people in the High PMR provide a strong ratification of their results.

A general summary of studies analysing PTB by SES factors, in the general sense, excluding those aiming only at specific aspects of SES (i.e. education, sanitation), can be seen in Table 5.1. It includes the two studies first mentioned in this section, in addition to another study by Ochoa et al. (2021), which uses a different strategy, thus being relatively harder to be fairly compared to our study, although a simplistic analysis of their results would proceed to commentaries similar to those made in the comparison to Adhikari et al. (2019). The table shows the distinction and similarities between the studies, and provides a brief description of its main findings regarding SES and PTB.

5.3 Overview

Although many studies have explored the subject analysed here, there was no such study found for comparison that employs precisely the 3 key points worked: Preterm birth, SES data and unsupervised learning. Our work also aimed at expanding such the knowledge by applying both linear and non-linear methods, namely k-Means and SOM. This work provides two methods that allow cluster analysis on high-dimensional datasets, and applies these methods to enable the analysis of PTB Rate through SES factors.

The proposed method was successful on clustering the country and finding sets of similar municipalities with considerably diverging preterm birth, but the general analysis shown here is an isolated view of SES factors only and should be followed by future analyses. Future studies could enrich these findings by working with some of the 104 features individually, or by discovering which ones are the main driving factors of PTB, or even by joining SES factors with genetic, vital, behavioural and climate data to approximate the actual impact of SES factors. Preterm birth analyses reach many areas of study, and SES is just one of the considered factors, an isolated study such as this is naturally limited in its results.

Table 5.1: Comparison of works relating SES and PTB, including studies that make use of a general SES view only.

Reference	Area	SES Variable	Time period	Purpose	Method	Findings
Adhikari et al. (2019)	Canada	Single - Aggregate	Cross-sectional	Prediction	Multi-level Logistic Regression	5.72% of PTB variance attributed to SES
Ochoa et al. (2021)	Netherlands	Single - Based on Income	Longitudinal	Correlation analysis	Multi-level Logistic Regression	Significantly higher PTB risk in neighbourhoods with Stable Low SES (1.12 OR) or SES declining to low (1.09 OR)
Deguen et al. (2018)	Paris, France	Single - Aggregate	Cross-sectional	Clustering	Spatial Clustering	Spatial clustering of high PTB risk blocks only significant when SES index was included
k-Means Method	Brazil	Multiple	Cross-sectional	Clustering	*k*-Means Clustering	Clusters of high PTB risk generally concentrated in areas of low SES, and vice-versa
SOM Method	Brazil	Multiple	Cross-sectional	Clustering	Self-Organising Maps	

Chapter 6

Conclusion

This chapter presents the main conclusions of this research. First, in Section 6.1, the main findings of the study were summarised. Section 6.2 presents a few implications of the results and methods presented. Finally, in Section 6.3, the main limitations are discussed and suggestions for future research to continue this work and solve these limitations are given.

6.1 Answers and Findings

The goal of this research was defined as:

> The main goals of this work are to stratify the risk of PTB in Brazil from SES factors alone, to obtain a general and feature-level view of factors that may affect PTB, to uncover which areas of Brazil put their women is higher risk of experiencing PTB, and to do all that automatically – leaving the decision of feature relevance entirely to the machine – using linear and non-linear algorithms.

Starting from the hypothesis that SES factors are one of the causes of PTB, which was strongly defended and demonstrated by several aforementioned authors, we approached the problem from a new perspective, trying to see if a pool of raw socioeconomic data with high dimensionality could be clustered in a way where areas of a singularly high level of PTB would be naturally discovered by their socioeconomic arrangement. That was successively achieved by both the k-Means Method and the SOM Method. The methods were able to find socioeconomic municipality segments with high and low levels of PTB.

