Crime Inference Using Machine Learning and Geographical Data

Miguel Roque

Dissertation written under the supervision of professor Nicolò Bertani.

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, January 4th, 2023.

ABSTRACT

Crimes are not random events in society, and eventually something must influence their occurrence. It is by characterizing the environment that it is possible to create algorithms that predict the criminal activity in a certain place and at some point in time, which allows its anticipation and prevention through decision-making in public policy.

This study focusses on finding the best way to predict crimes, that is, which types of features are the most important to consider while predicting crimes, and which methods are the most predictive.

An analysis of the city of Philadelphia, in the state of Pennsylvania (USA), is made, taking into account the urban, racial, demographic and socioeconomic characteristics of its different geographical blocks, and the number of criminal occurrences in each of them, over multiple years. The methods used are both linear and non-linear.

When non-linear methods are used, via machine learning techniques, it is evident that the prediction of the number of crimes is much more assertive for any type of variable, leading to the conclusion that the relationships studied here are not linear in nature, and therefore tree-based models (especially gradient boosting and random forest) represent the most suitable approach for this data. In this perspective, the models that consider only the socio-demographic characteristics of the neighborhoods are significantly more effective in forecasting than the entirely urban ones.

- **Title**: Crime Inference Using Machine Learning and Geographical Data
- **Author**: Miguel Francisco Frade Roque
- **Keywords**: crimes; socio-demographic; urban; linear; non-linear.

ABSTRATO

Os crimes não são eventos aleatórios na sociedade e, eventualmente, algo deve influenciar a sua ocorrência. É pela caracterização do ambiente que é possível criar algoritmos que preveem a atividade criminosa num determinado local e em algum momento no tempo, o que permite a sua antecipação e prevenção por meio das tomadas de decisão na política pública.

Este estudo foca-se em encontrar a melhor forma de prever crimes, ou seja, que tipos de características são as mais importantes a considerar na previsão de crimes, e que métodos são os mais preditivos.

É feita uma análise da cidade de Filadélfia, no estado da Pensilvânia (EUA), tendo em consideração as características urbanas, raciais, demográficas e socioeconómicas dos seus diferentes quarteirões geográficos, e o número de ocorrências criminais em cada um deles, ao longo de vários anos. Os métodos utilizados são lineares e não lineares.

Quando são utilizados métodos não lineares, através de técnicas de machine learning, fica evidente que a previsão do número de crimes é muito mais assertiva para qualquer tipo de variável, levando à conclusão de que as relações aqui estudadas não são de natureza linear e, portanto, modelos baseados em árvores de decisão (especialmente gradient boosting e random forest) representam a abordagem mais adequada para estes dados. Nessa perspetiva, os modelos que consideram apenas as características sociodemográficas dos bairros são significativamente mais eficazes na previsão do que os inteiramente urbanos.

- **Título**: Crime Inference Using Machine Learning and Geographical Data
- **Autor**: Miguel Francisco Frade Roque
- **Palavras-chave**: crimes; socio-demographic; urban; linear; non-linear.

CONTENTS

1. INTRODUCTION

Crimes are not random events in society, and eventually something must influence their occurrence. Although it is difficult to find a causal relationship between crime and certain factors, given that no one is forced to commit crimes, it is possible to identify what circumstances seem to make crimes possible or more frequent. For example, pickpocketing is probably higher in crowded and touristy places, and violent crimes are thought to be more likely to happen in poor neighborhoods. Therefore, it is by characterizing the environment that it is possible to create algorithms that predict the criminal activity in a certain place and at some point in time. This, in turn, could permit its anticipation and prevention through decision-making in public policy. There is a lack of formal investigation on crime inference based on the characterization of their environment, and therefore this study represents a potential contribution to the acknowledgement on this topic.

This study bases on crime forecasting, but its goal is not to predict how many crimes will occur in a place (or many), in which time periods they occur the most, or which places tend to be the most criminal. Instead, it focusses on finding the best way to predict crimes, that is, which types of features are the most important to consider while predicting crimes, and which methods are the most predictive. This way, it is possible to obtain findings that can be generalized to every city, in order to optimize the criminal forecast and help to prevent crimes.

An analysis of the city of Philadelphia, in the state of Pennsylvania (USA), is made, taking into account the urban, racial, demographic and socioeconomic characteristics of its different geographical blocks, and the number of criminal occurrences in each of them, over multiple years. The variables considered are classified as socio-demographic and urban. The first group is associated to the characteristics of people, for instance, racial distribution, population, age, gender, education, and socioeconomic aspects. The second group is about the characteristics of the place, using multiple variables representing the number of some specific public institution or business space, such as for example the number of schools or restaurants in a neighborhood.

To carry out this analysis, it is firstly necessary to know the nature of the panel and its variables, mainly with regard to the number of crimes. It is also important to understand which characteristics best correlate with crimes and how their distributions look like. For this reason, this paper presents a framework about the data, in which a narration of the construction process of the dataset to be used is made, the variables existing in it are shown and described, and the hypothesis of panel balance is tested. Then, before proceeding to modeling, a review of the

descriptive statistics is performed, in which an analysis of the correlation coefficients is made, complemented with a comparison between the socio-demographic and urban characteristics, as well as an association of time-related factors with criminal activity; and it is observed, through histograms and geographical maps, the distribution of several variables, including crimes, in a natural and logarithmic way.

Further ahead, a theoretical framework of the adopted modeling methods and the metrics to be considered in the evaluation of the models is also made. The methods used are both linear and non-linear, although in the feature selection section only linear regressions are tested. In the main analysis, several models, with variables of different types included in them, are compared in both perspectives. The objective is to understand which types of characteristics, between socio-demographic and urban, have a greater influence on the number of crimes both from a linear and non-linear relationship point of view, and in which of the perspectives the error in criminal forecasting is minimized.

When considering a linear relationship, using standard linear regression, it is notable that the urban characteristics have a greater influence on the criminal frequency, given that the models that only include urban variables present a much lower error than the exclusively socio-demographic models. When non-linear methods are used, via machine learning techniques (decision tree, gradient boosting and random forest), it is evident that the prediction of the number of crimes is much more assertive for any type of variable, leading to the conclusion that the relationships studied here are not linear in nature, and therefore tree-based models (especially gradient boosting and random forest) represent the most suitable approach for this data. In this perspective, the models that consider only the socio-demographic characteristics of the neighborhoods are significantly more effective in forecasting than the entirely urban ones. Even so, the inclusion of both socio-demographic and urban variables always turns out to be the best option when the number of crimes is represented in a natural scale, with a smaller prediction error, but adding urban features to the socio-demographic ones only decreases the average prediction error by 0.3% and the standard prediction error by 0.7%. Thus, perhaps surprisingly, considering the physical characterization of the environment does not seem to help us improve our ability to predict.

The rest of the manuscript is structured as follows: In the next (second) chapter, called *Background*, a context of the crime in the United States is done, as well as a literature review on studies related to crime prediction. In the chapter 3, *Data*, there is a description of the data

and the process of collection and building. The chapter 4, *Analysis*, includes some descriptive statistics, the theoretical explanation of the metrics and models, the feature selection, and the results. To finish the study, the chapter 5, *Conclusion*, summarizes the results and explains the consequent findings and conclusions, which are compared with other similar studies, and refers the limitations of this analysis. After that, a post-conclusion chapter, *References*, containing the origins of the external information and cited papers, is also available at the end of this manuscript.

The entire analytical part of this study is carried out using the program and programming language R. The entire code is available in the following GitHub link: https://github.com/MiguelRoque23/Dissertation.git

2. BACKGROUND

2.1. CONTEXT OF CRIME IN THE USA

Crimes have very negative effects on society. These effects can take the form of monetary costs (short term), physical or psychological distress (often long term), or loss of life [1]. The costs can be incurred in the direct theft of goods or money, in medical, security, court and funeral expenses, or even in moving out of fear [1]. But the physical and psychological harm caused to victims turns out to be the most aggravating factor in people's lives and can cause mobility problems (physical) or trauma (psychological) that affect the confidence to make new friends or live a life without fear [1].

In economic terms, there is also a negative impact of crime. Not only is the tourism sector severely affected in more conflicted cities, as tourists are more afraid and less interested, but productivity is also reduced as more victims are absent from work [1]. Furthermore, public spending on security and justice is clearly higher when there is a greater criminal tradition, spending that could be converted into funds to support the economy [1].

The United States of America is far from being a perfect country. Despite the country's growing economic and technological development on a global scale, it continues to be too imperfect in social aspects, with frequent criminal activity as one of the biggest problems. The country ends up becoming famous for crime through, for example, the recurrence with which we hear episodes of shootings in public places on the news, even if these are only a very small representation of a vast set of these practices [25].

According to the Gun Violence Archive, on average, more than 40,000 people die each year from firearms in the United States, which translates into 111 of these deaths per day [25]. In 2021, more than 45,000 were registered, including homicides and suicides [25]. In the same year, there were about 692 mass shootings, a value that corresponds to almost twice the number of days in the year, counting only those in which there were at least 4 victims, either injured or killed [25]. Considering homicides in general, according to the FBI, the homicide rate was 6.9 per 100,000 inhabitants in 2021, with an absolute record of almost 23,000 homicides [26].

