Physica A 505 (2018) 435–443

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Crime prediction through urban metrics and statistical learning

Luiz G.A. Alves [a,*], Haroldo V. Ribeiro [b], Francisco A. Rodrigues [a]

[a] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[b] Departamento de Física, Universidade Estadual de Maringá, Maringá, PR, Brazil

HIGHLIGHTS

- Predictive analysis is applied to crime data.
- Correlation between crime and urban metrics.
- Quantification of the importance of urban metrics in predicting crime.

ARTICLE INFO

Article history:
Received 23 November 2017
Received in revised form 26 February 2018
Available online 4 April 2018

Keywords:
Crime
Statistical learning
Complex systems
Random forest
Regression

ABSTRACT

Understanding the causes of crime is a longstanding issue in researcher's agenda. While it is a hard task to extract causality from data, several linear models have been proposed to predict crime through the existing correlations between crime and urban metrics. However, because of non-Gaussian distributions and multicollinearity in urban indicators, it is common to find controversial conclusions about the influence of some urban indicators on crime. Machine learning ensemble-based algorithms can handle well such problems. Here, we use a random forest regressor to predict crime and quantify the influence of urban indicators on homicides. Our approach can have up to 97% of accuracy on crime prediction, and the importance of urban indicators is ranked and clustered in groups of equal influence, which are robust under slightly changes in the data sample analyzed. Our results determine the rank of importance of urban indicators to predict crime, unveiling that unemployment and illiteracy are the most important variables for describing homicides in Brazilian cities. We further believe that our approach helps in producing more robust conclusions regarding the effects of urban indicators on crime, having potential applications for guiding public policies for crime control.

1. Introduction

Social phenomena increasingly attract the attention of physicists driven by successful application of methods from statistical physics for modeling and describing several social systems, including the collective phenomena emerging from the interactions of individuals [1], the spread of ideas in social networks [2], epidemic spreading [3], criminal activity [4], political corruption [5], vaccination strategies [6], and human cooperation [7]. In the particular case of crime, this interest trace back to works of Quetelet, who coined the term "social physics" in the 19th century [8]. On the one hand, traditional physics methods have proved to be useful in understanding phenomena outside conventional physics [9,10]. On the other hand, recently,

[*] Corresponding author.
E-mail address: lgaalves@usp.br (L.G.A. Alves).

https://doi.org/10.1016/j.physa.2018.03.084
0378-4371/© 2018 Elsevier B.V. All rights reserved.

several problems from physics have been addressed through the lenses of machine learning methods, including topics related to phases of matter [11], quantum many-body problem [12], phase transitions [13], phases of strongly correlated fermions [14], among others. As physicists had added these new tools to the box [15], naturally, social physics problems could also be addressed using such ideas. In particular, in this work, we are interested in understanding the relationships between crime and urban metrics by using statistical learning.

Crime and violence are ubiquitous in society. Throughout history, organized societies have tried to prevent crime following several approaches [16]. In this context, understanding the features associated with crime is essential for achieving effective policies against these illegal activities. Studies have linked crime with several factors, including psychological traits [17,18], environmental conditions [19,20], spatial patterns [4,21,22], and social and economic indicators [23–26]. However, it is easy to find controversial explanations for the causes of crimes [27]. Methodological problems in data aggregation and selection [28,29], errors related to data reporting [30], and wrong statistical hypothesis [27] are just a few issues which can lead to misleading conclusions.

A significant fraction of the literature on statistical analysis in criminology tries to relate the number of a particular crime (*e.g.* robbery) with explicative variables such as unemployment [31] and income [32]. In general, these analyses are carried out by using ordinary-least-squares (OLS) linear regressions [33]. These standard linear models usually assume that the predictors have weak exogeneity (error-free variables), linearity, constant variance (homoscedasticity), normal residual distribution, and lack of multicollinearity. However, when trying to model crime, several of these assumptions are, often, not satisfied. When these hypotheses do not hold, conclusions about the factors affecting crime are likely to be misconceptions.

Recently, researchers have promoted an impressive progress on the analysis of cities, where one of the main findings is that the relationship between urban metrics and population size is not linear, but it is well described by a power-law function [34–40]. Crime indicators scale as a superlinear function of the population size of cities [33,35,37]. Other indicators (commonly used as predictors in linear regression models for crime forecasting) also exhibit power-law behavior with population size. These metrics are categorized into sub-linear (*e.g.* family income [37]), linear (*e.g.* sanitation [37]), and super-liner (*e.g.* GDP [33,34,37,41]), depending on the power-law exponent characterizing the allometric relationship with the population size [34]. In addition, the relationships between crime and population size as well as urban metrics and population size have some degree of heteroscedasticity [34], and most of these urban indicators also follow heavy-tailed distributions [38,42]. Thus, it is not surprising to find controversial results about the importance of variables for crime prediction when so many assumptions of linear regressions are not satisfied.

A possible approach to overcome some of these issues is to apply a transformation to the data in order to satisfy the assumptions of linear regressions. For instance, Bettencourt et al. [35] (see also [33,37]) employed scaled-adjusted metrics to linearize the data and provide a fair comparison between cities with different population sizes. By considering these variables and applying corrections for heteroscedasticity [43], it is possible to describe 62% of the variance of the number of homicides in function of urban metrics [37]. Also, by the same approach, researchers have shown that simple linear models account for 31%–97% of the observed variance in data and correctly reproduce the average of the scale-adjusted metric [33]. However, the data still have co-linearities which can lead to misinterpretation of the coefficients in the linear models [33,37].

A better approach to crime prediction is the use of statistical learning methods (*e.g.* [44]). Regression models based on machine learning can handle all the above-mentioned issues and are more suitable for the analysis of large complex datasets [45]. For instance, decision trees are known to require little preparation of the data when performing regression [46–49]. Tree-based approaches are also considered a non-parametric method because they make no assumption about the data. Among other advantages, these learning approaches map well non-linear relationships, usually display a good accuracy when predicting data, and are easy to interpret [46–49].

Here, we consider the random forest algorithm [46,50–52] to predict and quantify the importance of urban indicators for crime prediction. We use data from urban indicators of all Brazilian cities to train the model and study necessary conditions for preventing underfitting and overfitting in the model. After training the model, we show that the algorithm predicts the number of homicides in cities with an accuracy up to 97% of the variance explained. Because of the high accuracy and easy interpretation of this ensemble tree model, we identify the important features for homicide prediction. Unlike simple linear models adjusted through OLS, we show that the importance of the features is stable under slight changes in the dataset and that these results can be used as a guide for crime modeling and policymakers.

2. Methods and results

2.1. Data

For our analysis, we choose the number of homicides at the city level as the crime indicator to be predicted. Homicide is the ultimate expression of violence against a person, and thus a reliable crime indicator because it is almost always reported. In Brazil, the report of this particular crime to the Public Health System is compulsory, and these data are aggregated at the city level and made freely available by the Department of Informatics of the Brazilian Public Health System — DATASUS [53]. As possible predictor variables of crime, we select 10 urban indicators (also at the city level) available from the Brazilian National Census that took place in 2000. They are: child labor (fraction of the population aged 10 to 15 years who is working or looking for work), elderly population (citizens aged 60 years or older), female population, gross domestic product (GDP), illiteracy (citizens aged 15 years or older who are unable to read and write at least a single ticket in the language they