# Feature Selection for Clustering of Homicide Rates in the Brazilian State of Goiás

**Samuel Bruno da Silva Sousa**
Federal University of São Paulo, Institute of Science and Technology,
São José dos Campos, Brazil, 12247-014
*samuel.bruno@unifesp.br*

**Ronaldo de Castro Del-Fiaco**
State University of Goiás, Campus Henrique Santillo,
Anápolis, Brazil, 75132-903
*ronaldo.delfiaco@ueg.br*

**Lilian Berton**
Federal University of São Paulo, Institute of Science and Technology,
São José dos Campos, Brazil, 12247-014
*lberton@unifesp.br*

**Abstract**

Homicide is recognized as one of the most violent types of crime. In some countries, it is a hard problem to tackle due to its high occurrence and the lack of research on it. In Brazil, this problem is even more complex, since this country is responsible for about 10% of the homicides in the world. Some Brazilian states suffer from the rise of homicide numbers, like the state of Goiás, in which its homicide rate increased from 24.5 cases per 100,000 inhabitants in 2002 to 42.6 cases per 100,000 inhabitants in 2014, becoming one of the five most violent states of Brazil at the end of this period, despite of having few population compared to other Brazilian states. This paper aims at applying clustering algorithms and feature selection models on data concerning homicides and socio-economic variables in the state of Goiás. We employed three clustering algorithms: K-Means, Density-based, and Hierarchical; as well as two feature selection models: Univariate Selection and Feature Importance. Our results indicate that homicide cases are more recurrent in large urban centers, although these cities have the best socio-economic indicators. Population and the educational level of the adult population were the variables which most influenced the results. K-Means clustering brought the optimum outcomes, and Univariate Selection better selected attributes of the database.

**Keywords:** Clustering, K-Means, Homicides, Machine Learning, Feature Selection.

## 1 Introduction

Violent behaviors are commonly noticed in human history, and even in the current days, it is not difficult to see news concerning violence in both physical and psychological forms. Cases of terrorism, homicides, robbery, among other forms of violent behaviors, change the way the people interact with the place in which they live [1]. Violent acts are noticed within households, communities, neighborhoods, and other social agglomerations, but, when human beings rise against the lives of each other, a major social problem emerges. Homicide is one of the most violent forms of crime and has become a major social problem in Brazil [2], since this is the country with the highest homicide rates in the world.

Homicides as recognized as security and public health problems. As a public health problem, it results in administrative expenses with hospitalizations, medical treatments, medicines, exams, among others; and as a public security problem, there are losses of human resources and expenses with poling and social security [2]. Excluding war zones, the count of homicides by year is low in most parts of the world. In 2017, the

global homicide rate per 100,000 inhabitants was 6.2. 78% of the victims were male, and the countries with the highest homicide cases by groups of 100,000 inhabitants were Honduras (85.5), Venezuela (53.7), and Belize (44.7). On the other hand, Monaco, Liechtenstein, and Andorra did not register any case of homicide which made their rates per 100,000 inhabitants reach the value 0. Most of the countries on the top of the ranking of homicide rates are in Latin America or the Caribbean [3]. These regions historically registered many homicide cases, due to drug trafficking, corruption in governments, and cartels operations [4].

In absolute numbers, Brazil is the world leader on killings. In 2018, the country broke its record for murder cases when 63,880 people were killed in just one year [5]. About 10% of overall homicides in the world happen in Brazil [4]. This country is also responsible for most of the killings of transgender people [6] and the fifth most elevated number of female deaths by hate crimes, being surpassed only by El Salvador, Colombia, Guatemala, and Russia [7]. The homicide rates per 100,000 inhabitants increased in some Brazilian regions in the first two decades of the twenty-first century while decreased in others. In 2002, the national rate was 27.9 per 100,000, while in 2014 it hit 29.7 per 100,000 [3]. The main reasons for this rise are related to drug trafficking, corruption, and lack of policing.

In the Brazilian state of São Paulo, the decrease in the homicide rates is a result, overall, of gun control policies [8]. In the state of Goiás, otherwise, the homicide rate increased from 24.5 per 100,000 in 2002 to 42.6 per 100,000 in 2014, totaling 24,300 victims in the period and becoming one of the most violent Brazilian states [3]. This state surrounds the Brazilian capital, Brasília, and had an estimated population at around 6,7 million people in 2017 [9]. Its largest city is the state capital, Goiânia, with 1,4 million inhabitants. The state is formed by 246 municipalities whose social-economic features vary a lot from one to another. In 2014, Goiás concentrated the fifth highest homicide rate amongst the Brazilian states, although having a low population.

Due to this scenario in the rates of homicides in Goiás, some questions appear, such as:

- Which variables present correlations to the increase of homicide cases in this state?
- How can Machine Learning (ML) and Artificial Intelligence (AI) algorithms be useful to understand and model the occurrence of homicides?
- Which works have already been proposed in this topic of research?
- Since ML models are built, what can be done to improve them?

To address the questions above, we conducted the present study.

The homicides in Brazil are a research topic mainly for medicine and social sciences. Few works were done using ML. Recently published studies in Brazil analyzed homicide and suicide mortality rates in different states [10, 11, 12, 13]. In our previous work [14], we have used clustering algorithms from unsupervised ML to stratify the state areas, according to the occurrence of killings in its municipalities between 2002 and 2014. Other approaches also include the employment of the Random Forest algorithm to predict killing cases [15].

Organizing data into groups is the most popular task of unsupervised ML. Cluster analysis can be defined as "the formal study of methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity, without prior knowledge of the number of clusters or any other information about their composition" [16]. There are several clustering algorithms, such as K-Means, Density-Based Clustering (DBSCAN), and Hierarchical Clustering, each of these implements a different strategy. Cluster analysis does not use category labels that tag objects with prior identifiers, such as class labels, which distinguishes unsupervised data ML from supervised ML [16]. Nevertheless, to measure the influence of each attribute of the database in this task outcomes can be a problem if they are numerous or if the information held by them is small [17].

Feature selection is a task of ML pipeline used to choose the most meaningful attributes of a database. There are two kinds of methods for this task, according to the number of variables they handle: univariate models and multivariate models [18]. The former considers the variables one by one, as they do not have any relations of dependency or correlation. Otherwise, the latter uses sub-samples or groups of variables jointly to measure how informative they are. As examples of these classes of methods, there are: Univariate Selection technique based on the chi-square test and the multivariate method of Feature Importance. In this work, we used both approaches to measure the influence of each socio-economic variable on the results of K-Means clustering, since it was the best in our experiments.

This paper aims at extending the experiments of our previous work [14], applying feature selection algorithms to measure the influence of each attribute on clustering results. We used the groups found by K-Means with K=4 as input to feature selection methods. As results, the variables related to population, educational level of adult population, and homicide rates themselves were the most influential on K-Means results. The main contributions of this paper are: 1) evaluate the correlations between homicide and socio-demographic variables for Goiás state; 2) employ three clustering algorithms (K-Means, Hierarchical and

Density-based clustering) to cluster the cities and to identify critical areas for homicides; 3) present the map which depicts the pattern of homicides distribution in the state of Goiás with K-Means algorithm, in which the highest homicide rates are presented in cities neighbors of large urban centers, like Brasília and Goiânia, despite these cities have the highest per capita income in the country. 4) measure how much each social-economic variable influenced on clustering outcomes by way univariate and multivariate algorithms for feature selection analysis.

The article structure is organized as follows: Section 2 presents related works concerning the use of ML and statistical approaches in the research of patterns in crime data. Section 3 describes the objectives, the data, the clustering algorithms, and the feature selection methods in details. Section 4 presents the results of our study and its implications for the discussion on crime analysis. Finally, Section 5 presents the final remarks and the perspectives to be followed in the future.

## 2 Related Works

The literature of crime studies based on ML approaches comprises works developed during the last 20 years. These works can be divided into two main groups: the ones which employ unsupervised ML approaches and the ones which employ supervised ML approaches. There is also a third group which comprises the works done without any approach of ML. Usually, the papers in the last group describe the use of statistics and mathematical tools. The difference between supervised and unsupervised ML is defined according to the feedback given by a human during the learning process and the nature of the data set [19]. Given a data point $x_i \in X$, if there is assigned to it a value $y_i \in Y$, where $X$ represents the whole data set, and $Y$ is the set of labels, then there is a supervised approach of ML. Otherwise, if there is no $y_i$ value, that is an unsupervised ML method. In supervised ML, the human feedback is done in the form of the label $y_i$ and the knowledge used to validate the algorithms, while in unsupervised ML no previous knowledge about the data is required. Among the unsupervised tasks of ML, it highlights the cluster analysis and the mining of association rules; and among the supervised tasks of ML, there are the tasks of classification and regression.

The three groups of related works are presented in the following subsections. The Subsection 2.1 summarizes the papers found in the literature which makes use of unsupervised ML algorithms. The subsequent subsection (2.2) describes the works that employed supervised ML, and the last one (2.3) gives an overview of the articles which instead of using ML algorithms used common statistical approaches and tools.

### 2.1 Unsupervised Approaches

This subsection describes the works which have employed algorithms from unsupervised tasks of ML. The most frequent approach found in these papers was clustering, and the most used algorithms were K-Means and Hierarchical Clustering. This can be due to the facility at applying them on data without previous knowledge and the interesting findings they can easily lead to. Many of these papers lack of validation measures to evaluate ML algorithms outcomes, such as Silhouette index and Gap statistic.