Among the most developed cities in the United States, the city of Philadelphia, in the state of Pennsylvania, is one of those with the highest crime record. According to official data for the first half of 2022, the crime rate in this city is 55% higher than the national average [22], a value that could be ambiguous if most of the registered crimes were of low gravity, something that

does not correspond to reality. Filtering only for violent crimes such as rape, murder, robbery and non-negligent manslaughter, the crime rate is 139% higher than the American average [23], which makes the statistics even less favorable for the one that is known as the city of brotherly love. The same data indicates that, in Philadelphia, the risk of being a victim of a violent crime is 1 in 25 [22].

In the year 2021, Philadelphia broke an obscure record at city level, which dates back to 1990 and corresponds to 500 homicides, registering 562 [23], a value that even in absolute terms exceeds the two largest American metropolises, New York and Los Angeles [24]. The homicide rate was, in the same year, 35 per 100,000 inhabitants, equivalent to five times the national rate, which was 6.9 [23].

However, it is only risky to visit the city if you circulate in the wrong neighborhoods, given that crime is concentrated in some places considered more dangerous and not spread equally throughout the territory, something that is in line with what is perceived in this study, and shown later, through a geographical map of the city with the distribution of the number of crimes by the respective neighborhoods.

The high quantity of crimes throughout the territory reduces the population's quality of life. Since the role of any State is to guarantee the general well-being, it is important to resort to certain public policies and, for these to be effective, it is essential to first understand which factors influence the number of crimes.

2.2. LITERATURE REVIEW ON FORECASTING CRIME

The relationship between crime and the various factors that may influence it is a topic that has already been addressed in several studies. A large part of these reports associate crime with aspects of a social or demographic nature, such as the level of education [2], unemployment rate [3], average salary [4], ethnic distribution [5], or population density [6]. Some analyzes also point to the criminal history of the localities as a potential determinant of the occurrence of crimes [7, 8].

Existing studies on this relationship can be inserted in 3 types of theories: Those based on the time-related factors to find correlations with crime (time-centric paradigm), those that use locations as targets in forecasting (place-centric paradigm), and those that focus on

understanding what individual or community features lead to crime (population-centric paradigm) [9].

Basing on the first paradigm (time-centric), researchers use self-exciting point processes to predict crimes and take conclusions about the temporal trends or seasonality that may be associated to them, thus a pure time series analysis is implied. There is an example of a study that uses these methods to understand the temporal dependence of burglary [7]. Also, there is another one that adopts a travel and opportunity model to find the time-related constraints on criminal activity, confirming the hypothesis that crimes are driven by the accessibility of opportunities existing in the routine lives of criminals [12].

Many studies focus on the second paradigm (place-centric), in which the goal is to predict the geographical location of crimes, with the designation of hotspot. An example of this type of research is one that uses the history of criminal occurrences in space and time to find patterns of both dimensions in the correlation with crime, adopting many quantitative methods from mathematics and physics [13]. A second case is where a simple model understands the dynamics of the hotspots and detects the stable ones, using criminals as random walkers [14]. A particular study uses mobile network data to profile people behavior, and also demographic information simultaneously, to guess the hotspots, with random forest ending up being the best classifier [15]. There is also a very complex one that uses natural language processing on Twitter data in this type of research [16]. And some other examples of studies adopt an alternative statistical method called kernel density estimation to identify the most criminal zones [17, 18, 19].

Regarding the third paradigm (population-centric), to profile criminals in the perspective of an individual, a study identifies the crimes that were executed by a specific person, among a set of many crimes from multiple authors, with a model named Series Finder, created to characterize the way criminals act [20]. In the perspective of a community, there is an example of a paper where mobile phone data from London is used to associate people dynamics with the criminal activity, with the aid of adequate computational techniques [21].

Although the resources available in the present study, in terms of data, allow an analysis based on any of these two paradigms, this research falls into the third category. Many geographical areas are analyzed throughout multiple time periods, but each location is characterized by different attributes, not only at the social and demographic level, but also urban, through the categorization of the neighborhoods according to their distribution of several specific

institutions or business spaces. Instead of predicting crimes, the goal is to discover which of these types of features have a higher weight on crime forecasting.

A very interesting analysis, and similar to what is done in this study, is that of a paper that adds the taxi flow and the points of interest of the city of Chicago to the usual social and demographic features [9]. The points of interest refer to the number of structures belonging to a specific institutional or business category present in a geographical block, such as nightclubs, restaurants, or schools, but are complemented by the popularity and reviews of each one. Some preceding studies had already verified a large importance of the points of interest in profiling the neighborhoods [10], which helps in criminal forecasting, and that paper also concludes that the inclusion of those type of variables greatly reduces the error of the prediction (by 5%) [9]. Its literature review also indicates that previous studies conclude that a neighborhood, despite not being influenced by another geographically bordering neighborhood [6], is affected by neighborhoods that are close in terms of ease of access [11]. Thus, that study uses taxi flow to consider this indirect proximity between locations. In the end, it turns out that there is little influence on the part of geographic factors, but that taxi flow significantly reduces the error in predicting the crime rate (by 5% if only taxi flow is added to the demographic features, and by 17.6% if both taxi flow and points-of-interest are added.) [9].

It is very important to focus on this specific study, as the use of points of interest is comparable with the inclusion of the so-called urban variables, even though factors linked to popularity are only used in the first one. One of the differences between the two studies is in the use of the taxi flow and in the consideration of the possible geographic influence, because in the present study the variables related to the coordinates of each neighborhood are not included, nor the relationship of proximity between them. In addition, the methods used are also different, since in this reference the basis is the linear relationships between crime and the other factors, through the exclusive use of techniques such as linear regressions and negative binomial regressions, while in the present analysis some machine learning techniques are considered as an alternative to linearity. Therefore, the focus has more to do with the conclusions that are drawn at the end regarding the points of interest, which are the target of comparison with the results of this study and may lead to an interesting relationship with the methods applied in each one.

3. DATA

For the purpose of this study, I coalesce crime data (number of crimes) with sociodemographic information (related to racial distribution, quality of life, education, and population) and urban geography features (such as the location of a park, a restaurant, or a hospital). To prepare the inputs of the posterior analysis, multiple resources were initially used:

- The crime frequency per block and per month, from 2016 to 2019, which is provided by the Philadelphia Police Department website [27], inserted into a publicly accessible database with the vast criminal history of the region. Despite the enormous detail that characterizes the data, only information on the total number of crimes is collected.
- The information associated to sociodemographic parameters, from 2016 to 2019. This data comes from the surveys of the US Census Bureau [28], which has the purpose of controlling and updating information about the localities and their households, helping in the public decision making. These numbers only change annually, therefore they keep constant across the different months of the same year.
- One shape file with the geographical coordinates and shape details of 1336 blocks belonging to the city of Philadelphia, to be possible to build and visualize a map of the city with different variables by neighborhood through a color scale. Also, the location of multiple categorized urban objects is used, so the count of each of them can be made per block. This data is origin from the OpenStreetMap project [29].

The geographical covariates that are used are the counts of object of a given type (for instance, restaurants) in each block (for instance, block 749 where the Liberty Bell is). These geographical covariates are obtained from the urban objects as follows. Each of these urban objects contains geographical details that determine their exact location. Since the shape and geographical borders are also known, the distance between a specific urban point (for example, one restaurant) and each block is calculated, by identifying the closest point of the polygon to the isolated point. After all distances are collected, it is possible to classify each urban object as inside or outside each geographical block, and after that, the count of each type of urban object for each geographical block (for example, the number of restaurants in one specific neighborhood) is done. This count originates the urban variables that are used in this study, which are specified below.

The criminal, socio-demographic and urban data is joined by the common columns into a single and complete dataset, having all the information per neighborhood and month, and serving as

the input of descriptive statistics and modeling analysis. This data frame is also joined to the shape file and grouped by block, excluding the time periods, and averaging all the values, in order to have a completed shape file, for geographical visualization of the data. The final data frame has 64 128 rows (1336 blocks and 12 months of 4 years) and 57 columns in total, which are the following:

- General columns:
  1. ***block***: city neighborhood, identified by a number from 1 to 1336.
  2. ***year.month***: a code to identify the month of a specific year from 2016 to 2019.
  3. ***year***: year of *year.month*.
  4. ***month***: month of *year.month*.
  5. ***crimes***: number of crimes (key variable of this study).
- Socio-demographic columns:
  1. ***white***: proportion of white residents in a block.
  2. ***black***: proportion of black residents in a block.
  3. ***asian***: proportion of Asian residents in a block.
  4. ***latino***: proportion of Hispanic residents in a block.
  5. ***nilf***: proportion of unemployed residents in a block.
  6. ***hsol***: proportion of residents with high school or lower education in a block.
  7. ***age***: average age of the residents in a block, in years.
  8. ***male***: proportion of male residents in a block.
  9. ***pop***: total population in a block.
  10. ***area_kmsq***: total area of a block, in squared kilometers.
  11. ***pop_density***: population per squared kilometer in a block.
  12. ***poverty***: proportion of residents below the poverty level in a block.
  13. ***income***: average yearly income of the residents in a block, in US dollars.
- Urban columns:

  airport, bar, food, university, school, library, parking_lot, bank, hospital, healthcare_*professional, pharmacy, social_center, theater, club, law_enforcement, fire_station, church, industrial, military, cemetery, park, indoor_sport, outdoor_sport, golf, office, station*, *liquor_store, food_and_beverages, mall, clothing, health_and_beauty, DIY, furniture, electronics, car_related_shops, art_and_hobbies, books_and_gifts, tourist_attraction, tourist_accommodation* (the names of the variables are expressive,

which means that, for example, *airport* refers to the number of airports in a block, and *food* implies the number of restaurants, snack bars, or cafés).