We were able to extract meaningful information on municipal clusters of high and low PTB through our clustering methods – both of which employed a combination of different clustering techniques and dimensionality reduction strategies based on unsupervised learning – and the feature-rich municipalities' socioeconomic dataset. The results of the K-Means method and the SOM method suggest a clear socioeconomic contrast between clusters with high and low risk of PTB, with high-risk clusters predominantly located in regions with the worst social indexes. Most clusters were regional, and even those of high PTB (or low PTB) had considerably different characteristics compared to those of the same type. The North region was the most outstanding geographical focus as it was

– almost in its entirety – a single high PTB cluster, always present in the manifold tested clustering scenarios and having the highest PTB overall.

This research saw how the quality of life and quality of public services may affect, in a positive way, the reduction of PTB occurrences among the Brazilian population in such a way that these factors should always be taken into account in any general mainstream study on PTB.

6.2 Implications

For research and health

The fact that we were able to uncover these clusters across the country using only socioeconomic factors is a good indication that previous works in the area that indicated such relations were indeed correct – or at least on the right path – in their findings.

Previous research had not yet explored machine learning algorithms for clustering, predominantly using popular statistical methods, such as multivariate logistic regression, to draw conclusions. Here, we brought to the SES-PTB studies a set of powerful techniques that allows us the use richer datasets and that are able to discover the studied relationship in a completely different manner. We also presented two distinct clustering methods, one linear and one non-linear, based on k-Means and SOM. Having multiple algorithms come to a similar conclusion for a non-trivial problem – even using very different strategies – strengthens, even more, the idea of socioeconomic influence on PTB. SES factors were shown once again to have a relationship to PTB, being it reasonable to consider such factors when trying to understand or predict PTB.

From a public health perspective, the implications of this research go beyond studies on the subject. The results not only imply that areas of worse socioeconomic status will have a higher risk of PTB but also define the regions most likely suffering from this risk. Decision-making personnel from public health institutions could use this information to help decide what actions to take to reduce PTB occurrences in the country and where to take them.

For clustering methods

The problem tackled here isn't a simple clustering problem, as has been mentioned. Only finding the "best" clustering scenario would not necessarily (and didn't) work to uncover the clusters we were aiming to find. The methods developed and described in this work were designed to solve extraordinary clustering scenarios.

The k-Means method, for instance, is a workaround that allows not only k-Means (linear) clustering in high dimensions, but it also includes a control variable, excluded from the clustering, a variable presumably to be studied in order to be compared to the clustering result. This method can be used for any clustering problem of a similar nature, expressly: when wanting to find clusters in a particular dataset that are somehow related and interfere with another variable, but without including the latter in the clustering. Including the comparable variable in the clustering would result only in biased results.

6.3 Limitations and Further Research

Even though it is the central idea of this research, working only with socioeconomic factors is the primary and most relevant limitation of our results. PTB is a complex and multifactorial phenomenon, and the search for its causes demands analyses of several different aspects. Some external factors, behavioural or meteorological, targeting only specific regions, could have helped altering the PTB on some of the regional clusters discovered. And, of course, the primary causes of PTB are biological. Using municipality-level data may help reduce some of the bias that would come from these other factors, and yet an analysis that doesn't take into account all influencing aspects will always miss something and therefore can and should be improved, confirmed, and/or corrected by future research.

Time is another point for improving. This study was based on some rich and (thus) large datasets kindly provided by the Federal Government of Brazil. Preparing and processing these datasets can sometimes be problematic, as it was when this research was still in its inception period and the choice was made to use the most up-to-date set of datasets, which were the ones from 2018. Not including other years is not a problem *per se*, most related works also use a single fixed period, but using data from other years could bring interesting insights into the trends of PTB and even create possibilities for the application of sequential machine learning algorithms, such as RNNs, to this problem.

Bibliography

Adhikari, Kamala, Scott B Patten, Tyler Williamson, Alka B Patel, Shahirose Premji, Suzanne Tough, Nicole Letourneau, Gerald Giesbrecht & Amy Metcalfe (2019), 'Does neighborhood socioeconomic status predict the risk of preterm birth? a community-based canadian cohort study', *BMJ open* **9**(2), e025341.