Nath [20] used an enhanced implementation of K-Means clustering to detect patterns on crime data and speed up the process of solving crimes. The data were obtained from a sheriff's office, and the feature selection was done manually by detectives, which gave importance to each attribute in the data set, according to their previous knowledge of crimes. K-Means was running over the data set and produced 6 clusters, each of these groups was related to a different scenario of crimes occurrence. All of them were validated by detectives as well. Some limitations were found by the author, such as the influence of the data quality, with missing attributes, noise, and outliers, as well as the need for human interpretation for the results.

Chandra et al. [21] applied Hierarchical Clustering with single-linkage method over data from Indian police for finding similar crime trends. The crime data contained records of seven kinds of crime, such as murder, attempt to murder, kidnapping, assault, hurt, etc. These data were collected from 29 districts of an Indian state. After applying the clustering algorithm, it was formed five clusters with similar crime trends. No validation measure was computed, and the approach was intended to handle time series clustering problems which rely on weights for each dimension of the data set.

Argawal et al. [22] used K-Means clustering to analyze a crime data set collected from the Indian government. They have found five groups in the data, and few analysis was done on the results. The authors pointed out that clustering techniques are useful for detecting peaks in the distribution of homicide occurrence, as well as designing solutions intended to predict future cases. Nevertheless, clustering algorithms are not meant to predict future occurrences of a variable. Common approaches intended to do so are based on supervised approaches of ML, such as Alves et al. [15].

In our previous work [14], we have conducted an exploratory study concerning the rise in the number of killings in the Brazilian state of Goiás between 2002 and 2014, applying three clustering algorithms (K-Means, DBSCAN, and Hierarchical Clustering) on a data set collected from government agencies and the

United Nations Program for Development's website. In this study K-Means has brought the best result, finding four clusters related to different scenarios of homicides distribution. To validate the results, three measures of how well the data within the clusters are grouped were used. These measures are Silhouette index, the sum of the squares within, and the Gap statistic. Cities close to large metropolis tend to present the most elevated numbers of homicide, despite having the highest income per capita. Among the work's limitations, the authors highlight the interpretation of results relying on human perception [23].

### 2.2 Supervised Approaches

Predicting values of a variable is a common approach for regression, one of the main tasks of supervised ML. As well as in other domains, regression can be successfully used to predict crime occurrence. In this subsection, we summarize some papers which performed this task. It is possible to notice that Random Forest (RF) algorithms are popular in the literature and have brought high values for performance measures of supervised ML.

Bogomolov et al. [24] presented a novel approach to predict crime using multiple data sources as a mobile phone and demographic data. The data set comprised anonymized data collected from mobile network activity in the London area. The forecasting problem was handled as a classification task for which several algorithms were trained, such as logistic regression, support vector machines, neural networks, decision trees, and different kinds of ensembles of tree classifiers with different parameters. The authors also have used PCA to exclude features from the data set, ranking the features by Gini coefficient values. The model which yielded the best performance was decision tree classifier based on the Breiman's RF algorithm. This work showed that the usage of human behavioral data improves prediction accuracy when compared to the use of households census or demographic data.

Alves et al. [15] used an RF regressor to predict crime based on urban indicators of homicide cases. The data set were obtained from the Brazilian Public Health System — DATASUS — and the Brazilian National Census that took place in 2000. The model yielded 97% of accuracy on crime prediction, and the features provided to the algorithm were clustered in groups. Unemployment and illiteracy were assigned as the most critical variables for describing homicides in Brazilian cities. This work was intended to produce conclusions concerning the influence of urban indicators on crime rates and to help authorities from the government to implement policies to control the criminality.

### 2.3 Statistical Approaches

Several works have been developed on homicide rate analysis since the last half of the 20th century. The fields of sociology and criminology were the first to start researching this theme. The main purposes of those works were related to investigating whether demographic, economic, ecological, and social variables maintained some correlations to the variation in homicide rates across time and space [25]. Variables as resident racial segregation, racial inequality, extreme poverty, social capital, and unemployment rate were used for some well-succeeded findings [26].

Larsen et al. [27] used scan spatial statistics to analyze clusters of the gunshot occurrences within the city of Syracuse, New York, since in the United States of America, gunshot violence is responsible for about 34,000 deaths annually. Amongst the results, it was noticed that the highest violence rate was related to environmental and economic disparities.

In Central America, where the homicide rates are historically elevated, some works have been developed on the analysis of possible causes. Carcah [28] purposed that in El Salvador, the clusters of homicides may be related to drug trafficking and organized crime. And in Mexico, Gonález-Pérez [29] explained that the spatial variation of homicides was linked to firearms possession, drug trafficking and social exclusion.

Some papers analyzed mortality by homicide in Brazilian states. Souza et al. [10] take an ecological approach to the situation of homicides in all the municipalities of Bahia for the male population aged 15 to 39, considering the health macro-region. Lozada et al. [11] analyzed the homicide mortality trend from 1979 to 2005 for males aged 15 to 49, living in the State of Paraná. Souza et al. [12] made an ecological study which analyzed the violence-related death records of women aged 10 years and older, in the Brazilian geographic regions, between 1980 and 2014. Bando and Lester [13] evaluated correlations between suicide, homicide and socio-demographic variables by an ecological study for the Brazilian states. Peres et al. [8] described homicide mortality in the municipality of São Paulo according to social-economic features of the victims and type of weapon.

The techniques and tools commonly used on this topic are: estimation of regression coefficients [25, 26]; spatial scan statistics [27]; SaTScan software methods [30]; Bayesian approach with Monte Carlo Markov Chain algorithm [28]; Moran's Global index [10]; descriptive statistics or correlation techniques [11, 13];

estimable functions and negative binomial regression [12]; SSPS software tools [8]; and multiple regression analysis (stepwise method) [29].

## 3 Materials and Methods

Since the use of unsupervised ML to analyze data from crime occurrences is a topic which lacks formalism in the evaluation of the results due to works that did not employed validation metrics and bring few analysis on their results [21, 22], there is space in the literature for new works which employ validation measures and better discuss their findings. Another reason to analyze crime data, especially concerning homicides, is the high number of deaths in Brazil year by year that generates expenses with public health, policing, justice, social security, among others [2].

This section describes the general and specific objectives of the study (3.1), the clustering algorithms (3.2), the data sets used (3.4) and the evaluation measure employed in the experiments (3.5)

### 3.1 Objectives

We aim at applying three clustering algorithms (K-Means, DBSCAN, and Hierarchical clustering) over crime data concerning homicides in the Brazilian state of Goiás to select the one which better fits to exploiting meaningful patterns in the data. To decide which algorithm brings the best performance, we compute three evaluation metrics for clustering (silhouette index, some of the squares within, and Gap statistic). After this, we measured the influence of each attribute on clustering algorithms outcomes, using the feature selection methods of Univariate Selection and Feature Importance. As far as we know no paper has done so for data from the state of Goiás. This is also an empirical study of homicide mortality in the State of Goiás for the period from 2002 to 2014. This period was chosen in function of data availability and justified by the enormous rise in the number of homicides registered in it.

The specific objectives are:

- Collect data about mortality from homicide rates (MHR) in Goiás from 2002 to 2014.
- Apply cluster algorithms to group similar data, specifically, K-Means, Hierarchical and Density-based algorithms.
- Extend the results on the identification of patterns in the clustering results to delimit dangerous areas.
- Plot a map with the risk areas identified by the clustering algorithms.
- Based on validation metrics, define which algorithm has better fixed to handling with crime data.
- Describe the influence of each feature of the data set on the algorithms outcomes.

### 3.2 Clustering Algorithms

The presence of clusters was evaluated by K-Means, Hierarchical and Density-based algorithms. The software used for building and pre-processing the database and for computing the clustering methods was R.[1]

#### *3.2.1 K-Means Clustering*

This algorithm aims to partition $n$ data instances into K clusters in which each instance belongs to the cluster with the nearest mean, serving as a cluster's prototype [31]. This is a very popular algorithm in data mining. It proceeds by alternating between two steps, given an initial set of K centroids $m_1$, $m_2$... $m_n$. The first step comprehends the assignment of each instance to the cluster with the least squared Euclidean distance. In the second step, new means are calculated to be the centroids of the new clusters. During the iterative process, it is common that some instances skip from one cluster to another. The end of the iterations occurs when the instances are established and, at this time, the K-Means are indeed the means of the groups formed [31].

Let $\mu_k$ be the centroid of cluster $c_k$. The squared error between $\mu_k$ and the points in cluster $c_k$ is defined as in Equation 1.

$$G(c_k) = \sum_{\vec{x_i} \in c_k} ||\vec{x_i} - \mu_k||^2 \tag{1}$$

[1] https://www.r-project.org/

The goal of K-Means is to minimize the sum of the squared error overall K clusters, as defined by Equation 2.

$$G(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \tag{2}$$

Automatically determining the number of clusters is a difficult problem in data clustering. K-Means is run independently for different values of K and the partition that appears the most meaningful to the domain expert is selected.

#### *3.2.2 Hierarchical Clustering*

This is a method for cluster analysis which seeks to build a hierarchy of clusters [32]. Its representation consists of a hierarchical system (dendrogram) like the taxonomic hierarchy in biology. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a "bottom up" approach where each observation starts in its cluster and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top-down" approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Formally, it is assumed that when there is a sequence of $n$ elements, represented by values from 1 to $n$, there is also a sequence of clusters with length $m+1$ ($C_0$, $C_1$, $C_2$, ..., $C_m$) [32]. Each cluster has a number $\alpha_i$ with its value. The cluster $C_0$ is considered the weakest clustering of $n$ elements ($\alpha_0 = 0$), and the cluster $C_m$ is considered the strongest clustering. The numbers $\alpha_i$ increase ($\alpha_i \leq \alpha_{i+1}$) as well as the clusters $C_i$ ($C_i \leq C_{i+1}$), which means each cluster $C_i$ is the merging (or union) of clusters $C_{i-1}$ [32]. In general, the merges and splits are determined greedily. The results of hierarchical clustering are usually presented in a dendrogram. The most common algorithms are single-linkage and complete-linkage clustering.