Some blocks such as parks, industrial zones, airports, or military bases, have no population. For those ones, since the socio-demographic information is always a ratio or an average, it is not possible to have a value, but they are important observations because of their interesting urban profile, therefore they cannot be removed from the data set. To fix this problem without affecting the models, the yearly average of each socio-demographic feature is used in these neighborhoods, and only the variables *pop* and *pop_density* have "zero" values on these observations. Exceptionally, this method is not applied in the case of the geographic representation of these variables for these specific blocks. Since it does not make sense to present colors corresponding to an average value in these cases, as this would confuse the interpretation of the map, no value is assigned to them.

Since we are dealing with panel data, we have key variables by which the data is grouped, and it is crucial to check whether the panel is balanced or not, so that the information is complete, and all elements of the sample have equal participation in the analysis. In this case, those variables are *block* (space) and *year.month* (time). As previously referred, this study is analyzing a total of 1336 neighborhoods and 4 entire years (48 months). If the data frame is grouped by *block*, we can count exactly 48 rows per each, and, if it is grouped by *year.month*, there are 1336 observations per month. Therefore, it is confirmed that this panel is balanced and there is no missing data.

4. ANALYSIS

4.1. DESCRIPTIVE STATISTICS

4.1.1. CORRELATION

In an analysis of the correlation between *crimes* and the remaining features, it is observable that there are no socio-demographic variables correlated with the number of crimes. The highest correlation coefficient found is that of *poverty* (0.230), which is the only one above 0.2 (or under -0.2). Among the urban variables, some more significant correlation coefficients, comparing to the previous ones, are already observable. The biggest one is that of *station* the biggest (0.452), followed by *food*, *tourist_accomodation* and *bank* above 0.4, and 5 others above 0.3. There is only one urban variable that is negatively correlated with crimes, which is *golf*, and it has the weakest of all coefficients (-0.003).

Figures 1 and 2 compare both best correlations of each type of features with *crimes*. The black straight lines represent the linear relation between the two variables in each plot, and they show that the poverty rate influences the number of crimes much less than the number of stations does, as the second has a much higher slope than the first, using the same scale in the *crimes* axis. But even the second plot seems to indicate a non-linear relation with the number of crimes, since the points are very dispersed in relation to the trend line. Therefore, this suggests that non-linear models might be better suited to this data and consequently lead to a more assertive prediction.

Figure 1: Relation between poverty rate and number of crimes

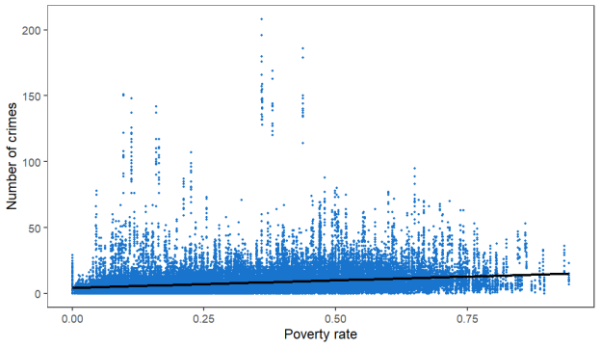
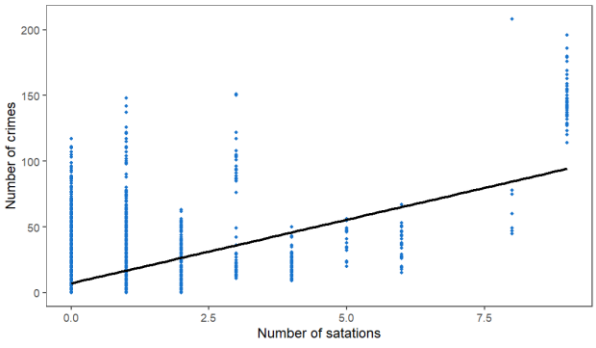


Figure 2: Relation between number of stations and number of crimes



Examining the correlation coefficients between the independent variables, excluding *crimes*, and starting by the socio-demographic features, it is possible to verify that the variables *white* and *black* are strongly correlated (-0.908), therefore, to avoid a collinearity problem, they cannot be included in a model at the same time. The *income* variable also shows a significant correlation with *hsol* (-0.774) and *poverty* (-0.660), that is, if the first one is included, the other

two are not, but, given that *hsol* and *poverty* are not so correlated with each other (0.534), these may be together.

Figures 3 and 4 show the strong correlation between *white* and *black*, and between *income* and *hsol*, respectively. In both plots the points follow the direction of the trend line, which evidences clear linear relations.

Figure 3: Relation between white rate and black rate

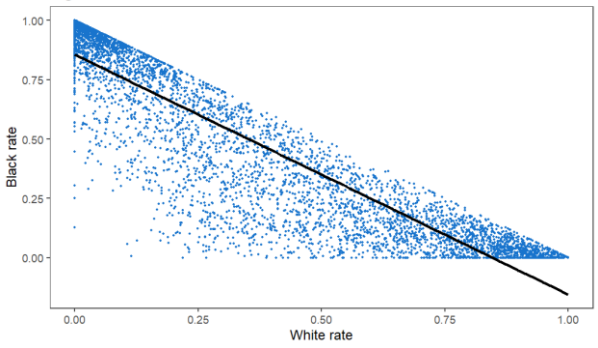
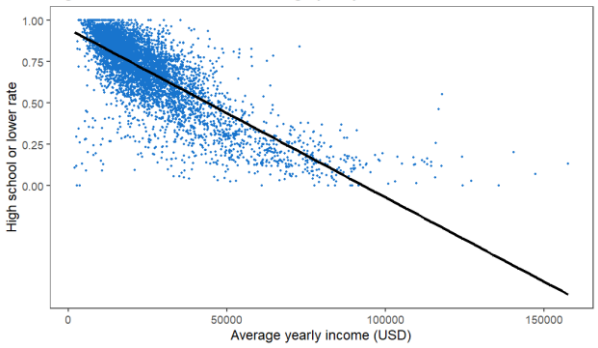


Figure 4: Relation between average yearly income and low education rate



Regarding the time association with the number of crimes, when visualizing the average number of crimes per month, it is noticeable that the occurance of crimes is lower during the colder periods, that is, between november and march, with special highlight to february. This may be related to the fact that people spend more time indoors and less on the street. This not only decreases the number of crimes in open spaces, but also reduces the number of burglaries in houses, because these are under greater surveillance. In addition, tourist activity is also reduced in winter months, which reduces the influx of people in the city and the use by robbers.

Over the 4 years of analysis, the total number of crimes has not shown any trend, nor peaks or drops, therefore it can be assumed that crime has remained constant in this city.

Figure 5: Average number of crimes per month

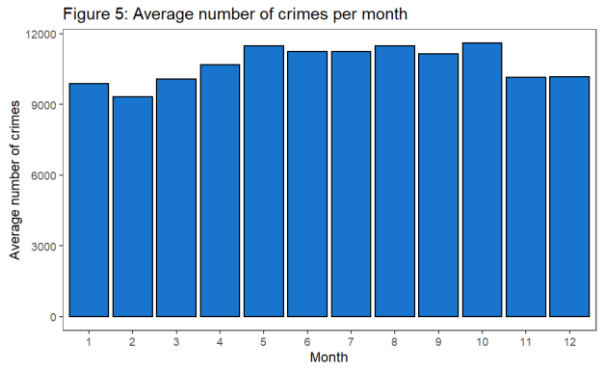
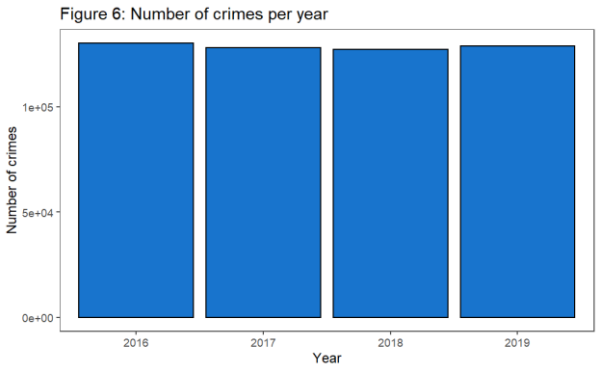