Alleman, Brandon, Amanda Smith, Heather Byers, Bruce Bedell, Kelli Ryckman, Jeffrey Murray & Kristi Borowski (2013), 'A proposed method to predict preterm birth using clinical data, standard maternal serum screening, and cholesterol', *American journal of obstetrics and gynecology* **208**.

Baker, Kelly, William Story, Evan Walser-Kuntz & M. Bridget Zimmerman (2018), 'Impact of social capital, harassment of women and girls, and water and sanitation access on premature birth and low infant birth weight in india', *PLoS ONE* **13**, e0205345.

Beeckman, Katrien, Sabine Putte, Koen Putman & Fred Louckx (2009), 'Predictive social factors in relation to preterm birth in a metropolitan region', *Acta obstetricia et gynecologica Scandinavica* **88**, 787–92.

Ben-Dor, Amir, Ron Shamir & Zohar Yakhini (1999), 'Clustering gene expression patterns', *Journal of computational biology : a journal of computational molecular cell biology* **6**, 281–97.

Borgen, Fred & David Barnett (1987), 'Applying cluster analysis in counseling psychology research', *Journal of Counseling Psychology* **34**, 456–468.

Buen, Mariana, Eliana Amaral, Renato Souza, Renato Passini, Giuliane Lajos, Ricardo Tedesco, Marcelo Nomura, Tabata Dias, Patrícia Rehder, Maria Sousa & Jose Cecatti (2020), 'Maternal work and spontaneous preterm birth: A multicenter observational study in brazil', *Scientific Reports* **10**.

Cantarutti, Anna, Matteo Franchi, Matteo Monzio Compagnoni, Luca Merlino & Giovanni Corrao (2017), 'Mother's education and the risk of several neonatal outcomes: An evidence from an italian population-based study', *BMC Pregnancy and Childbirth* **17**.

Catley, Christina, Monique Frize, C Walker & Dorina Petriu (2006), 'Predicting high-risk preterm birth using artificial neural networks', *IEEE transactions on information*

technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society **10**, 540–9.

Chen, Mengfan, Ni Xie, Zhaoxia Liang, Tingting Qian & Danqing Chen (2020), 'Early prediction model for preterm birth combining demographic characteristics and clinical characteristics'.

Datasus (n.d.), 'SINASC - Sistema de Informações de Nascidos Vivos'.
URL: *http://www2.datasus.gov.br/DATASUS/index.php?area=060702*

de Oliveira, Kelly, Edna de Araújo, Keyte de Oliveira, Cesar Augusto Casotti, Carlos da Silva & Djanilson Santos (2018), 'Association between race/skin color and premature birth: a systematic review with meta-analysis', *Revista de Saúde Pública* **52**.

DeFranco, Emily, Min Lian, Louis Muglia & Mario Schootman (2008), 'Area-level poverty and preterm birth risk: A population-based multilevel analysis', *BMC public health* **8**, 316.

Deguen, Severine, Nina Ahlers, Morgane Gilles, Arlette Danzon, Marion Carayol, Denis Zmirou-Navier & Wahida Kihal-Talantikite (2018), 'Using a clustering approach to investigate socio-environmental inequality in preterm birth—a study conducted at fine spatial scale in paris (france)', *International journal of environmental research and public health* **15**(9), 1895.

Esplin, Michael, Tracy Manuck, Bryce Christensen, Joseph Biggio, Radek Bukowski, Samuel Parry, Heping Zhang, Michael Varner, William Andrews, George Saade, Yoel Sadovsky, Uma Reddy & John Ilekis (2015), 'Cluster analysis of spontaneous preterm birth phenotypes identifies potential associations among preterm birth mechanisms', *American Journal of Obstetrics and Gynecology* **212**, S107–S108.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander & Xiaowei Xu (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *em* 'KDD'.