#### *3.2.3 Density-Based Clustering*

DBSCAN is one of the most common density-based clustering algorithms. It was designed to discover clusters and noise in a spatial data set, requiring only one parameter for input whose value is determined by the user [33]. This algorithm works as follows: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

The DBSCAN algorithm can be abstracted into the following steps:

- Find the $\epsilon$ neighbors of every point and identify the core points with more than a predefined number of neighbors.
- Find the connected components of core points on the neighbor graph, ignoring all non-core points.
- Assign each non-core point to a nearby cluster if the cluster is an $\epsilon$ neighbor, otherwise, signalize it as noise.

DBSCAN is an effective algorithm to discover clusters that present unusual shapes even in large spatial databases [33].

### 3.3 Feature Selection Methods

Selecting the most relevant features of a data set is needed when they are numerous or do not have enough information, leading ML models to be slow or inaccurate [18]. To perform this task, several classes of methods have been proposed, such as univariate and multivariate methods. Univariate methods ignore the possibility of relations among the attributes, such as correlation. Usually, this kind of feature selection techniques uses statistical tests as a tool to assign scores to each attribute in the database. Multivariate methods, otherwise, use samples or subsets of features joined to measure their importance.

A common univariate technique for feature selection is selecting the attributes, according to their scores for the chi-squared test ($X^2$), in the form:

$$X^2 = \sum_{i=1}^{n} \frac{(O_1 - E_i)^2}{E_i}, \tag{3}$$

where $E_i$ is the expected number of observations in the group $i$, if there is no relationship between the feature and the target; and $O_i$ represents the number of observations in $i$. The insight took into account to use this test is the assumption of independence between the features and the group target. So for each attribute in the database, chi-squared is computed, and the ones with the highest values for $X^2$ scores are selected.

Among the multivariate methods, Feature Importance is a very popular algorithm based on bagged decision trees [34] to estimate how meaningful an attribute is. The algorithms work building totally randomized trees which their structures are independent of the outcome values of the sample selected for the learning process. On the algorithm outputs, there are scores for each feature of the database. The highest ones are selected.

We have compared the results of both methods to select the most relevant features which influenced the results of the best clustering algorithm for our work. The analysis of feature selection results is presented in Section 4.

### 3.4 Data Set

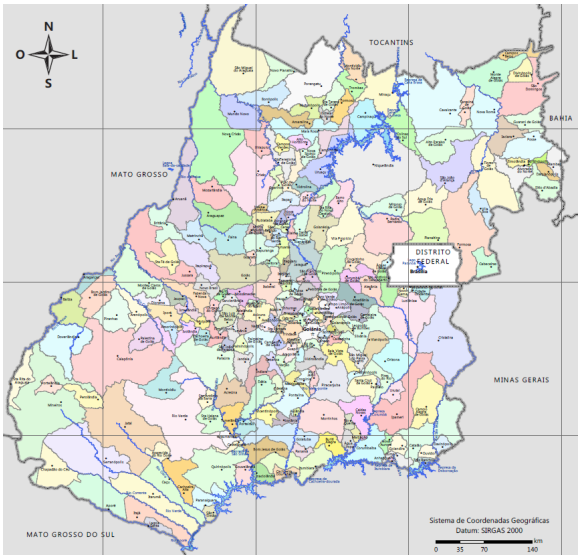



Figure 1: Map of the Municipalities of the State of Goiás Adapted from Mauro Borges Institute

Data from 246 municipalities of the state of Goiás (Figure 1) were used as analysis units. They were collected between January and March of 2018 by downloads from official websites of organizations maintained by the Brazilian government and the United Nations. All data used in this analysis are publicly available for download on the Internet. They are highly reliable since they are published by government agencies and by internationally recognized organizations.

The data on demographics consist of the population estimate for each municipality and its demographic density. They were extracted from the website of the Brazilian Institute of Geography and Statistics (IBGE) [9]. The values used in the query consisted of the municipalities names. Data on mortality were extracted from the Mortality Information System (SIM) [35]. This system is maintained and updated by the Brazilian Ministry of Health with data from mortality in general. In the query, we aimed to extract only the MHR, that included deaths by aggression, under the CID-10 codes X85 and Y09, and by legal intervention, under the codes Y35 and Y36. The selected values in the query were: the area of influence of the state of Goiás, the municipalities names, the period in the extraction was between the years 2002 and 2014, and the codes of CID-10 mentioned above.

Basic Education Development Index (IDEB) data [36] were extracted from the website of Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), an agency of the Brazilian Ministry of Education. This index measures the quality of basic education in Brazil, according to international standards. The values for the index range from 0 to 10. In the query, we selected to consult by: municipality, public (federal, state or municipal) administrative dependency, series corresponding to the final years of Secondary School (8th and 9th years in Brazil).

Social and economic features data were collected from the website of the United Nations Development Program (UNDP). In Brazil, UNDP develops an Atlas of development in the country, named Human Development Atlas in Brazil, which covers all the Brazilian States, the Brazilian Federal District, and its municipalities. In this study, we collected the data that comprehended the Municipal Human Development

Index (MHDI) [37] and its variables, like life expectancy, education index, the percentage of total income appropriated by the 10% richest, and Gini index. These data were available to be downloaded on a Microsoft Excel file.

Table 1 contains the name and description of all variables selected to compose the data set. For cluster analysis, the data set was pre-processed and analyzed in R. We have made use of Factoextra, Amap, FPC, Cluster, and DBSCAN packages.

Table 1: Description of the Social-Economic Variables in the Data Set

| **Variable** | **Description** |
|---|---|
| MHR | Total number of MHR since 2002 until 2014. |
| POPULATION | Population counting in IBGE 2010's census. |
| DEMOGDENSITY | Municipality population by its total area. |
| IDEB2005 | Basic Education Development Index in 2005. |
| IDEB2007 | Basic Education Development Index in 2007. |
| IDEB2009 | Basic Education Development Index in 2009. |
| IDEB2011 | Basic Education Development Index in 2011. |
| IDEB2013 | Basic Education Development Index in 2013. |
| LIFEEXPECT | Life expectation in 2010. |
| GINI | Gini coefficient in 2010. |
| INRICHEST10 | Rate of the overall income held by richest 10%. |
| EDUCLEVEL | Education level of adult population in 2010. |
| MHDI | MHDI in 2010. |
| MHDIE | MHDI (Education) in 2010. |
| MHDIL | MHDI (Longevity) in 2010. |
| MHDII | MHDI (Income) in 2010. |

### 3.5 Evaluation of Results

The process of evaluating the results obtained from a clustering algorithm is commonly called validation. There are two main types of grouping validation indexes: i) external indexes, which compare the group structure discovered with a previously known group structure; ii) internal indexes, which analyzes the structure of groups discovered concerning some criterion, such as, for example, compactness or separability.

We employ the external index Silhouette, and the internal measures Gap statistic and Sum of Squares Within (SSW) to validate the clustering results. The Silhouette index (SIL) is calculated per data, and the SIL of a group is the average of the SIL of all the data in the group. So the clustering SIL is the mean of the SIL of the groups. The higher the index value, the better. It is described by Equation 4.

$$SIL = \frac{a_i - b_i}{max\{a_i, b_i\}} \tag{4}$$

where $a_i$ is the average distance from the data $i$ to all other data in its group and $b_i$ is the minimum distance of the data $i$ to all other data that do not belong to its group.

The Gap statistic compares the overall intra-cluster variation for a different number of clusters with their expected values under a null reference data distribution. The Gap statistic for a given value $k$ is defined as;

$$Gap_n(k) = E_n^* log(W_k) - log(w_k), \tag{5}$$

where $E_n^*$ is an expectation in a sample of size $n$ from the reference distribution, obtained by bootstrapping by computing the average $log(W_k)$, and by generating $B$ copies of the data set. The Gap statistic measures the deviation of the observed $W_k$ value from its expected value under a null hypothesis.

SSW is a measure of compactness of a cluster structure, given by the distance between the centroid of a cluster and the instances assigned to it. The SSW equation has the form:

$$SSW_M = \sum_{i=1}^{N} ||x_i - c_{Pi}||^2, \tag{6}$$

where $x_i$ is an instance in each cluster whose centroid is $c_{Pi}$.

# 4 Results and Discussion

Subsection 4.1 presents the analysis of the variables and their correlation with homicide rate. Pearson's, Spearman's and Kendall's indexes were employed to measure correlation. Subsection 4.2 outlines the cluster analysis with K-Means, DBSCAN, and Hierarchical clustering. In 4.3 the results of feature selection approaches are reported. In the end, Subsection 4.4 presents the discussion of the results.

## 4.1 Socio-Economic Variables Analysis

From 2002 to 2014, 24,300 people were murdered in the State of Goiás [3]. The number of MHR did not vary a lot between 2002 and 2007, but from 2007 to 2013 it increased year-to-year and reached the highest mark in 2013 (Figure 2). In 2014, it was noticed a slight decrease in the number of MHR, but it was still the second highest record in the period.