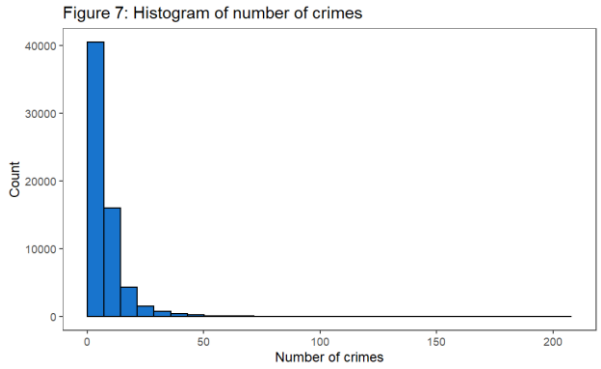
Figure 6: Number of crimes per year

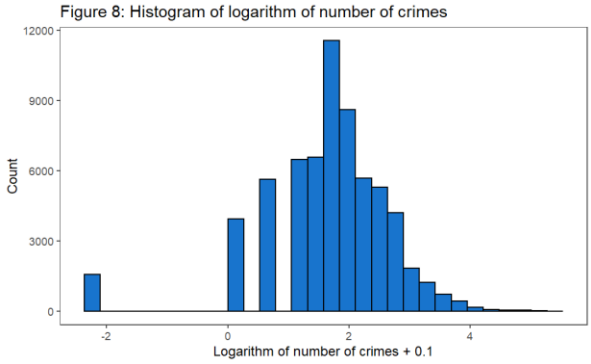


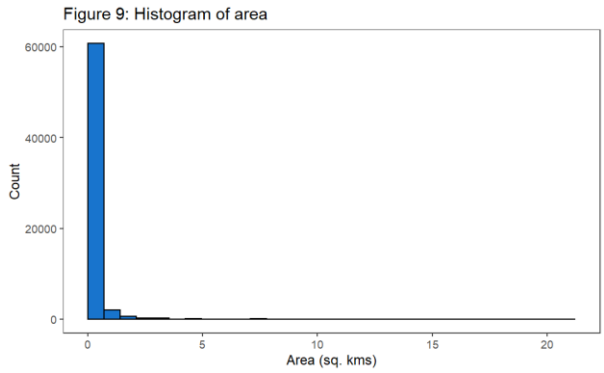
This correlation between months and the number of crimes, as well as for years, is taken into account throughout the analysis, as the control for temporal factors are always included in the models.

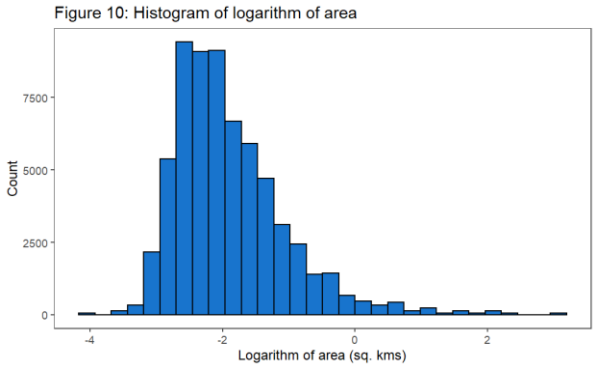
4.1.2. VARIABLES DISTRIBUTIONS

Through histograms, the distributions of socio-demographic variables whose value is not between 0 and 1 are observed, in order to understand whether they should be included in the models in their natural or logarithmic form. The distribution of the main variable, *crimes*, is very concentrated on the left, on values up to 15. However, there are multiple observations with higher values, reaching 208. The distribution of the variable *area_kmsq* has a similar shape and is also not a normal distribution. Because of that, the number of crimes and the block area are probably more predictable in their logarithm form. The first feature has some "zero" values, then we must add 0.1, so that the transformation is possible. By watching their histograms in this form, an approach to a normal distribution is noticeable in both cases.









Regarding the variables representing the average age, population, populational density and average income, all histograms show distributions that are closer to normal, with special highlight to *age*, therefore they can be used in their natural form in the models, which facilitates the interpretation of their coefficients.

Figure 11: Histogram of average age

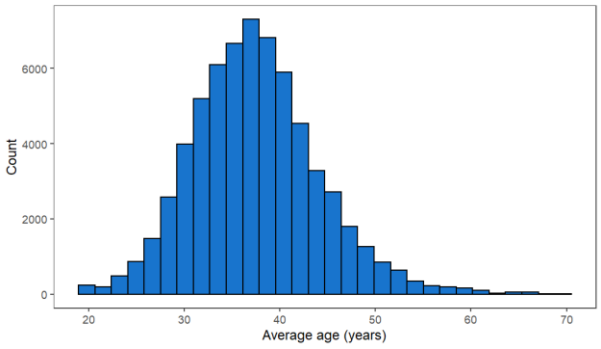


Figure 12: Histogram of population

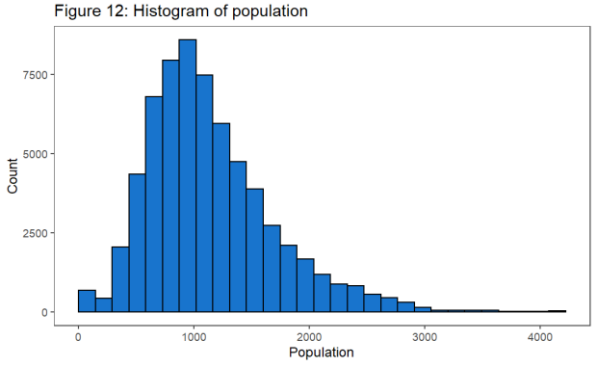


Figure 13: Histogram of population density

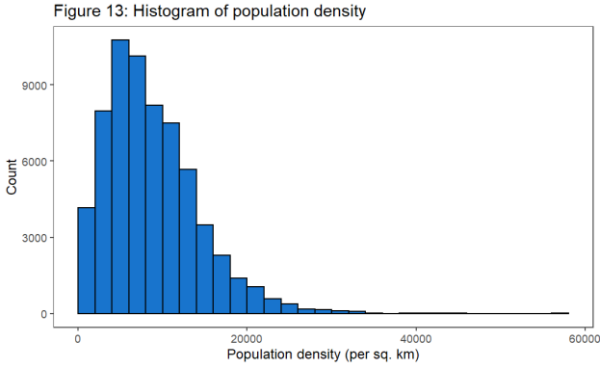
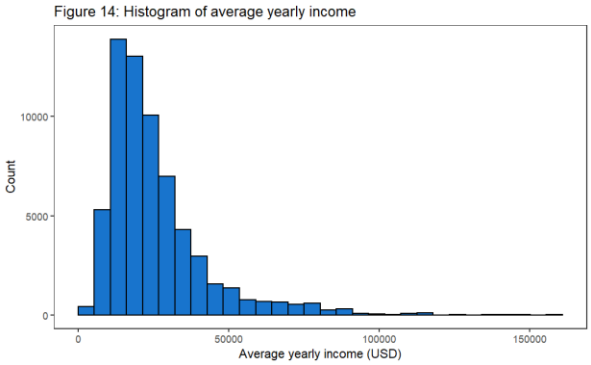
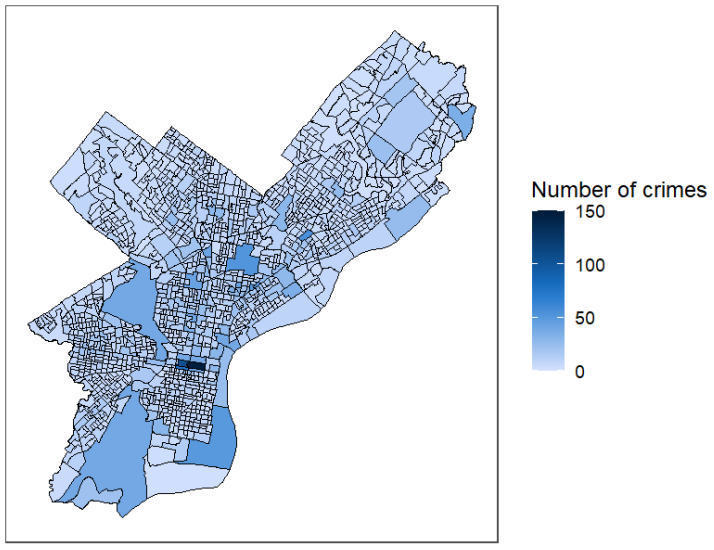


Figure 14: Histogram of average yearly income



Visualizing the distribution of the monthly average of the number of crimes by the geographic map of the city of Philladelphia, in Figure 15, there is a concentration in the downtown, with the neighborhoods in dark blue having particularly high crime, even though there are other neighborhoods around the territory with high criminal frequency. The most prominent of all has an average of over 147 crimes per month, followed only by another with an average higher than 100. Including these two, 7 of the 1336 neighborhoods record an average of above 50 crimes per month, which is translated in much more than one crime per day.

Figure 15: Monthly average number of crimes per block



Figures 16 to 19 portray the share of residents per block by ethnicity. In Figures 16 and 17 there is an obvious inverse proportionality between the share of white and black people living in a block, since these are the two main racial groups in the city and in the country, thus when a block is represented by a dark color in one map, it has a light color in the other map, and vice versa. There are some observeable exceptions, particularly in some neighborhoods in the center of the city, where they keep light colors in both maps because the majority in those blocks is some other ethnicity.