França, Elisabeth, Sônia Lansky, Maria Rego, Deborah Carvalho Malta, Júlia Santiago França, Renato Teixeira, Denise Porto, Marcia Almeida, Maria De Fatima Marinho de Souza, Célia Szwarcwald, Meghan Mooney, Mohsen Naghavi & Ana Vasconcelos (2017), 'Principais causas da mortalidade na infância no brasil, em 1990 e 2015: estimativas do estudo de carga global de doença', *Revista Brasileira de Epidemiologia* **20**, 46–60.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville (2016), *Deep Learning*, MIT Press.
URL: *http://www.deeplearningbook.org*

Granese, Roberta, Eloisa Gitto, Gabriella D'Angelo, Raffaele Falsaperla, Giovanni Corsello, Donatella Amadore, Gloria Calagna, Ilaria Fazzolari, Roberta Grasso & Onofrio Triolo (2019), 'Preterm birth: Seven-year retrospective study in a single centre population', *Italian Journal of Pediatrics* **45**.

Grjibovski, Andrej, Lars Bygren, Agneta Yngve & Michael Sjostrom (2005), 'Large social disparities in spontaneous preterm birth rates in transitional russia', *Public health* **119**, 77–86.

Group, ESHRE Capri Workshop (2005), 'Fertility and ageing', *Human Reproduction Update* **11**(3), 261–276.
URL: *https://doi.org/10.1093/humupd/dmi006*

Hill, Jacquelyn, M Campbell, Guang Zou, John Challis, Gregor Reid, Hiroshi Chisaka & Alan Bocking (2008), 'Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers', *American journal of obstetrics and gynecology* **198**, 468.e1–7; discussion 468.e7.

Huang, Jin, Yating Qian, Mingming Gao, Hongjuan Ding, Lei Zhang & Ruizhe Jia (2020), 'Analysis of factors related to preterm birth: a retrospective study at nanjing maternity and child health care hospital in china', *Medicine* **99**, e21172.

IBGE (n.d.), 'Estimativas da população'.
URL: *https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html*

Institute of Medicine (2007), *Preterm Birth: Causes, Consequences, and Prevention*, The National Academies Press, Washington, DC.
URL: *https://www.nap.edu/catalog/11622/preterm-birth-causes-consequences-and-prevention*

Istvan, Marion, Florence Rouget, Léah Michineau, Christine Monfort, Luc Multigner & Jean-François Viel (2019), 'Landfills and preterm birth in the guadeloupe archipelago (french west indies): A spatial cluster analysis', *Tropical Medicine and Health* **47**.

Kaufman, J., F. Alonso & P. Pino (2008), 'Multi-level modeling of social factors and preterm delivery in santiago de chile', *BMC Pregnancy and Childbirth* **8**, 46 – 46.

Kawachi, Ichiro & Lisa F Berkman (2003), *Neighborhoods and health*, Oxford University Press.

Kim, Yun-Sook (2019), 'Analysis of spontaneous preterm labor and birth and its major causes using artificial-neural-network', *Journal of Korean Medical Science* **34**.

Kohonen, Teuvo (2014), *MATLAB Implementations and Applications of the Self-Organizing Map*, Unigrafia Oy, Helsinki, Finland.

Langfelder, Peter, Bin Zhang & Steve Horvath (2007), 'Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R', *Bioinformatics* **24**(5), 719–720.
URL: *https://doi.org/10.1093/bioinformatics/btm563*

Lee, Jennifer, Jinjin Cai, Fuhai Li & Zachary Vesoulis (2021), 'Predicting mortality risk for preterm infants using random forest', *Scientific Reports* **11**, 7308.

Lopes Jr, Márcio (2022), 'PTB and SES Research'.
URL: *https://github.com/marciojunior159/ptb_ses_research*

Marsland, Stephen (2009), *Machine Learning - An Algorithmic Perspective.*, Chapman and Hall / CRC machine learning and pattern recognition series, CRC Press.