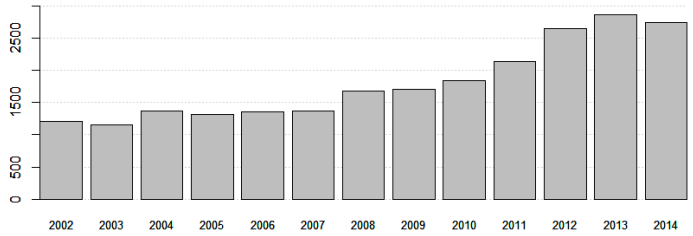



Figure 2: MHR in the State of Goiás by Year

The Gini coefficient provides a measure of income distribution among individuals or households within a territory, such as a country, a state, a municipality, etc. A value 0 represents perfect equality, and a value 100 represents total inequality. In our analysis, we used a normalized range of Gini coefficient from 0 to 1, in the way it was presented on the data source. Figure 3 presents the distribution of Gini coefficient among the municipalities in the State of Goiás. Fewer than 20 municipalities have presented values for Gini index under 0.4. Thus, it is possible to notice that the State of Goiás presents high-income disparities among its municipalities and within most of the municipalities too. For instance, the state's largest city, Goiânia, presented a Gini value of 0.58.

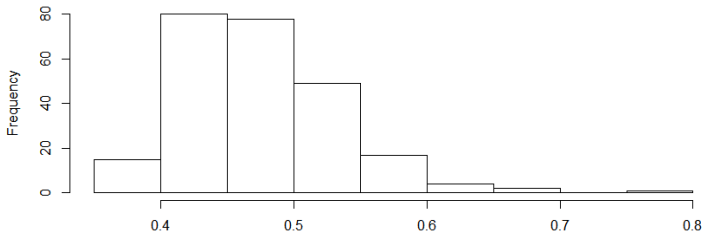



Figure 3: Distribution of Gini coefficient in Goiás (2010)

The educational variable EDUCLEVEL is related to the percent of the adult population that has completed the last year of Secondary School. In most of the municipalities, fewer than 60% of the adult population completed Secondary School (Figure 4). This situation is due to the fact of most of the population lived in rural areas until the decade of 1970 [38], when schools and universities were concentrated in the largest cities. So the rural population was aside from formal education.

The variable IDEB consists of the Basic Education Development Index. In this study, we analyzed the values for IDEB in the last years of Secondary School for public schools. The means for IDEB from 2003 to 2013, are shown in Figure 5. It is possible to notice that the values of IDEB were increasing through the years, which means improvements in the quality of public basic education. In 2013, the IDEB mean for the state reached 4.5. In this year, the municipalities with the highest IDEB means were Perolândia (6.2),

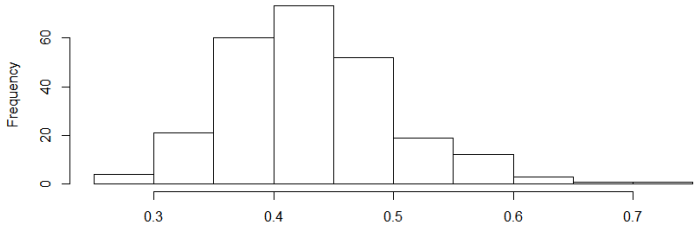



Figure 4: Distribution of Education Level of Adult Population (2010)

Córrego do Ouro (5.9), and Ivolândia (5.9), that are all small cities. On the other hand, the cities with the lowest IDEB means in 2013 were Águas Lindas de Goiás (3.4), Água Fria de Goiás (3.5), and Cavalcante (3.6), where the first two are located in the Surroundings of Federal District.


Figure 5: IDEB Variation Through the Years

In 2012, UNDP started calculating MHDI for all municipalities in Brazil. This index is calculated in the same way as HDI, which considers three dimensions (income, education, and longevity). Despite having the seventh highest HDI in Brazil (0.735) in 2010 [9], the State of Goiás presents a significant disparity in the values of HDI among its municipalities (Figure 6). More than half of the municipalities of Goiás have values for MHDI under 0.70, i.e., medium and low human development.

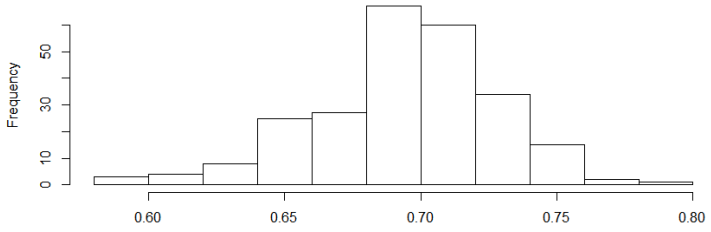



Figure 6: Distribution of Municipal Human Development Index in Goiás (2010)

Table 2 presents the Pearson's, Spearman's and Kendall's coefficient correlations between the social-economic variables considered and the MHR. The values of the correlations are between -1 and +1. The signal indicates the direction, whether the correlation is positive or negative, and the absolute value of the variable indicates the strength of the correlation. Values between 0.7 and 0.9 indicate a strong correlation. We noticed that POPULATION has a strong correlation with MHR as indicated by the three coefficients. DEMOGDENSITY only has correlation indicated by Pearson's coefficient, suggesting a linear relation between it and the variable MHR.

Before applying the cluster algorithms, we computed the distance between the data set columns (Figure 7), to visualize the (dis)similarity within the data set. It is possible to notice that values for demographic

Table 2: Correlations between the Total of MHR and Social-Economic Variables

| | Correlation Coefficients | | |
|---|---|---|---|
| **Variable** | **Pearson's c.** | **Spearman's c.** | **Kendall's c.** |
| POPULATION | **0.9915637** | **0.8932006** | **0.7385185** |
| DEMOGDENSITY | **0.7262442** | 0.3446707 | 0.2389337 |
| IDEB | -0.1440371 | -0.2019128 | -0.1417871 |
| LIFEEXPECT | 0.09612802 | 0.1669409 | 0.1119476 |
| GINI | 0.1134491 | 0.2435379 | 0.1748277 |
| INRICHEST10 | 0.1039698 | 0.2740375 | 0.1939559 |
| EDUCLEVEL | 0.3994967 | 0.4042891 | 0.287134 |
| MHDI | 0.248962 | 0.1948997 | 0.133404 |
| MHDIE | 0.206129 | 0.08632882 | 0.05902461 |
| MHDIL | 0.096134 | 0.1656033 | 0.1120611 |
| MHDII | 0.2649425 | 0.2717906 | 0.1852401 |

density, population, and MHR are close to each other. At the same time, these variables present values far from life expectation, IDEB, and MHDI values, showing to us that in the state of Goiás the population growth was not accompanied by educational improvements, and the distribution of MHR is related to the population number.

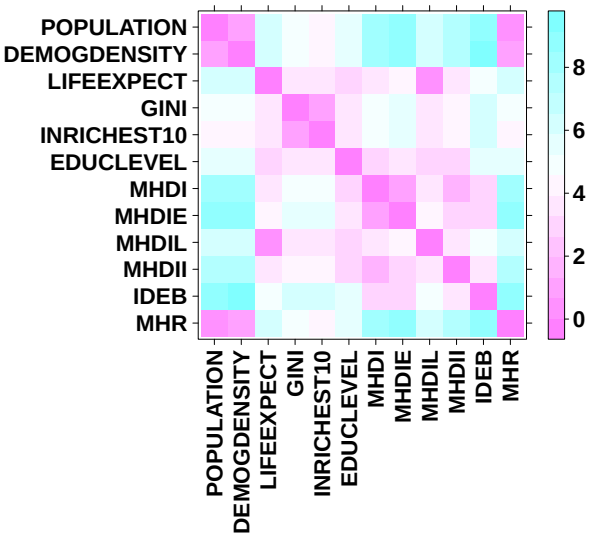



Figure 7: Data Set Distance Matrix Using Manhattan Distance (Color online)

### 4.2 Cluster Analysis

This section presents the results of applying the Hierarchical, Density-based and K-Means algorithms in the municipalities of Goiás. We validate the clustering results with the Silhouette measure, Gap statistic, and SSW.

#### *4.2.1 Hierarchical Clustering*

The hierarchical clustering algorithm, with the single-linkage method and Manhattan distance, generated clusters with few municipalities. We have cut the dendrogram into three partitions, which is shown in Figure 8. The optimal number of clusters to cut the tree was decided by the results of Silhouette index and Gap statistic. We used the value which presented the biggest difference to its successor, and the optimum Gap statistic is given by $Gap_{(k)} > Gap_{(k+1)} - S_{k+1}$, being $S_{k+1}$ the standard deviation multiplied by the square of 1 plus the inverse of B copies of the reference data set done by bootstrapping.

This algorithm assigned to the first cluster (in red) the largest city of the state, Goiânia; assigned to the second cluster (in yellow) the city with the highest demographic density, Valparaíso de Goiás; and assigned to the third cluster (in purple) the other cities in the state, despite of the dissimilarity between them.

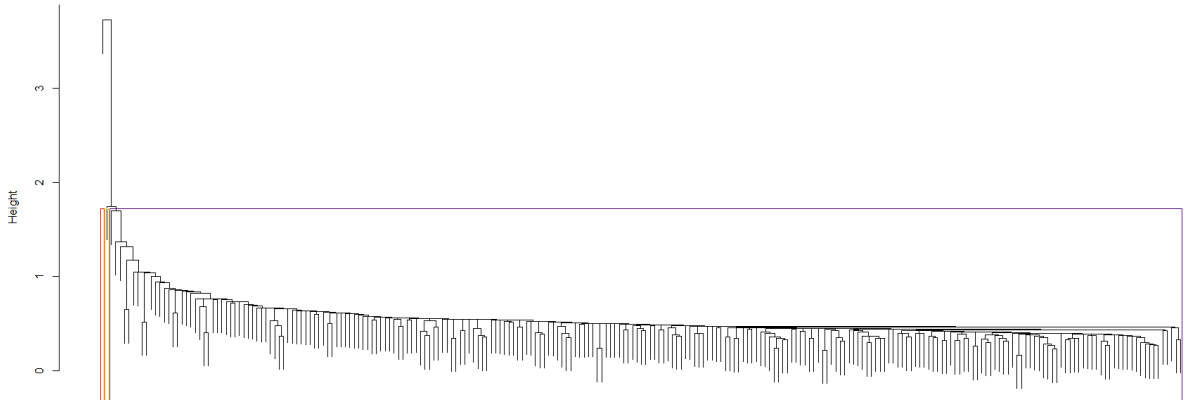



Figure 8: Results of Hierarchical Agglomerative Algorithm with Single-Linkage Method and Manhattan Distance (Color online)

### *4.2.2 Density-based Clustering*

The DBSCAN algorithm generated just one cluster and assigned as outliers 20 municipalities, amongst these outliers, there are: the largest cities of the state, the municipality with the highest demographic density, and small cities whose values for MHR are 0 or close to it. This algorithm did not generate a meaningful result, because of the dissimilarity between the instances of the data set. Most of the municipalities assigned as outliers hold the highest values for the variables population, demographic density, and MHR. The results of DBSCAN are shown in Figure 9.