In Figures 18 and 19 it is represented the distribution of two minor ethnicities in the city, showing the proportion of Asians and Hispanics per neighborhood. The Hispanic community is concentrated on the center of the map, in multiple adjacent blocks, while the Asians are spread across the city, from north to south.

Figure 16: Proportion of white residents per block

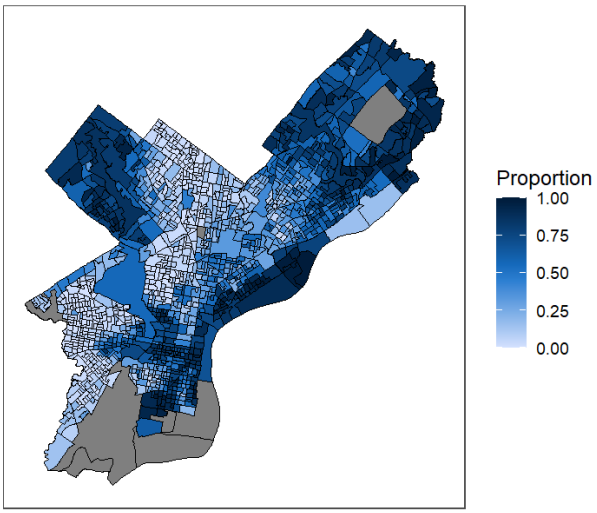


Figure 17: Proportion of black residents per block

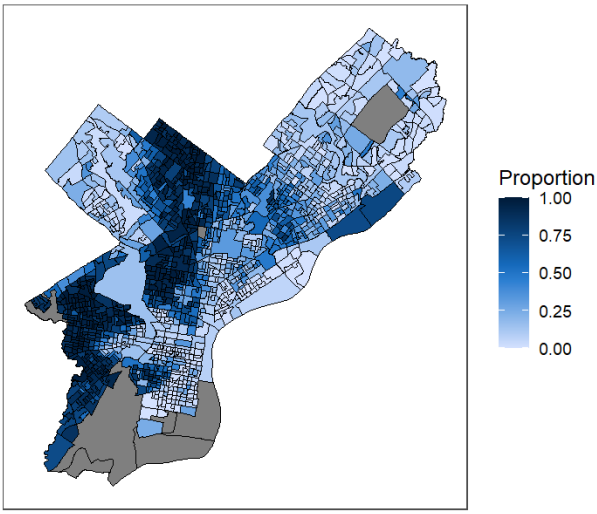


Figure 18: Proportion of Asian residents per block

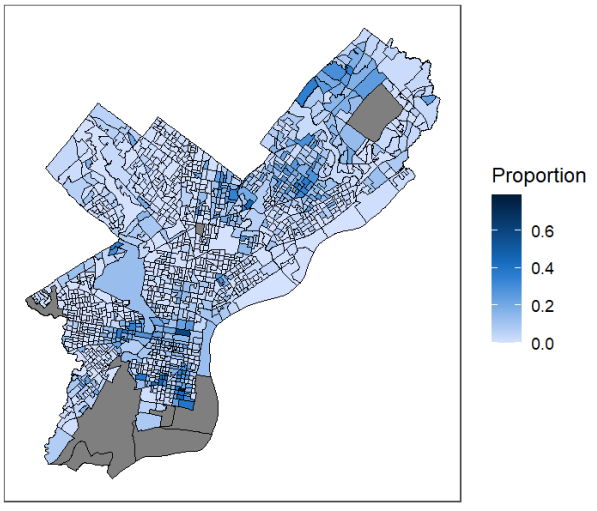
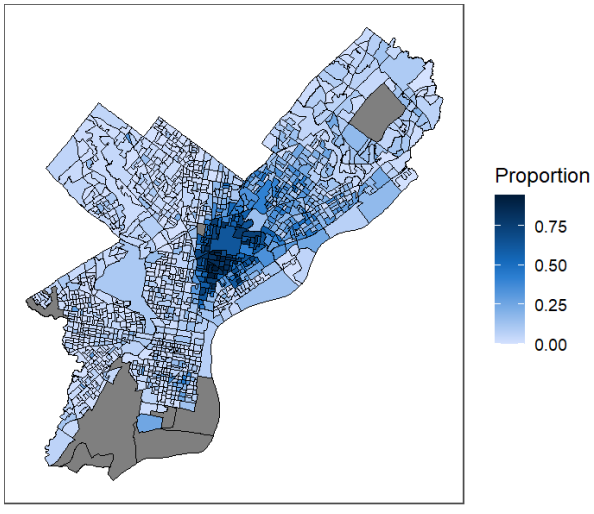


Figure 19: Proportion of Hispanic residents per block



4.2. METRICS

For all models that are run in this study, from feature selection to the main analysis, the k-fold cross validation method is adopted instead of hold-out and thus it is not necessary to split the data into a single train and test set. The entire data is randomly split into 5 folds, and 4 of them are alternately used as the training data, with the model being validated on the remaining set, repeating this process until the validation is done on all folds. This way, the models are trained and tested in many different samples, avoiding the hypothesis of a single lucky draw, and allowing the contact with very diversified sets, which improves the generalization to unseen data.

Subsequently, the performance of the models is evaluated and certain metrics such as MAE (mean absolute error) and RMSE (root mean squared error) are compared between them. Before moving on to the stage where they are present, it is important to understand what they mean, what they are measuring, and how to interpret them.

- **MAE** (mean absolute error) corresponds to the average of the absolute difference between the original and predicted values, equaling a negative error to a positive one. It measures the mean of the residuals.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$y_i = actual\ value; \hat{y}_i = predicted\ value; n = number\ of\ observations; i = observation$

- **RMSE** (root mean squared error) refers to the square root of the average of the squared difference between the actual and predicted values. It measures the standard deviation of the residuals.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$y_i = actual\ value; \hat{y}_i = predicted\ value; n = number\ of\ observations; i = observation$

When a large error is squared, the value increases much more than when the same is done on a small one. Therefore, RMSE penalizes the largest errors in the prediction, comparing to MAE, since they are squared at first and only the average is then rooted. Meanwhile, MAE is easier to interpret because it is simply an average. For both metrics, the smallest the value the best it is, meaning that the predictions are closest to the actual numbers.

4.3. MODELS

To understand which types of features should be used to make a better prediction of the number of crimes, two different analytical paths are followed.

The first is via linear regressions. Different types of models are created: one including only socio-demographic regressors, another exclusively with the urban variables, and a third with a mixture of both previous ones. For every of the 3 types, there is also a second version using only the variables that prove to be most statistically significant.

As the number of crimes may not have a linear relationship with the other variables, the second path is done alternatively through the application of more modern techniques, by using several tree-based machine learning methods suitable for this regression problem, such as decision tree, gradient boosting, and random forest. The same combinations of regressors as in the first stage are used in this phase as well.

It is important to always include the temporal control variables *year* and *month* in the models as dummies, since months influence crimes, and the majority of features remain constant during all months of one year.

Just as it is important to understand what the metrics consist of, it is also crucial to understand how the different proposed models work, even if a detailed interpretation of each one is not carried out and only their performance is analyzed. In this section, the different techniques used in this study are explained in a brief and simple way, with no need to go into much depth.

4.3.1. LINEAR REGRESSION

A linear model evaluates the influence that the variation in one of the independent variables has on the dependent variable, controlling and keeping all the others constant. A linear relationship between the two variables is assumed and therefore the average effect is taken into account, ignoring the location of the first in its total range. In this study it is adopted the OLS (ordinary least squares) method for linear regression, which only predicts continuous variables.

This method is effective in case the independent variables actually have an almost linear relationship with the dependent one. In cases where the correlation between both differs depending on the value of the first, considering the average effect leads to a high error in the forecast.

The plots displayed below show an example of a model where a linear analysis is effective, on the left, and one where the prediction error is very high if the average relation between both variables is used as reference, on the right. In the first model, the values of *y* are very close to the average trend, represented by the black straight line, therefore the prediction errors are small. In the second, the values of *y* are concentrated in 3 distinct ranges of *x*, with *y* smaller on both extreme intervals of *x* and larger on the intermediate interval, which means that assuming the same effect for all situations leads to a huge ineffectiveness of the model.