Metcalfe, Amy, Parabhdeep Lail, William A Ghali & Reg S Sauve (2011), 'The association between neighbourhoods and adverse birth outcomes: A systematic review and meta-analysis of multi-level studies', *Paediatric and perinatal epidemiology* **25**(3), 236–245.

Ministério da Cidadania (n.d.), 'Base desidentificada do Cadastro Único com marcação do Bolsa Família'.
URL: *https://aplicacoes.mds.gov.br/sagi/portal/index.php?grupo=212*

Modell, Bernadette, RJ Berry, Coleen A Boyle, Arnold Christianson, Matthew Darlison, Helen Dolk, Christopher P Howson, Pierpaolo Mastroiacovo, Peter Mossey & Judith Rankin (2012), 'Global regional and national causes of child mortality', *The Lancet* **380**(9853), 1556.
URL: *https://www.sciencedirect.com/science/article/pii/S0140673612618789*

Ochoa, Lizbeth, Loes Bertens, Pilar García-Gómez, Tom Van Ourti, Eric Steegers & Jasper Been (2021), 'Association of neighbourhood socioeconomic trajectories with preterm birth and small-for-gestational-age in the netherlands: a nationwide population-based study', *The Lancet Regional Health - Europe* **10**, 100205.

Oliveira, Adelaide Alves de, Marcia Furquim de Almeida, Zilda Pereira da Silva, Paula Lisiane de Assunção, Ana Maria Rigo Silva, Hellen Geremias dos Santos & Gizelton Pereira Alencar (2019), 'Fatores associados ao nascimento pré-termo: da regressão logística à modelagem com equações estruturais', *Cadernos de Saúde Pública* **35**.

Organization, World Health (2012), 'Born too soon: the global action report on preterm birth'.

Padhi, Bijaya K., Kelly K. Baker, Ambarish Dutta, Oliver Cumming, Matthew C. Freeman, Radhanatha Satpathy, Bhabani S. Das & Pinaki Panigrahi (2015), 'Risk of adverse pregnancy outcomes among women practicing poor sanitation in rural india: A population-based prospective cohort study', *PLOS Medicine* **12**(7), 1–18.
URL: *https://doi.org/10.1371/journal.pmed.1001851*

Passini, Jr, Renato, Jose G. Cecatti, Giuliane J. Lajos, Ricardo P. Tedesco, Marcelo L. Nomura, Tabata Z. Dias, Samira M. Haddad, Patricia M. Rehder, Rodolfo C. Pacagnella, Maria L. Costa, Maria H. Sousa & for the Brazilian Multicentre Study on Preterm Birth study group (2014), 'Brazilian multicentre study on preterm birth

(emip): Prevalence and factors associated with spontaneous preterm birth', *PLOS ONE* **9**(10), 1–12.
URL: *https://doi.org/10.1371/journal.pone.0109069*

Patel, Ratna, Ajay Gupta, Chauhan Shekhar & Dhananjay W. Bansod (2019), 'Effects of sanitation practices on adverse pregnancy outcomes in india: a conducive finding from recent indian demographic health survey', **19**.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay (2011), 'Scikit-learn: Machine learning in python', *Journal of Machine Learning Research* **12**(85), 2825–2830.
URL: *http://jmlr.org/papers/v12/pedregosa11a.html*

Ruiz, Milagros, Peter Goldblatt, Joana Morrison, Lubomír Kukla, Jan Švancara, Marjo Riitta-Järvelin, Anja Taanila, Marie-Josephe Saurel-Cubizolles, Sandrine Lioret, Chryssa Bakoula, Alexandra Veltsista, Daniela Porta, Francesco Forastiere, Manon Eijsden, Tanja Vrijkotte, Merete Eggesbø, Richard White, Henrique Barros, Sofia Correia & Hynek Pikhart (2015), 'Mother's education and the risk of preterm and small for gestational age birth: A drivers meta-analysis of 12 european cohorts', *Journal of epidemiology and community health* **69**.