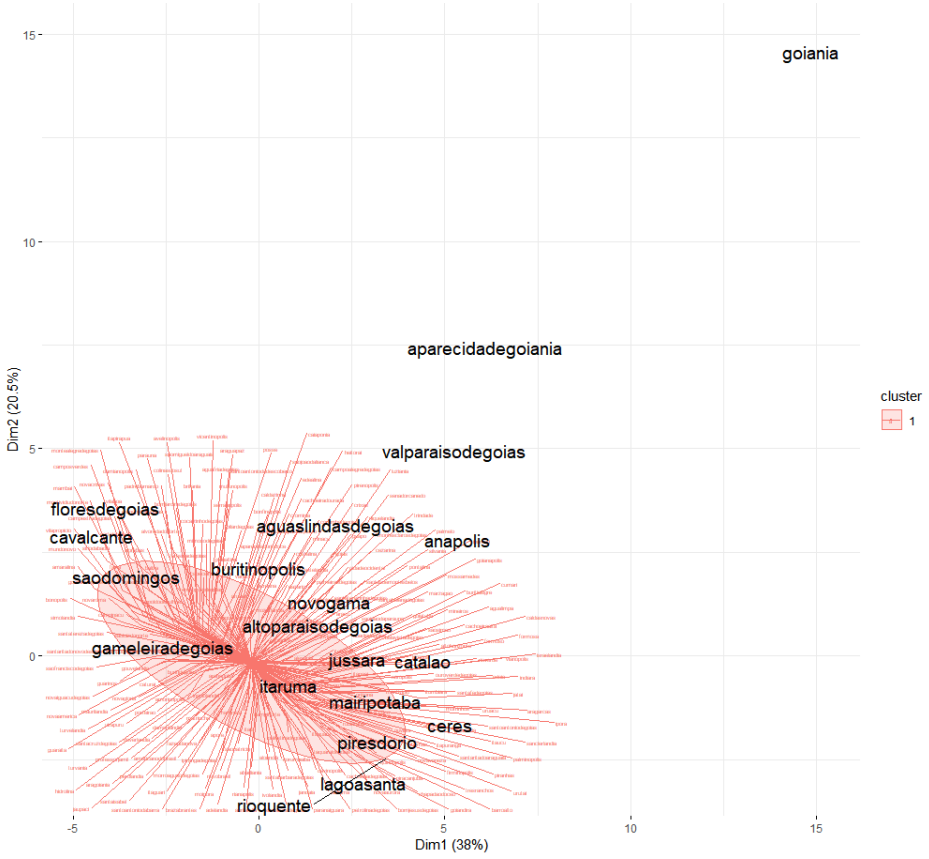



Figure 9: Results of DBSCAN (Color online)

### *4.2.3 K-Means Clustering*

We employed different K values (from 1 to 5) in the K-Means clustering algorithm. We have also employed different distance measures, such as Euclidean, Manhattan, Pearson, and Canberra. The one that provided

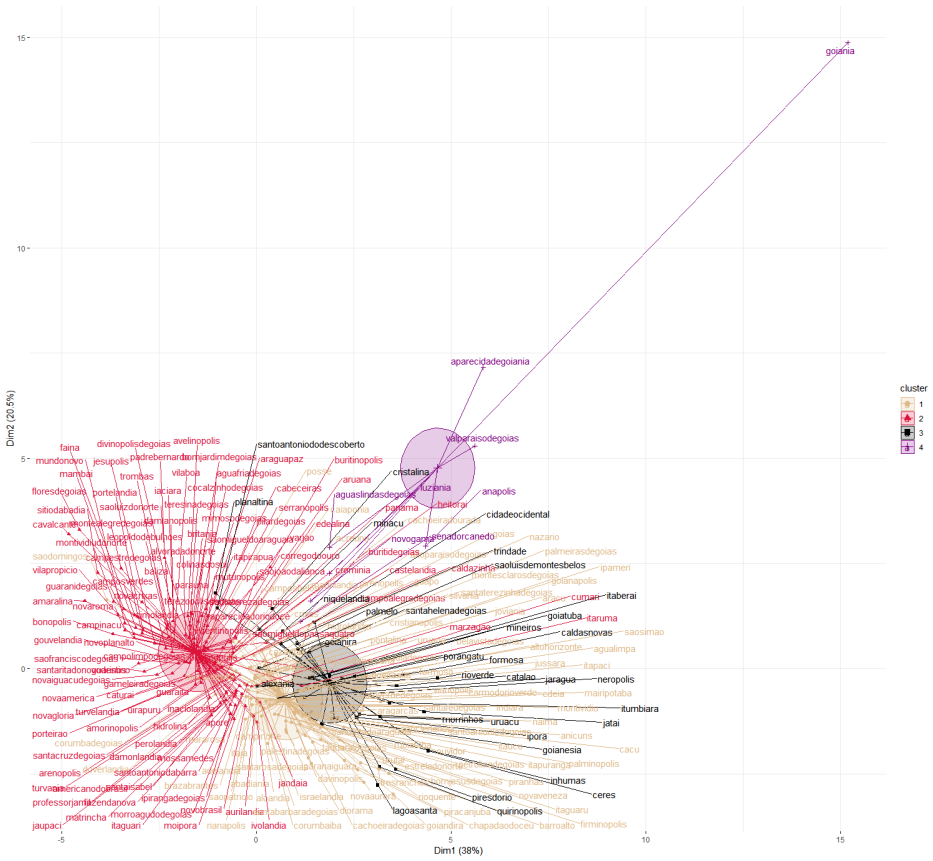



Figure 10: Results of K-Means Clustering with K = 4 (Color online)

the most meaningful set of clusters was Canberra since this measure is not sensible to outliers presence. It was decided to start K with five because the state has five regions and maybe this should influence the distribution of MHR. With K = 2, 3, and 5, the results were not interesting. In every running, the algorithm assigned Goiânia to the same cluster as small cities. The K = 4 value generated groups more interesting as shown in Figure 10. The results suggested different dynamics in the distribution of MHR among the municipalities of Goiás. The results of K-Means algorithm are shown in Figure 10.

In cluster 1 (beige), there are 96 municipalities predominantly small cities. The cluster presented the second lowest mean for the number of MHR and variables, such as population, demographic density, life expectation, Gini index, educational level of adult population, MHDI and its tree dimensions (Education, Longevity, and Income). Otherwise, this cluster has the second highest mean for income held by the richest 10% and the highest mean for IDEB index amongst all the clusters.

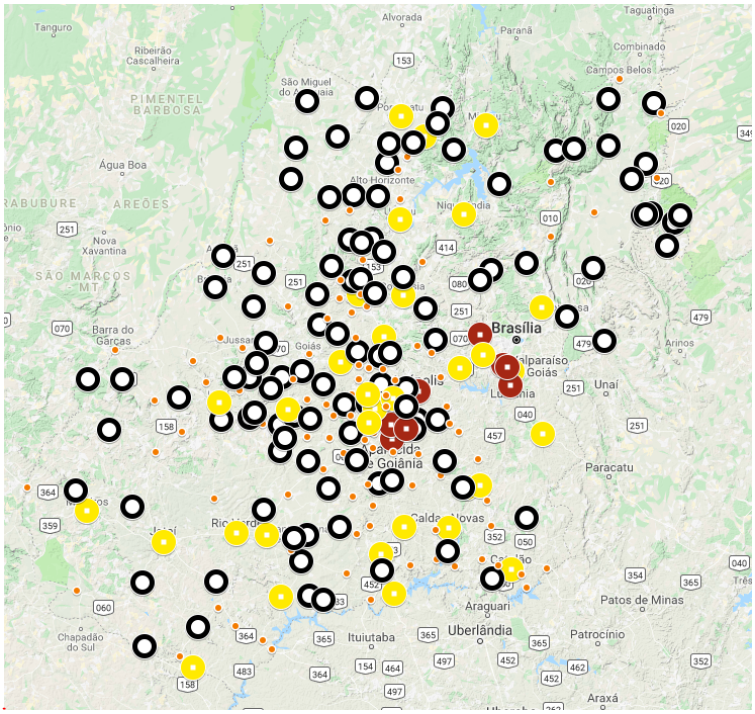
In cluster 2 (pink), there are 108 municipalities, which are also small cities close to the lowest means for population, demographic density, income held by the richest 10%, educational level of adult population, MHDI and its three dimensions and MHR amongst all the clusters. Furthermore, this cluster has the second lowest IDEB and Gini index means. As the cities in this cluster are predominantly small, far from the two metropolia in the region, the educational level of the adult population tends to be low because, as cited in Sec. 4.1, most of the state population was aside from formal education until the decade of 1970.