Figure 20: Relation between x and y

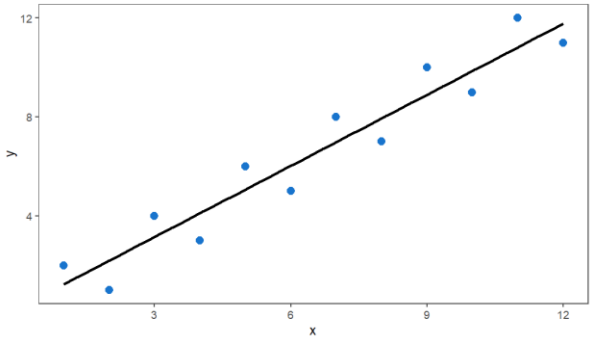
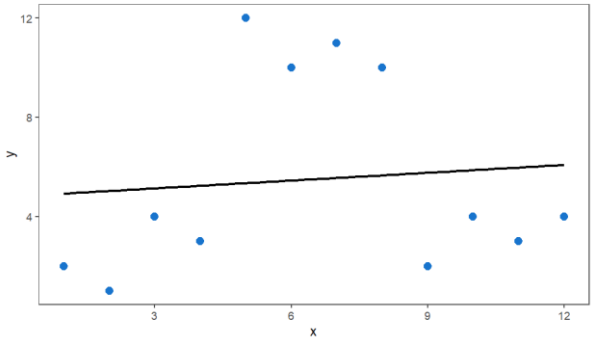


Figure 21: Relation between x and y



When there is a case like the one on the right, in which there are clearly multiple intervals of values of *x* for which the influence of this variable on *y* is different, a non-linear method must be applied and good examples are the tree-based models, which are explained below.

4.3.2. DECISION TREE

The simplest tree-based model is the decision tree itself, which is a machine learning technique that consists of creating branches using the various independent variables to draw a path to the final decision, that is the prediction of the value of the dependent variable. The decision tree can be used to predict categorical or continuous variables, assuming the designations of classification tree or regression tree, respectively (in this case, it is the second).

Being a tree, it has a root, which corresponds to the split of the range of values of a variable that allows the best differentiation of the outcome. This root represents the first node, and each node gives rise to two branches, each one representing a subset of values resulting from the division and connecting the previous node to a new one that repeats the process until an average outcome (leaf) is reached.

Picking the example of the last plot seen in the linear regression section, the observations with x lower than 5 would follow a different path than the remaining ones, and those others would be split into a subset of values lower than 9 and the last group. This way, 3 different average values of y would be assigned to these 3 different intervals and the predictions errors would be low.

Figure 22: Relation between x and y

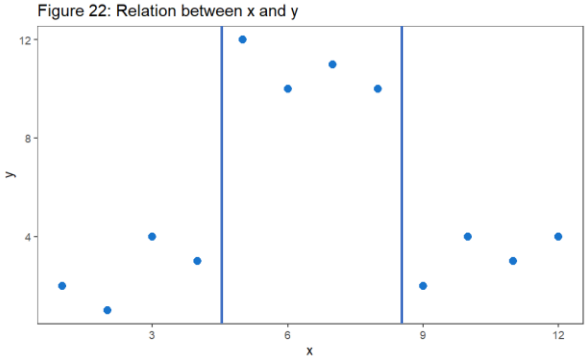
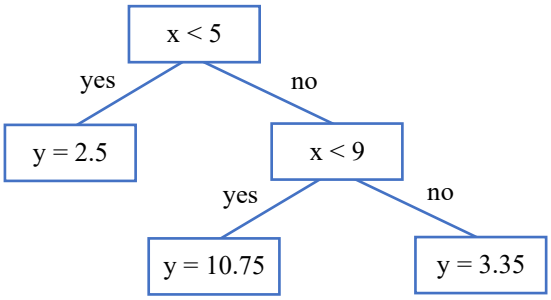


Figure 23: Decision tree based on Figure 22



The greater the number of nodes, the more complex and consequently more specific the tree becomes, which impairs the generalization of the model for different data. Therefore, this study tunes the parameter of the maximum depth of the tree to try to understand which one reduces the errors in the validation the most. Still, when too many variables are included, this method makes the models either too individual, if the tree is deep, or fails to consider many important factors, if it is short, ending up reducing its effectiveness. Therefore, other more modern techniques have been developed to use trees that can be successfully applied to very diverse data.

4.3.3. GRANDIENT BOOSTING

Like all boosting methods, the gradient boosting tree joins a series of weak models with the objective of creating a strong one, thus translating into an ensemble model, in this case, of several decision trees (weak models).

It starts by creating the simplest decision tree possible, with only one leaf corresponding to the average value of the dependent variable in the entire dataset, which represents the first forecast for all observations. Following, all prediction errors are collected. As an error results from the difference between the actual and the predicted value, a residual is positive when the estimation is lower than the reality, and negative when it is higher.

Afterwards, a normal decision tree is created considering all variables, but instead of predicting the outcome value, as usual, it predicts the residuals obtained in the first step, thus each leaf corresponds to the average error of the respective subset. Once this tree is built, new predictions are made by summing the initial guess (average value) to the product of a defined learning rate (between 0 and 1) and the error indicated by the leaf where each observations fits, as demonstrated below.

$$new\ prediction_i = average\ value + learning\ rate \times predicted\ error_i$$

The new residuals are logically smaller this time. All of them are collected again, and a new tree is created to predict them. Although a new one is built, the previous tree is not forgotten, and the subsequent prediction combines the preceding one with the product of the same learning rate and the error indicated by the new tree, as in the following formula.

$$new\ prediction_i = previous\ prediction_i + learning\ rate\ \times\ new\ predicted\ error_i$$

The residuals keep decreasing and the predictions slowly approach the actual values. This process is repeated until a new tree reveals to be indifferent for the forecast, which means that the correct values are predicted and no errors are verified. Also, it can reach the maximum number of trees specified.

This method is an improvement of the simple decision tree models, and it is much better because it contains a learning rate that allows the model to learn slowly and build multiple trees in an effective way, which leads to a gradual successful combination of their results.

4.3.4. RANDOM FOREST

Random forest is also an ensemble model of several decision trees, such as gradient boosting, but the difference is in the way these are combined. While in the previously seen method, multiple trees are joined in sequence and are dependent on each other, in this one they are created in parallel and the average value of the combined predictions of each one is assumed as the final prediction of the model.

Of course, using the same dataset for all, the trees are homogeneous, and the average of the results does not differ from that of a simple decision tree. Therefore, two processes occur inside this model to diversify the trees:

The first is called “bootstrapping” and consists of creating distinct samples with the same size as the original dataset for each tree, by randomly replacing every observation with another in the data, making many of them appear repeated. This allows the trees to form differently. The other step has the designation of feature randomness and consists of randomly filtering the variables to be considered in each node of each tree, which means that their nodes are automatically different between them.

After all the trees are created, all the data is used to obtain the predictions, which are naturally different between them, as a result of the differentiation of their nodes. The aggregation of the multiple predictions is called bootstrap aggregation or bagging.

It is this creation of multiple trees at random that originates the name “random forest”. This method is more effective than a single decision tree because, by creating many trees with different samples and sets of variables, it prevents the model from individualizing and not fitting to data that has never been seen before, thus the bigger the number of trees the best.

4.4. FEATURE SELECTION

Before advancing to the main analysis, it is crucial to make a general feature selection, considering the previously assessed correlation coefficients between the covariates, to avoid having strongly correlated regressors together and consequent collinearity or multicollinearity in the models. To select the best combination between two or more correlated variables, multiple linear regressions are performed, and the two already mentioned evaluation metrics are considered to choose the best option.

The two different ways in which the dependent variable *crimes* can be represented, natural or logarithm scale, are also compared in the same way in these linear models. It had previously been assumed, in the analysis of the histograms of the number of crimes, that the best option would probably be to use it in the logarithmic form, however also its natural representation should be tested in parallel.

As previously clarified, these linear regressions are run using the k-fold cross validation method, by splitting the data into 5 folds, and the models are compared through the same metrics, RMSE and MAE.

This feature selection is performed separately in 3 different sorts of models. Firstly, only models with exclusively socio-demographic features are compared. After that, the models only contain urban variables. To finish this section, models with both types of features are used to select the best way to represent the number of crimes.

4.4.1. SOCIO-DEMOGRAPHIC FEATURES

To start, only socio-demographic covariates are included. As it was previously stated in the correlation coefficients analysis among these variables, *white* and *black* should not be included in the models simultaneously, as well as *income* must be isolated from *hsol* and *poverty*. There are thus 4 different combinations to be tested.

The tables below display the metrics of the linear models that are tested in this segment, which serve as criterion for selecting the variables for the most advanced phase of the analysis. Although the representation *log(crimes)* is written, it corresponds to *log(crimes + 0.1)*, since there are some "zero" values in *crimes* among the observations, and the first way could not be used. Moreover, the errors and consequently the RMSE and MAE of the logarithmic models are at same scale as the others, by exponentializing the prediction values, to allow a better comparison.

Table 1: Metrics of linear models including only socio-demographic features

Socio-Demographic Model	RMSE	MAE
crimes ~ white + hsol + poverty + ...	8.369	4.675
crimes ~ white + income + ...	8.421	4.791
crimes ~ black + hsol + poverty + …	**8.367**	**4.676**
crimes ~ black + income + ...	8.413	4.787
log(crimes) ~ white + hsol + poverty + ...	8.673	4.308
log(crimes) ~ white + income + ...	8.762	4.408
log(crimes) ~ black + hsol + poverty + …	**8.673**	**4.306**
log(crimes) ~ black + income + ...	8.757	4.401

This table indicates that the best option is to include the variables *hsol* and *poverty* and exclude *income* from the models. When using the first two, the RMSE and MAE is always lower, favoring them, once compared to the inclusion of the other variable instead. These metrics remained practically constant between the use of *white* and *black*, yet slightly better when using the second one in almost all models, which leads to the choice of *black* instead of *white*.