Santoso, Noviyanti & Sri Wulandari (2018), 'Hybrid support vector machine to preterm birth prediction', *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)* **8**, 191.

Saurel-Cubizolles, Marie-Josephe, Jennifer Zeitlin, Nathalie Lelong, Emile Papiernik, Gian Renzo & Gérard Bréart (2004), 'Employment, working conditions, and preterm birth: Results from the europop case-control survey', *Journal of epidemiology and community health* **58**, 395–401.

Stylianou-Riga, Paraskevi, Panayiotis Kouis, Paraskevi Kinni, Angelos Rigas, Thalia Papadouri, Panayiotis Yiallouros & Mamas Theodorou (2018), 'Maternal socioeconomic factors and the risk of premature birth and low birth weight in cyprus: a case–control study', *Reproductive Health* **15**.

Sun, Jiajia & Yaoguo Li (2015), 'Multidomain petrophysically constrained inversion and geology differentiation using guided fuzzy c-means clustering', *Geophysics* **80**, ID1–ID18.

Sun, Shengzhi, Kate Weinberger, Keith Spangler, Melissa Eliot, Joseph Braun & Gregory Wellenius (2019), 'Ambient temperature and preterm birth: A retrospective study of 32 million us singleton births', *Environment International* **126**.

Taha, Zainab, Ahmed Ali Hassan, Ludmilla Wikkeling-Scott & Dimitrios Papandreou (2020), 'Factors associated with preterm birth and low birth weight in abu dhabi, the

united arab emirates', *International Journal of Environmental Research and Public Health* **17**.

Vettigli, Giuseppe (2018), 'MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map'.
URL: *https://github.com/JustGlowing/minisom/*

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt & SciPy 1.0 Contributors (2020), 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods* **17**, 261–272.

Włodarczyk, Tomasz, Szymon Płotka, Przemysław Rokita, Nicole Sochacki-Wójcicka, Jakub Wójcicki, Michał Lipa & Tomasz Trzciński (2020), 'Spontaneous preterm birth prediction using convolutional neural networks'.

Appendix A

Full list of features

Table A.1: $\mathbf{A}_0$ Features: Target, Living Conditions and Race

Feature	Type	Min	Max	Mean	Median	SD
PTB Municipal Rate	Target	0.00	0.02	0.00	0.00	0.00
Coated Bricks House	Living Conditions	0.00	1.00	0.65	0.71	0.26
Uncoated Brick House	Living Conditions	0.00	0.89	0.13	0.10	0.11
Dirt Floor	Living Conditions	0.00	0.84	0.04	0.01	0.09
Wooden House - Improper Wood	Living Conditions	0.00	0.70	0.02	0.00	0.06
Wooden House - Proper Wood	Living Conditions	0.00	0.96	0.11	0.00	0.20
Straw House	Living Conditions	0.00	0.66	0.00	0.00	0.02
Full Paving	Living Conditions	0.00	1.00	0.46	0.43	0.27
Coated Rammed Earth Hous	Living Conditions	0.00	0.35	0.01	0.00	0.03
Community Electricity Meter	Living Conditions	0.00	0.69	0.06	0.02	0.08
Individual Electricity Meter	Living Conditions	0.02	1.00	0.83	0.88	0.15
No Electricity Meter	Living Conditions	0.00	0.73	0.03	0.01	0.06
Oil/Gas/Kerosene Illumination	Living Conditions	0.00	0.60	0.01	0.00	0.03
Candle Illumination	Living Conditions	0.00	0.41	0.01	0.00	0.02
No Pavement	Living Conditions	0.00	1.00	0.43	0.44	0.25
Incomplete Pavement	Living Conditions	0.00	0.91	0.06	0.03	0.09
Wooden Floor - Improper Wood	Living Conditions	0.00	0.75	0.02	0.00	0.07
Uncoated Rammed Earth House	Living Conditions	0.00	0.73	0.01	0.00	0.05
Wooden Floor - Proper Wood	Living Conditions	0.00	1.00	0.07	0.00	0.16
Private Dwelling	Living Conditions	0.09	1.00	0.95	1.00	0.11
Ceramic Tile/Stone Floor	Living Conditions	0.00	1.00	0.41	0.39	0.25
Concrete House	Living Conditions	0.00	1.00	0.40	0.38	0.25
Rooms per Household	Living Conditions	1.92	7.06	4.76	4.79	0.57
# of Rooms for Sleeping	Living Conditions	0.01	3.88	2.01	2.00	0.27
Urban	Living Conditions	0.00	1.00	0.68	0.70	0.23
Improvised Private Dwelling	Living Conditions	0.00	0.91	0.05	0.00	0.10
Collective Dwelling	Living Conditions	0.00	0.44	0.00	0.00	0.02
People in Dwelling	Living Conditions	2.17	5.80	3.40	3.37	0.30
Carpet Floor	Living Conditions	0.00	0.08	0.00	0.00	0.00
Yellow/Asian	Race	0.00	0.32	0.00	0.00	0.01
White	Race	0.00	1.00	0.38	0.32	0.27
Pardo	Race	0.00	1.00	0.54	0.58	0.26
Black	Race	0.00	1.00	0.07	0.05	0.07