In cluster 3 (black), there are 33 municipalities mid-sized like Rio Verde, Cidade Ocidental, Catalão, Trindade, and Itumbiara; and some small cities. This cluster presented the highest means of life expectation, Gini index, income held by the richest 10%, MHDI, MHDI (Longevity), and MHDI (Income). Furthermore, it was noticed in this cluster the second highest means for the variables population, demographic density, educational level, MHDI (Education), IDEB, and MHR.

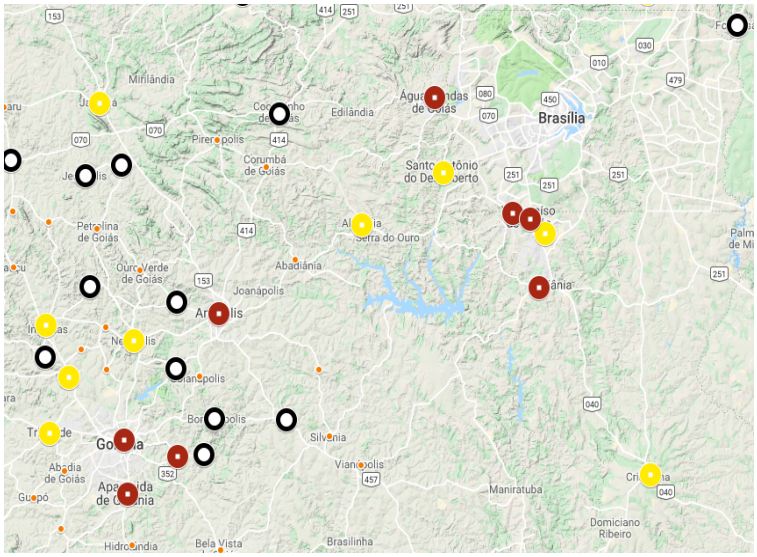
The cluster 4 (purple) aggregates only eight municipalities, all of them big or mid-sized cities. In this cluster, there are the municipalities of Goiânia, Aparecida de Goiânia, Anápolis, Valparaíso de Goiás, Águas Lindas de Goiás, Senador Canedo, Novo Gama, and Luziânia. These cities compose the region between the capital of the state of Goiás, Goiânia, and the capital of Brazil, Brasília. The means of the population, demographic density, educational level, MHDI (Education), and MHR were the highest among the 4 clusters. At the same time, the cluster presented the second highest means for life expectation, Gini index, MHDI, MHDI (Longevity), and MHDI (Income). The mean income held by the richest 10% was the third highest.

Although having the highest value for the educational level of the adult population, the index which measures the quality of basic education – IDEB – presented the lowest mean among all the clusters. Furthermore, this cluster held 59.59% of the MHR in the state from 2002 until 2014.

In Fig.11, we plotted the results of K-Means clustering. We noticed that all the cities in the fourth cluster (the one which aggregates the cities with most of the MHR) are located close to the capital of Goiás, Goiânia, or the capital of Brazil, Brasília.




(a) The Four Groups Found by K-Means with K=4 (Color online)




(b) The Fourth Group in Big Red Dots (Color online)

Figure 11: The groups formed by K-Means clustering on the map of Goiás (Color online). The little orange dots represent the first group. The big white and black dots depict the second group. The big yellow dots mark the cities in the third group, and the big red dots represent the municipalities in the fourth cluster.

### 4.3 Feature Selection Methods

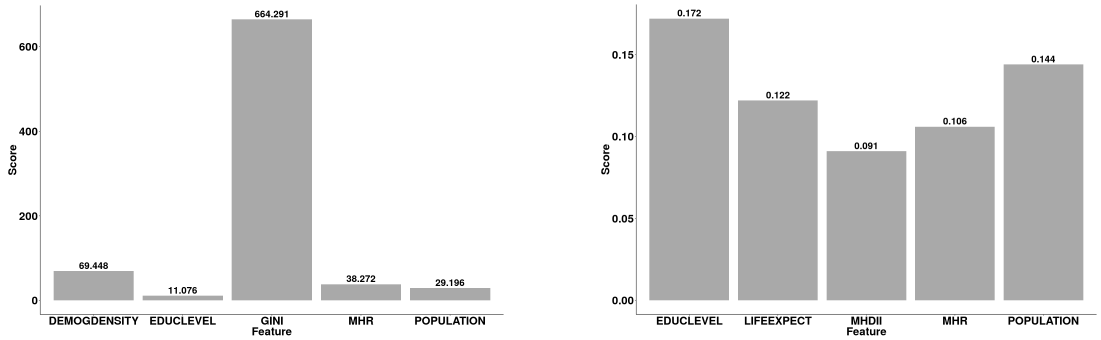
We considered the K-Means clustering with K=4 to apply feature selection methods since it was the best outcomes among the different clustering algorithms analyzed. Univariate Selection and Feature Importance were run, and the scores for each attribute on the data set are shown in Table 3. The variables with scores among the five highest are highlighted in bold. POPULATION, EDUCLEVEL, and MHR had high scores

for both methods. So we can induce that they have influenced the results more than the rest of the features.

Table 3: Feature Selection Scores for Each Data Set Attribute on K-Means Results by Univariate Selection and Feature Importance

| | **Feature Selection Methods** | |
|---|---|---|
| **Attribute** | **Univariate Selection** | **Feature Importance** |
| POPULATION | **29.196** | **0.144** |
| DEMOGDENSITY | **69.448** | 0.073 |
| IDEB | 1.884 | 0.031 |
| LIFEEXPECT | 6.663 | **0.122** |
| GINI | **664.291** | 0.027 |
| INRICHEST10 | 0.517 | 0.031 |
| EDUCLEVEL | **11.076** | **0.172** |
| MHDI | 5.127 | 0.087 |
| MHDIE | 4.176 | 0.057 |
| MHDIL | 6.781 | 0.053 |
| MHDII | 4.323 | **0.091** |
| MHR | **38.272** | **0.106** |

However, the algorithms assigned different variables as the most relevant. Univariate Selection assigned GINI, which is related to income inequality, as the most relevant. Otherwise, Feature importance assigned EDUCLEVEL, that represents the educational level of the adult population, with the highest score. This is due to the fact that the first method uses a statistical test which assumes no correlation between the variables and the second one uses a decision tree algorithm to calculate the importance of attributes. As most of the cities have Gini coefficient values (Figure 3) between 0.4 and 0.55, the instances in the clusters have Gini coefficient values close one another as well. So the chi-squared test presented a score for this variable 956.53% higher than the second best. Feature importance, which takes into account the hypothesis of correlation, did not select DEMOGDENSITY, since this feature has the same distribution as POPULATION, in other words, they are redundant. Furthermore, as shown in Table 2, GINI did not present a correlation with MHR, for this reason, Feature Importance did not select it as relevant. The five features with the highest scores assigned by the feature selection methods are shown in Figure 12.




(a) Top five features by Univariate Selection

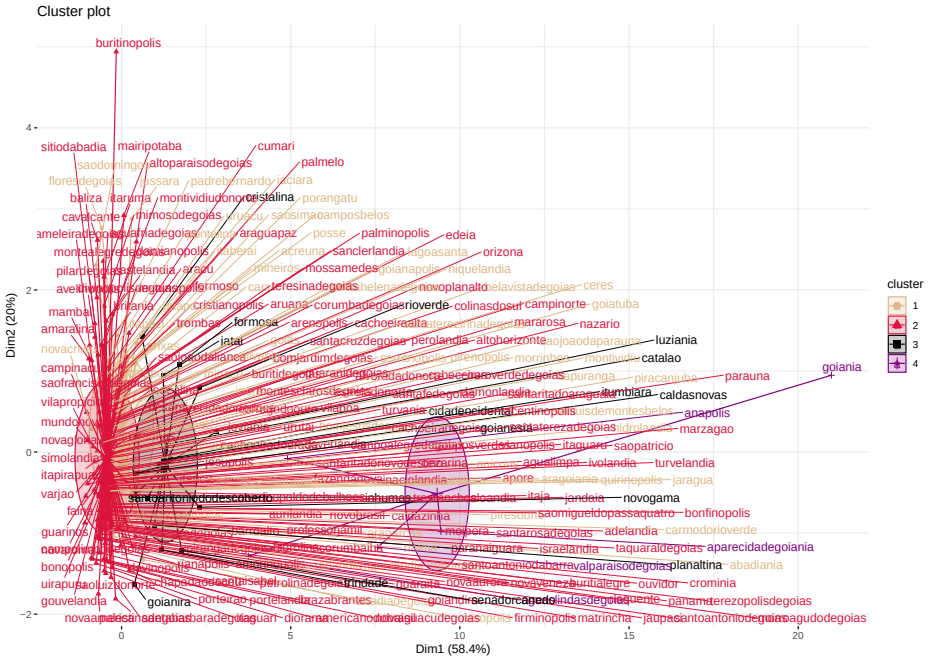
(b) Top five features by Feature Importance