It is verified that the linear regressions in which the number of crimes is not used in logarithm scale are always the ones that have a higher MAE, thus it could be confirmed that the models with *crimes* in the logarithmic form perform better, however the RMSE shows an opposite behavior. This way, it is not clear in which form the dependent variable should be represented in the main analysis, therefore both ways are tested for the socio-demographic features models.

4.4.2. URBAN FEATURES AND MIXED MODELS

Among the urban variables, there is no presence of too strong correlations, therefore there is no need to exclude any of them. This time, only urban covariates are used in the models, and two versions are again created, modifying only the form in which the dependent variable is represented, natural and logarithmic.

Table 2: Metrics of linear models including only urban features

Urban Model	RMSE	MAE
crimes ~ all urban features	**6.709**	**4.378**
log(crimes) ~ all urban features	15.772	4.600

The model that shows the best metrics is, by far, the one in which the number of crimes is not transformed, with a huge difference in the RMSE and MAE. There is no need to use the logarithm of *crimes* in the models with exclusivity of urban variables from now on.

Table 3: Metrics of linear models including both socio-demographic and urban features

Mixed Model	RMSE	MAE
crimes ~ all features	**6.139**	**3.979**
log(crimes) ~ all features	13.984	4.150

When comparing the mixed features models, which include the previously selected socio-demographic variables and all the urban features, the same conclusions are taken. Both metrics favor the first model, where there is no transformation of the dependent variable, therefore its logarithm is dispensable in the next analysis for all mixed models as well.

4.5. RESULTS

4.5.1. LINEAR MODELS

From the previous analysis, 4 linear models were selected, two with only socio-demographic regressors, one exclusively urban, and another with both types included. From those models, the statistical significance of each coefficient was analyzed to pick the most relevant features and create a simpler version of the models. Given the large number of regressors and very large number of observations, in order to differentiate the models with all variables and with only the relevant ones, a smaller than customary threshold of was stipulated for the p-values to be considered statistically significant in this analysis. Only coefficients with p-value larger than 10^{12} were considered significant (basically, barely larger than machine zero). It is implied that this process was applied to the model that includes all the variables as well, and some of the coefficients show different levels of significance compared to the observed in the smaller models.

In the next tables, when the expression "all features" is used, only the variables chosen in the feature selection are being considered, thus not corresponding to all existing ones in this study. In this case, the filter applies exclusively to socio-demographic features. As in the preceding section, from *log(crimes)* it is implied *log(crimes + 0.1)*, and the errors are also all adapted to the same scale, for the same reasons previously explained.

After training and testing the linear regressions for the 8 models, in the way that was previously explained in the methodology, the verified values of the metrics are as follows:

Table 4: Metrics of linear regression models

Linear Regression Model	RMSE	MAE
crimes ~ all socio-demographic features	8.367	4.676
crimes ~ significant socio-demographic features	8.375	4.680
log(crimes) ~ all socio-demographic features	8.673	4.306
log(crimes) ~ significant socio-demographic features	8.681	4.307
crimes ~ all urban features	6.709	4.378
crimes ~ significant urban features	6.718	4.383
crimes ~ all features	**6.139**	**3.979**
crimes ~ significant features	6.154	3.985

In the model that includes all variables, without any restriction of type or significance, the analysis metrics MAE and RMSE showed the lowest values. That said, it is assumed that, from a perspective of linear relationship between crimes and the remaining variables, the most complete model is the best to predict the number of crimes.

It is also possible to observe that the metrics of the models with urban variables present much more favorable values than those of the models with socio-demographic variables. This indicates, similarly to what had already been verified in the analysis of the correlation coefficients, which also measure a linear relationship between the variables, that the urban characteristics present a much more linear influence on the number of crimes than the socio-demographic ones.

4.5.2. TREE-BASED MODELS

In the second part of the analysis, as an alternative to linear regressions, some machine learning models are implemented to escape from the assumption of a linear relationship between crimes and the other regressors. The first tested technique is the decision tree, and in Table 5 are the metrics achieved by each model.

For these models, only the maximum tree depth hyperparameter is tuned, by grid searching between 2 and 30. The best models from those with only socio-demographic features have a maximum tree depth of 14, and 4 in the case of the logarithmic ones. The best value of this hyperparameter is 9 for both urban models, 13 for the most complete model, and 11 for the one with only significant variables.

Table 5: Metrics of decision tree models

Decision Tree Model	RMSE	MAE
crimes ~ all socio-demographic features	6.935	4.411
crimes ~ significant socio-demographic features	6.935	4.411
log(crimes) ~ all socio-demographic features	8.917	4.441
log(crimes) ~ significant socio-demographic features	8.917	4.441
crimes ~ all urban features	6.902	4.546
crimes ~ significant urban features	6.934	4.548
crimes ~ all features	**6.393**	**4.201**
crimes ~ significant features	6.472	4.287

In relation to what is verified in the linear regressions, this method leads to a slight worsening in the results in all models where urban variables are included. It is also evidenced that the use of *crimes* in the logarithmic form is ineffective in this case. Considering only the models where the dependent variable is in the natural scale, the ones consisting entirely of socio-demographic regressors show a great performance improvement, with more favorable values for all metrics, having decreased both RMSE and MAE. Although the ranking of models does not change from one method to the other, there is clear evidence that the types of features are related to crimes in an opposite way when a non-linear analysis perspective is adopted.

As warned in the theoretical framework, in overly complex models, decision trees turn out to be ineffective. In this way, it is very likely that more evolved techniques based on these trees lead to better quality and more conclusive results. Following, to support the last phrase, more non-linear models are run. The second method addressed is the gradient boosting, and the metrics are shown in Table 6.

For these models, the tuned hyperparameters are the number of iterations and the maximum tree depth, using grid search. The grid contains values between 200 and 1000, jumping from 50 to 50, for the first hyperparameter, and the values 5, 8, 11 and 14 for the other one. The learning rate is constant at 0.1, there is no regularization, the minimum number of instances required in a child node is 1, the proportion of both features and observations supplied to a tree is 1. The best maximum tree depth is always 8 for the normal models and 5 for the logarithmic ones, and the best tunning value of the number of iterations for each model is respectively 400, 450, 1000, 1000, 500, 200, 350 and 400, following the order presented in Table 6.

Table 6: Metrics of gradient boosting models

Gradient Boosting Model	RMSE	MAE
crimes ~ all socio-demographic features	3.761	2.635
crimes ~ significant socio-demographic features	3.771	2.637
log(crimes) ~ all socio-demographic features	**3.697**	**2.385**
log(crimes) ~ significant socio-demographic features	3.806	2.430
crimes ~ all urban features	5.001	3.432
crimes ~ significant urban features	5.327	3.633
crimes ~ all features	**3.732**	**2.614**
crimes ~ significant features	3.767	2.642

All models see their performance being improved with this method, comparatively to the previous one. Again, contrary to what happens in the linear regressions, the urban variables lose influence in predicting the number of crimes. The metrics of the socio-demographic regressors show much more favorable values than the urban ones, and the most complete logarithmic model, that is also exclusively constituted by socio-demographic covariates, records even lower metrics than both mixed models, being the best one. Among the models where *crimes* is in the natural form, the one that includes all variables is the most assertive, but the one that contains all socio-demographic features is very close to it.

Finally, random forests are performed, and the results are presented in Table 7. In this method, the hyperparameters corresponding to the number of features to consider at any given split and the minimum number of observations in a terminal node (complexity) are tuned using grid

search. The grid values of the first hyperparameter are 5, 10 and 15, and the values of the second mentioned are 10, 25 and 50. The number of trees is constant at 500. The best tuning values for both hyperparameters are, respectively, 15 and 25 for the socio-demographic models where *crimes* is not in logarithm, 15 and 50 for those where the dependent variable is logarithmic, 10 and 10 for the model that includes all urban covariates, 10 and 25 for the one that contains only significant urban features, and 15 and 25 for the mixed models.

Table 7: Metrics of random forest models

Random Forest Model	RMSE	MAE
crimes ~ all socio-demographic features	3.692	2.588
crimes ~ significant socio-demographic features	3.692	2.588
log(crimes) ~ all socio-demographic features	**3.471**	**2.318**
log(crimes) ~ significant socio-demographic features	3.490	2.319
crimes ~ all urban features	4.917	3.384
crimes ~ significant urban features	5.277	3.592
crimes ~ all features	**3.667**	**2.579**
crimes ~ significant features	3.671	2.581

This method is the one with the lowest metrics values, being the best technique to predict the number of crimes with any type of variables. Once more, the most complete logarithmic model is the one that records the best metrics. The great contradiction with the linear regressions is, as in the previous methods, in the performance of the models with only socio-demographic features, which values are very close to the mixed models, which are the best among those where *crimes* is not in logarithm. This reinforces that, in a non-linear relationship, the urban regressors do not have a large influence on the number of crimes, and the mixed models base their predictions practically on socio-demographics covariates.