Table A.2: $\mathbf{A}_0$ Features: Education, Sanitation and Income

Feature	Type	Min	Max	Mean	Median	SD
Adult Literacy Program	Education	0.00	0.49	0.03	0.02	0.03
Higher Education	Education	0.00	0.50	0.03	0.02	0.04
Precollege Prep Course	Education	0.00	0.50	0.03	0.02	0.04
Preschool	Education	0.65	1.00	0.98	0.99	0.02
No Education	Education	0.00	0.33	0.00	0.00	0.01
High School	Education	0.00	1.00	0.53	0.53	0.11
Primary/Middle School	Education	0.65	1.00	0.98	0.98	0.02
Literacy-level at School	Education	0.65	1.00	0.98	0.99	0.02
Public Education	Education	0.00	0.79	0.18	0.17	0.06
Never Attended School	Education	0.00	0.29	0.02	0.01	0.02
Attended School	Education	0.21	1.00	0.80	0.80	0.07
Private Education	Education	0.00	0.33	0.01	0.00	0.02
Finished Courses	Education	0.00	0.89	0.53	0.53	0.11
Literate	Education	0.67	1.00	0.96	0.96	0.03
Rubbish Dumped on Wasteland	Sanitation	0.00	0.37	0.01	0.00	0.03
Rubbish Dumped on Sea/River	Sanitation	0.00	0.07	0.00	0.00	0.00
Rubbish Burnt or Buried	Sanitation	0.00	1.00	0.22	0.17	0.20
Indirect Rubbish Collection	Sanitation	0.00	0.93	0.05	0.02	0.09
Household with Bathroom	Sanitation	0.17	1.00	0.92	0.98	0.12
Open-Air Sewage Ditch	Sanitation	0.00	0.88	0.02	0.00	0.05
Wastewater to River/Lake/Sea	Sanitation	0.00	0.79	0.01	0.00	0.05
Sewage System	Sanitation	0.00	1.00	0.32	0.16	0.34
Septic Tank	Sanitation	0.00	1.00	0.17	0.08	0.22
Piped Water	Sanitation	0.01	1.00	0.87	0.94	0.18
Cesspit	Sanitation	0.00	1.00	0.35	0.28	0.31
Cistern Water	Sanitation	0.00	0.87	0.04	0.00	0.09
Other Water Sources	Sanitation	0.00	0.83	0.04	0.01	0.07
Direct Rubbish Collection	Sanitation	0.00	1.00	0.67	0.72	0.25
Well as Water Source	Sanitation	0.00	1.00	0.20	0.15	0.19
Water Supply Network	Sanitation	0.00	1.00	0.67	0.71	0.24
Income From Donations	Income	0.00	132.16	7.63	3.95	10.77
Avg Household Income	Income	1.56	926.29	169.38	157.87	90.45
Income From Pension	Income	0.00	318.00	29.82	25.33	24.68
Income From Alimony	Income	0.00	183.21	9.05	5.06	11.92
Yearly Personal Income	Income	0.00	14,859.69	3,506.72	3,240.63	2,012.74
Unemployment Benefits	Income	0.00	203.70	2.69	0.00	7.96
Salary	Income	0.00	1250.00	147.44	125.52	105.35
Income (Other Sources)	Income	0.00	279.43	7.27	1.23	16.99