Figure 12: Top Five Variables According to Feature Selection Methods

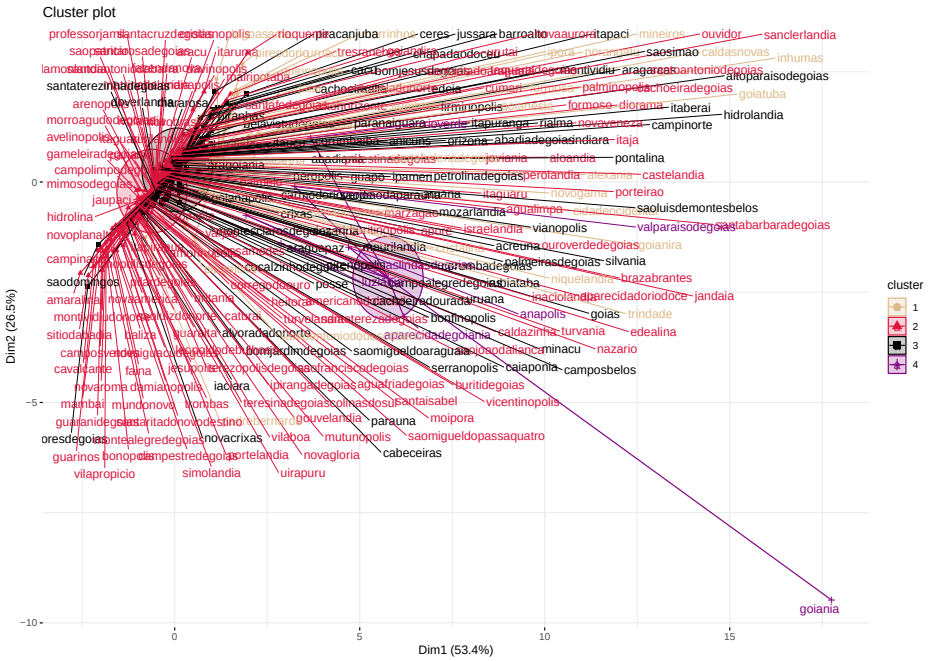
After selecting the five most relevant features with Univariate Selection and Feature Importance, we ran K-Means again. The results are shown in Figure 13. In the figure, we can notice that the features selected by Univariate Selection generated a better separated clustering (Fig. 13.a). Furthermore, the results for the attributes selected by Feature Importance (Fig. 13.b) were overlapping and inconsistent. For instance, the city of Lagoa Santa, which is a small city, was assigned to the same cluster as mid-sized cities as Novo Gama and Cidade Ocidental. In general, the results of K-Means with K=4 were enhanced after feature selection with Univariate Method. The four clusters are presented on the Table 4.

Table 4: Clusters Overview with Features Selected by Univariate Selection and Feature Importance

| Cluster | Univariate Selection | Feature Importance |
|---|---|---|
| 1 | 63 small cities with few MHR | 78 small cities with few MHR. |
| 2 | 160 smallest cities with almost 0 MHR | 131 smallest cities with almost 0 MHR. |
| 3 | 17 mid-sized cities with high GINI values | 29 mid-sized cities and some noise. |
| 4 | 5 largest cities with highest GINI values | 7 largest cities with highest number of MHR. |




(a) K-Means results with Attributes selected by Univariate Selection (Color online)




(b) K-Means results with Attributes selected by Feature Importance (Color online)

Figure 13: K-Means with K=4 after Application of Feature Selection Methods (Color online)

### 4.4 Discussion

By applying clustering algorithms, it was possible to notice that clusters of homicides in the state of Goiás are not merely random. Such clusters were considered as critical areas for homicides. It was also possible

to identify different dynamics in the distribution of MHR as well. The largest cities presented most of the homicide cases, despite its income and human development levels being higher than most of the other municipalities. On the other hand, these large cities present low values for IDEB index, suggesting that although having most people completing Elementary and Secondary School degrees, the quality of teaching in public schools is still low.

We also computed Silhouette measure, Gap statistic, and SSW to validate the clustering results (Table 5). The values in bold represent the optimum values for these measures. The optimum Silhouette value for each algorithm was assumed as the value with the biggest difference to its successor on the table. Optimum Gap statistic values were determined by selecting the lowest which satisfied the condition: $Gap_{(k)} > Gap_{(k+1)} - S_{k+1}$. The optimal values for SSW were assumed by analyzing the bend (knee) in the curves of SSW, i.e., the SSW values in bold were the bends (knees) in their respective curves of SSW.

Table 5: Scores of Silhouette, Gap, and SSW for Each Clustering Algorithm

| | **Validation Measures** | | |
|---|---|---|---|
| **Clustering algorithm** | **Silhouette** | **Gap** | **SSW** |
| K-Means (2) | 0.97 | 0.717 | 22.55 |
| K-Means (3) | 0.89 | 0.748 | 26.10 |
| K-Means (4) | **0.84** | **0.765** | **23.96** |
| K-Means (5) | 0.72 | 0.769 | 22.44 |
| Hierarchical (2) | 0.97 | 0.722 | 50.00 |
| Hierarchical (3) | **0.94** | **0.742** | 44.30 |
| Hierarchical (4) | 0.86 | 0.731 | 40.00 |
| Hierarchical (5) | 0.85 | 0.743 | **36.80** |
| Density-based | **0.97** | **0.705** | **66.71** |

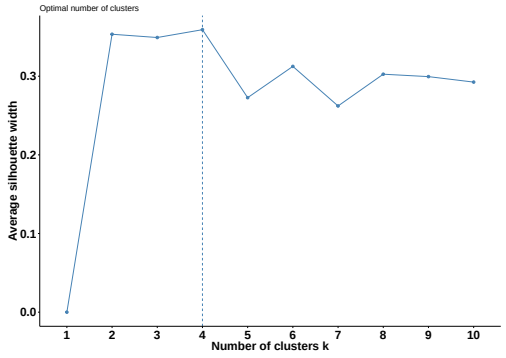
Measures such as Silhouette, Gap statistic, and SSW helps us to measure the clustering quality from a mathematical perspective. As clustering is an unsupervised task of data mining and ML, there is no label to be predicted/classified, and no training and test phases, to lately measure the error rate. So a meaningful clustering result is that one whose interpretation makes sense for a human being [23].

Hierarchical and K-Means clustering generates groups separating the large cities from small ones. The Silhouette measure indicates a high similarity between the instances clustered. However, when the number of clusters was increased, the Silhouette index values have decreased. The value 4 for K in K-Means clustering was assumed as optimal because of Gap statistic value for K = 4 is bigger than the value for K = 5, i.e., $Gap_{(k=4)} > Gap_{(k=5)} - S_{k=5}$.

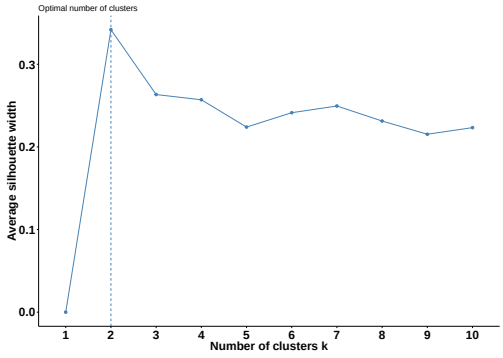
K-Means with K = 2 value achieves the highest Silhouette and the lowest SSW, but this clustering generates one group with 147 small cities and another group with 99 large cities. However, K = 4 gives a more interesting separation splitting the outliers cities, and K = 5 presents a low value for Silhouette, which suggests a weak clustering.

Density-based clustering resulted in just one cluster, and assigned as outliers 20 municipalities, among them the state largest cities. The Silhouette value for these results was also high, as well as the values for Gap statistic and SSW.

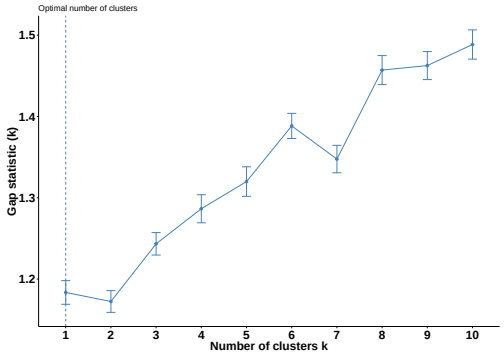
After applying the feature selection on the data set, and re-running K-Means, we also computed the three validation measures again. The results are presented in Figure 14. As noticed in the subfigures 14.a and 14.e, K-Means with K=4 run on attributes selected by Univariate Selection was the best clustering set up. Furthermore, the SSW values for this result was 73.04% less than the one presented in Table 5 for K-Means with K=4 and all attributes of the data set. Thus, this decrease in the SSW measure suggests a better clustering after feature selection.



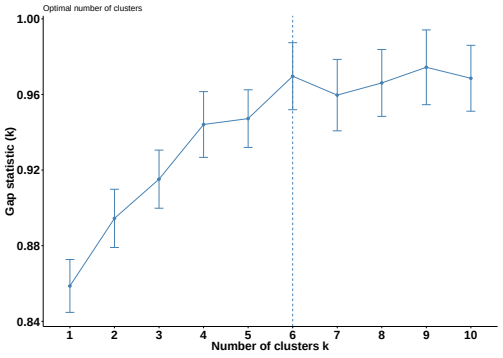

(a) Silhouette for K-Means + Univariate Selection



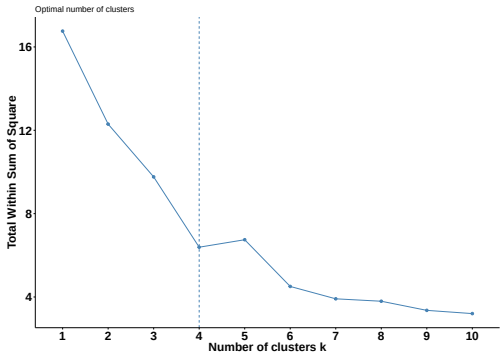

(b) Silhouette for K-Means + Feature Importance



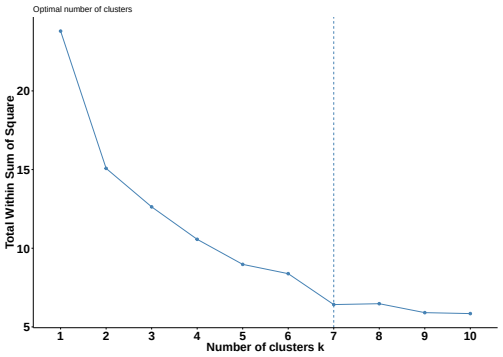

(c) Gap for K-Means + Univariate Selection




(d) Gap for K-Means + Feature Importance




(e) SSW for K-Means + Univariate Selection




(f) SSW for K-Means + Feature Importance

Figure 14: Validation Measures for K-Means Clustering after Feature Selection

Few studies bring information about all the municipalities of the state of Goiás, especially in the interior regions. Our contribution emphasizes the importance of the use of spatial analysis to understand violence indicators and geographic distribution. Another contribution is an empirical analysis of which features influenced the clustering algorithms.