5. CONCLUSION

Considering the results observed in the previous chapter, it is possible to conclude that, for any type of proposed relationship between the dependent and independent variables, and for any of the methods used to evaluate it, when the number of crimes is in a natural scale (the only way tested for every type of covariates), its prediction presents a smaller error when socio-demographic as well as urban characteristics are included in the models simultaneously. However, since this set of characteristics is precisely a combination of the two categories, it is equally important to understand which of them has a greater weight in this aggregate, that is, which of the two would lead to a more accurate prediction of the number of crimes in isolation.

From the point of view of a linear relationship between *crimes* and the regressors, through linear regressions, the urban variables are more effective in this prediction, with remarkable results compared to the socio-demographic variables. In a different perspective, in which the assumption of linearity is not assumed, resorting to more modern machine learning techniques to predict crimes, a reversal of those results can be noted, with the socio-demographic variables having a much higher performance than the urban ones. Despite drawing different conclusions, the second perspective overlaps the first, as the results of all non-linear models are clearly more favorable than the ones of the best linear model.

In this way, we conclude that the relationship between the variables used in this analysis and the number of crimes is not linear and, therefore, it is assumed that, to predict criminal activity, the characterization of a neighborhood at a social and demographic level is more important than the urban characterization. In the best modeling approach, that is the random forest, adding urban features to the socio-demographic ones only decreases the average prediction error (MAE) by 0.3% and the standard prediction error (RMSE) by 0.7%. Moreover, the best model of all is not even a mixed one. It is fully constituted by socio-demographic covariates and predicts the logarithm of the number of crimes. While these differences are small and might fail to be significant, it is a surprising result that the inclusion of urban features does not seem to help us improve our ability to predict crimes.

Regarding the study that had been referenced in the literature review [9], on which it was immediately indicated that an interesting association could be made in the conclusion, there is a similarity and a difference about the results obtained. That manuscript concludes that the inclusion of points of interest (equivalent to the so-called urban variables) significantly reduces the relative error in predicting crime, with a drop of 5%. This is in line with what is verified

here in the analysis of linear regressions, which makes perfect sense given that that study uses linear regression and negative binomial regression as the only forecasting methods, both based on linear relationships. In that same paper, no non-linear techniques are used to predict crimes, and it limits the analysis and does not allow a conclusion similar to the one drawn here: when non-linear methods are adopted, the inclusion of urban variables is not that much significant.

5.1. LIMITATIONS OF THE STUDY

Due to the type of methods used in the analysis, only correlation relationships are verified, which means that one of the limitations of this study is the lack of evidence of causal relationships.

Moreover, the focus of the study is on the number of crimes, but there is no specification of the type of crime. In this way, it is not noticeable whether a certain neighborhood is more or less dangerous than another by the value of that variable, in the same way that it is not known if a certain factor, having a significant influence on the number of crimes, affects more a specific type of crime.

Another limitation is present in the data collection, which, as it is carried out through annual census, does not allow the existence of distinct socio-demographic information for each analysis period (month). Although it is assumed that this information is the same for all months of each year, this does not correspond to reality, which may slightly influence the analysis and results of the study, even if the changes are not significant.

6. REFERENCES

[1] "Economic and Social Effects of Crime." Crime and Punishment in America Reference Library. *Encyclopedia.com*. December 20, 2022. https://www.encyclopedia.com/law/encyclopedias-almanacs-transcripts-and-maps/economic-and-social-effects-crime

[2] Ehrlich, Isaac. 1975. "On the Relation Between Education and Crime." *Education, Income, and Human Behavior* I: 313–38. http://www.nber.org/chapters/c3702.pdf.

[3] Freeman, Richard B. 1999. "Chapter 52 The Economics of Crime." *Handbook of Labor Economics*. doi:10.1016/S1573-4463(99)30043-2.

[4] PATTERSON, E. BRITT. 1991. "POVERTY, INCOME INEQUALITY, AND COMMUNITY CRIME RATES." *Criminology* 29 (4): 755–76. doi:10.1111/j.1745-9125.1991.tb01087.x.

[5] Best, Joel, and John Braithwaite. 1990. "Crime, Shame and Reintegration." *Social Forces* 69 (1). Oxford University Press (OUP): 318. doi:10.2307/2579648.

[6] Anselin, Luc. 2002. "Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27 (3). Wiley: 247–67. doi:10.1111/j.1574-0862.2002.tb00120.x.

[7] Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. 2011. "Self-Exciting Point Process Modeling of Crime." *Journal of the American Statistical Association* 106 (493): 100–108. doi:10.1198/jasa.2011.ap09546.

[8] Wang, Tong, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. 2013. "Learning to Detect Patterns of Crime." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8190 LNAI:515–30. doi:10.1007/978-3-642-40994-3_33.

[9] Wang, Hongjian, Daniel Kifer, Corina Graif, and Zhenhui Li. 2016. "Crime Rate Inference with Big Data." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:635–44. Association for Computing Machinery. doi:10.1145/2939672.2939736.

[10] Yuan, Jing, Yu Zheng, and Xing Xie. 2012. "Discovering Regions of Different Functions in a City Using Human Mobility and POIs." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 186–94. doi:10.1145/2339530.2339561.

[11] Graif, Corina, Andrew S. Gladfelter, and Stephen A. Matthews. 2014. “Urban Poverty and Neighborhood Effects on Crime: Incorporating Spatial and Network Perspectives.” *Sociology Compass* 8 (9). Wiley-Blackwell: 1140–55. doi:10.1111/soc4.12199.

[12] Ratcliffe, Jerry H. 2006. “A Temporal Constraint Theory to Explain Opportunity-Based Spatial Offending Patterns.” *Journal of Research in Crime and Delinquency* 43 (3): 261–91. doi:10.1177/0022427806286566.

[13] Toole, Jameson L., Nathan Eagle, and Joshua B. Plotkin. 2011. “Spatiotemporal Correlations in Criminal Offense Records.” *ACM Transactions on Intelligent Systems and Technology* 2 (4). doi:10.1145/1989734.1989742.

[14] Short, M. B., M. R. D’Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. 2008. “A Statistical Model of Criminal Behavior.” *Mathematical Models and Methods in Applied Sciences* 18 (SUPPL.): 1249–67. doi:10.1142/S0218202508003029.

[15] Bogomolov, Andrey, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. “Once upon a Crime: Towards Crime Prediction from Demographics and Mobile Data.” In *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, 427–34. Association for Computing Machinery, Inc. doi:10.1145/2663204.2663254.

[16] Wang, Xiaofeng, Matthew S. Gerber, and Donald E. Brown. 2012. “Automatic Crime Prediction Using Events Extracted from Twitter Posts.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7227 LNCS:231–38. doi:10.1007/978-3-642-29047-3_28.

[17] Chainey, Spencer, Lisa Tompson, and Sebastian Uhlig. 2008. “The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime.” *Security Journal* 21 (1–2). Springer Science and Business Media LLC: 4–28. doi:10.1057/palgrave.sj.8350066.

[18] John E. Eck, Spencer Chainey, James G. Cameron, Michael Leitner, and Ronald E. Wilson. 2005. “Mapping crime: Understanding hotspots.”

[19] Nakaya, Tomoki, and Keiji Yano. 2010. “Visualising Crime Clusters in a Space-Time Cube: An Exploratory Data-Analysis Approach Using Space-Time Kernel Density Estimation and Scan Statistics.” *Transactions in GIS* 14 (3): 223–39. doi:10.1111/j.1467-9671.2010.01194.x.

[20] Wang, Tong, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. 2013. “Learning to Detect Patterns of Crime.” In *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8190 LNAI:515–30. doi:10.1007/978-3-642-40994-3_33.

[21] Traunmueller, Martin, Giovanni Quattrone, and Licia Capra. 2014. "Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8851:396–411. Springer Verlag. doi:10.1007/978-3-319-13734-6_29.

[22] "Philadelphia's Top Ten Most Dangerous Neighborhoods." Electronic System for Travelling Authorization. August 6, 2022. https://estatousa.com/pt-pt/os-dez-bairros-mais-perigosos-de-filadelfia/

[23] "Crime rate in Philadelphia, Pennsylvania (PA)." City-Data. http://www.city-data.com/crime/crime-Philadelphia-Pennsylvania.html

[24] "Cidades americanas registram recordes de homicídios." Estado de Minas. December 20, 2021. https://www.em.com.br/app/noticia/internacional/2021/12/20/interna_internacional,1332631/cidades-americanas-registram-recordes-de-homicidios.shtml

[25] "Veja os assustadores números da violência armada nos EUA." Diário de Notícias. May 25, 2022. https://www.dn.pt/internacional/veja-os-assustadores-numeros-da-violencia-armada-nos-eua-14887734.html

[26] "Estados Unidos têm maior taxa de homicídios em 25 anos." Gazeta do Povo. October 5, 2022. https://www.gazetadopovo.com.br/mundo/estados-unidos-tem-maior-taxa-de-homicidios-em-25-anos/

[27] "Crime Maps & Stats." Philadelphia Police Department. https://www.phillypolice.com/crime-maps-stats/index.html

[28] "Explore Census Data." United States Census Bureau. https://data.census.gov/

[29] OpenStreetMap. http://openstreetmap.org