Table A.3: $\mathbf{A}_0$ Features: Household Type, Working Conditions and others

Feature	Type	Min	Max	Mean	Median	SD
Camping Family	Household Type	0.00	0.33	0.00	0.00	0.02
Fishermen	Household Type	0.00	0.75	0.01	0.00	0.04
Extractivist	Household Type	0.00	0.88	0.00	0.00	0.02
Quilombola Family	Household Type	0.00	0.79	0.01	0.00	0.04
Bolsa Família-assisted Family	Household Type	0.00	1.00	0.61	0.62	0.19
Family of Prisoner	Household Type	0.00	0.10	0.00	0.00	0.00
Farmer	Household Type	0.00	1.00	0.08	0.00	0.17
Land Reform Benefited	Household Type	0.00	0.50	0.01	0.00	0.04
Ribeirinha	Household Type	0.00	0.62	0.00	0.00	0.03
Indigenous Family	Household Type	0.00	0.99	0.01	0.00	0.06
PNCF Family	Household Type	0.00	0.20	0.00	0.00	0.00
Gargage Collector	Household Type	0.00	0.25	0.00	0.00	0.01
Gypsy Family	Household Type	0.00	0.20	0.00	0.00	0.00
Comunidade de Terreiro	Household Type	0.00	0.11	0.00	0.00	0.00
Family Harmed By Construction	Household Type	0.00	0.42	0.00	0.00	0.01
Has Worked Last Week	Working Conditions	0.00	1.00	0.27	0.26	0.14
Away From Work	Working Conditions	0.00	0.76	0.02	0.00	0.05
Employed in Agriculture	Working Conditions	0.00	1.00	0.21	0.11	0.24
Has Worked Last 12 Months	Working Conditions	0.00	1.00	0.30	0.29	0.15
Months Worked Of Last 12	Working Conditions	0.00	12.00	5.98	5.93	2.69
Self-Employed	Working Conditions	0.00	1.00	0.12	0.11	0.09
Regular Employment	Working Conditions	0.00	0.69	0.05	0.03	0.06
Regular Houseworker	Working Conditions	0.00	0.29	0.00	0.00	0.01
Irregular Houseworker	Working Conditions	0.00	0.29	0.01	0.00	0.02
Employer	Working Conditions	0.00	0.09	0.00	0.00	0.00
Intern	Working Conditions	0.00	0.20	0.00	0.00	0.01
Military or Public Employee	Working Conditions	0.00	0.62	0.02	0.01	0.03
Unpaid Worker	Working Conditions	0.00	0.64	0.01	0.00	0.05
Temp Rural Worker	Working Conditions	0.00	0.81	0.04	0.00	0.09
Trainee/Aprentice	Working Conditions	0.00	0.09	0.00	0.00	0.00
Irregular Employment	Working Conditions	0.00	0.30	0.02	0.01	0.03
Age	–	19.62	35.75	26.52	26.49	1.02
Disability	–	0.00	0.33	0.03	0.03	0.03