## 5 Conclusion

In this paper, we employed three clustering algorithms: Hierarchical, Density-based and K-Means to analyze spatial patterns of mortality by homicide in the Brazilian state of Goiás. Given the best clustering result, we applied two feature selection algorithms on it to measure the influence of each variable on the result and select the best five. Univariate Selection provided the best results for feature selection, leading to a expressive reduction of the SSW values, suggesting better clustering outcomes.

The state's three largest cities, Goiânia, Aparecida de Goiânia, and Anápolis, had the highest numbers of MHR. The municipalities in the region known as Surroundings of the Federal District also had elevated numbers of MHR. In the other municipalities, the numbers were low. All clustering algorithms separated these regions with high MHR from the others. However, K-Means with K = 4 results in better separation.

The means for population, Gini index, MHDI, IDEB, income concentration, and educational level of adult population varied among the municipalities as well as the number of MHR. The highest IDEB values were found in the smallest cities, which presented low homicide rates. In general, we noticed three scenarios for the occurrence of MHR in the state of Goiás. The first is characterized by a low occurrence of MHR in the

small cities, as shown by the clusters 1 and 2 in Subsection 4.2. The second scenario refers to the occurrence of MHR in mid-sized cities in municipalities with low demographic density and high-income concentration, as shown by the cluster 3. The last scenario is related to the occurrence of MHR in the state's municipalities with highest values of demographic density, low values of IDEB index, although having high MHDI and income, and comprising the metropolitan areas of Goiânia and Brasília, as noticed in the fourth cluster.

The use of clustering algorithms in the analysis of the distribution of MHR was an experimental approach, once we did not find studies that did it before. In general, the behavior of the algorithms was satisfactory. Hierarchical clustering and DBSCAN provided results influenced principally by the population size, although K-Means, with K = 4, showed an interesting separation whose quality was confirmed by its Silhouette measure, Gap statistic, and SSW. The feature selection analysis showed that population, educational level of adult population, and the number of homicides itself were among the variables which most influenced the results. Additionally, we expect this study to be useful to improve the knowledge about the distribution of MHR in the State of Goiás and to aid the state's government to define risky regions, where policing should be more effective. This study is also a source to know the variation of social-economic variables among the state's municipalities. Finally, as future work other analysis employing different variables to explain the rates of homicides in specific regions could be done, like precarious socio-economic conditions, social inequality, and so on.

## Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and São Paulo Research Foundation (FAPESP) grant 2018/09465-0.

## References

[1] J. Laplante, *La violence, la peur et le crime*. University of Ottawa Press, 2001.

[2] R. Cordeiro and M. R. C. Donalisio, "Homicídios masculinos na Região Metropolitana de São Paulo entre 1979 e 1998: uma abordagem pictórica," *Cadernos de Saúde Pública*, vol. 17, pp. 669–677, 2001.

[3] "Homicide Monitor." `http://homicide.igarape.org.br`, 2016. Accessed: 2016-11-21.

[4] "Daily chart: Revisiting the world's most violent cities." `http://homicide.igarape.org.br`, 2016. Accessed: 2016-11-21.

[5] "Brazil breaks own record for number of murders in single year as deaths hit 63,880." `https://www.independent.co.uk/news/world/americas/brazil-murder-rate-record-homicides-killings-rio-de-janeiro-police-a8485656.html`, 2018. Accessed: 2019-01-05.

[6] J. S. Hutta and C. Balzer, "Identities and citizenship under construction: Historicising the 'T' in LGBT anti-violence politics in Brazil," in *Queer presences and absences*, pp. 69–90, Springer, 2013.

[7] S. N. Meneghel and A. P. Portella, "Feminicídios: conceitos, tipos e cenários," *Ciencia & saude coletiva*, vol. 22, pp. 3077–3086, 2017.

[8] M. F. T. Peres, D. Vicentin, M. B. Nery, R. S. d. Lima, E. R. d. Souza, M. Cerda, N. Cardia, and S. Adorno, "Queda dos homicídios em São Paulo, Brasil: uma análise descritiva," *Revista Panamericana de Salud Publica*, vol. 29, pp. 17–26, 2011.

[9] "Goiás - IBGE Cidade." `https://cidades.ibge.gov.br/brasil/go`, 2018. Accessed: 2018-03-30.

[10] T. O. d. Souza, L. W. Pinto, and E. R. d. Souza, "Spatial study of homicide rates in the state of Bahia, Brazil, 1996-2010," *Revista de saude publica*, vol. 48, pp. 468–477, 2014.

[11] E. M. K. d. Lozada, T. A. d. F. Mathias, S. M. d. Andrade, and T. Aidar, "Homicide mortality trend in the state of Paraná, Brazil, per Health District, 1979 to 2005," *Revista Brasileira de Epidemiologia*, vol. 12, no. 2, pp. 258–269, 2009.

[12] E. R. d. Souza, K. C. Meira, A. P. Ribeiro, J. dos Santos, R. M. Guimarães, L. F. Borges, L. V. e Oliveira, and T. C. Simões, "Homicides among women in the different Brazilian regions in the last 35 years: an analysis of age-period-birth cohort effects," *Revista Ciência & Saúde Coletiva*, vol. 22, no. 9, 2017.

[13] D. H. Bando and D. Lester, "An ecological study on suicide and homicide in Brazil," *Ciencia & saude coletiva*, vol. 19, pp. 1179–1189, 2014.

[14] S. B. da Silva Sousa, R. de Castro Del-Fiaco, and L. Berton, "Cluster analysis of homicide rates in the Brazilian state of Goiás from 2002 to 2014," in *2018 XLIV Latin American Computer Conference (CLEI)*, pp. 445–454, São Paulo, Brazil, 2018.

[15] L. G. Alves, H. V. Ribeiro, and F. A. Rodrigues, "Crime prediction through urban metrics and statistical learning," *Physica A: Statistical Mechanics and its Applications*, vol. 505, pp. 435–443, 2018.

[16] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[17] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of machine learning research*, vol. 5, no. Aug, pp. 845–889, 2004.

[18] C. Lai, M. J. Reinders, and L. Wessels, "Random subspace method for multivariate feature selection," *Pattern recognition letters*, vol. 27, no. 10, pp. 1067–1076, 2006.

[19] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[20] S. V. Nath, "Crime pattern detection using data mining," in *Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on*, pp. 41–44, IEEE, 2006.

[21] B. Chandra, M. Gupta, and M. Gupta, "A multivariate time series clustering approach for crime trends prediction," in *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pp. 892–896, IEEE, 2008.

[22] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering," *International Journal of Computer Applications*, vol. 83, no. 4, 2013.

[23] A. Carvalho, K. FACELI, A. LORENA, and J. GAMA, *Inteligência Artificial–uma abordagem de aprendizado de máquina*. LTC: Rio de Janeiro, 2011.

[24] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 427–434, ACM, 2014.

[25] K. C. Land, P. L. McCall, and L. E. Cohen, "Structural covariates of homicide rates: Are there any invariances across time and social space?," *American journal of sociology*, vol. 95, no. 4, pp. 922–963, 1990.

[26] P. L. McCall, K. C. Land, and K. F. Parker, "An empirical assessment of what we know about structural covariates of homicide rates: a return to a classic 20 years later," *Homicide Studies*, vol. 14, no. 3, pp. 219–243, 2010.

[27] D. A. Larsen, S. Lane, T. Jennings-Bey, A. Haygood-El, K. Brundage, and R. A. Rubinstein, "Spatio-temporal patterns of gun violence in Syracuse, New York 2009-2015," *PloS one*, vol. 12, no. 3, p. e0173001, 2017.

[28] C. Carcach, "A spatio-temporal model of homicide in El Salvador," *Crime science*, vol. 4, no. 1, p. 20, 2015.

[29] G. González-Pérez, M. Vega-López, C. Cabrera-Pivaral, and A. Vega-Lopez, "P1-432 Violence and health: an epidemiological analysis of homicides in MÉxico, 1979–2008," *Journal of Epidemiology & Community Health*, vol. 65, no. Suppl 1, pp. A187–A187, 2011.

[30] A. M. Zeoli, S. Grady, J. M. Pizarro, and C. Melde, "Modeling the movement of homicide by type to inform public health prevention efforts," *American journal of public health*, vol. 105, no. 10, pp. 2035–2041, 2015.

[31] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[32] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[33] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, vol. 96, pp. 226–231, 1996.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[35] "TabbNet Win32 3.0: Mortalidade - Goiás." `http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10GO.def`, 2018. Accessed: 2018-03-30.

[36] "IDEB - Resultados e Metas." `http://ideb.inep.gov.br/resultado/home.seam?cid=2098801`, 2016. Accessed: 2016-11-21.

[37] "Atlas do desenvolvimento humano no Brasil 2013." `http://www.atlasbrasil.org.br/`, 2013. Accessed: 2018-03-30.

[38] "Estado de Goiás, Centro-Oeste e Brasil: população por situação de domicílio – 1950, 1960, 1970, 1980, 1991, 2000, 2008 – 12." `http://www.imb.go.gov.br/pub/Godados/2013/01-tab03.htm`, 2018. Accessed: 2018-03-30.