BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018

DATA-POP ALLIANCE

1. Socio-physical Vulnerability to Flooding in Senegal.
2. Characterizing and analyzing urban dynamics in Bogota.
3. Understanding the Relationship between Short and Long Term Mobility.
4. Large-Scale Mapping of Citizens' Behavioral Disruption in the Face of Urban Crime Shocks.

Partners:

DATA-POP ALLIANCE

Cloud to Street

FLOWMINDER.ORG

mit media lab

HARVARD HUMANITARIAN INITIATIVE

Funded by:

AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

BIG DATA
TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES
2018

DATA-POP ALLIANCE

CHARACTERIZING AND ANALYZING URBAN DYNAMICS IN BOGOTA

Marco De Nadai - University of Trento, denadai@fbk.eu
Emmanuel Letouzé - Data-Pop Alliance, eletouze@datapopalliance.org
Marta C. González - MIT, martag@mit.edu
Bruno Lepri - FBK, lepri@fbk.eu

Funded by:

DATA-POP ALLIANCE · FONDAZIONE BRUNO KESSLER · mit media lab · AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

CHARACTERIZING AND ANALYZING URBAN DYNAMICS IN BOGOTA

Marco De Nadai,
University of Trento,
denadai@fbk.eu

Emmanuel Letouzé,
Data-Pop Alliance,
eletouze@datapopalliance.org

Marta C. González, MIT,
martag@mit.edu

Bruno Lepri, FBK, lepri@fbk.eu

Please cite this paper as: DE NADAI, M., E. LETOUZÉ, M.C. GONZÁLEZ and B. LEPRI (2017), "Characterizing and analyzing urban dynamics in Bogota", *AFD Research Paper Series*, No. 2018-70, June.

Abstract

Containing crime without affecting the livability of the urban environment is a major challenge in our society. Traditionally, researchers relate crime to socio-economic disorganization and people's routine activity, as it influences effective control and suitable targets. An important open question is what the role the urban fabric plays. Although empirical research has shown that the physical urban environment is an essential factor for urban vitality and health, we lack evidence of any clear relationship between the structural characteristics (e.g. roads and land use mix) of neighborhoods and crime. Here, by using open data and mobile phone records, we explore this link with a spatial regression model that analyzes the environmental and the social conditions to which each part of the city is exposed. We found that physical characteristics of the city connected to higher urban diversity better explain the emergence of crime than traditional socio-economic conditions and, together, physical characteristics and socioeconomic conditions improve the performance of previous approaches. This result suggests that urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited in urban planning to reduce crime.

Key words: Welfare programs, Well-Being, Poverty, Family Structure and Planning, Economics of Minorities, Races, Indigenous Peoples and Immigrants, Non-labor Discrimination, Demand and Supply of Labor, Urban and Transportation Analysis, Housing, Land Use in Urban Economies

JEL Classification: I38, J12, J13, J15, J18, J21, J23, J61, J64, J68, J71, N36, O18, R14, R21, R23, R28, R41, R52

Original version: English

Accepted: June 2017

TABLE OF CONTENTS

Characterizing and analyzing urban dynamics in Bogota

Marco De Nadai[1,2], Andrés Clavijo[4], Marta C. González[3], Emmanuel Letouzé[4] and Bruno Lepri[2]

Containing crime without affecting the livability of the urban environment is a major challenge in our society. Traditionally, researchers relate crime to socioeconomic disorganization and people's routine activity, as it influences effective control and suitable targets. An important open question is what the role the urban fabric plays. Although empirical research has shown that the physical urban environment is an essential factor for urban vitality and health, we lack evidence of any clear relationship between the structural characteristics (e.g. roads and land use mix) of neighborhoods and crime. Here, by using open data and mobile phone records, we explore this link with a spatial regression model that analyzes the environmental and the social conditions to which each part of the city is exposed. We found that physical characteristics of the city connected to higher urban diversity better explain the emergence of crime than traditional socioeconomic conditions and, together, physical characteristics and socioeconomic conditions improve the performance of previous approaches. This result suggests that urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited in urban planning to reduce crime.

[1] University of Trento.
[2] FBK.
[3] MIT.
[4] Data–Pop Alliance

1. Introduction

The rapid growth of cities and the increase of population mobility have challenged our ability to understand crime. The primary focus of criminology research has been on people that commit crimes, and the reason they are involved. For crime to happen three conditions are supposed to be present and interact: the presence of a motivated offender who is willing to commit a crime, a suitable target, and the absence of guardians that would otherwise prevent the crime from taking place [6]. In this equation crime offenders, victims and guardians are all affected by socioeconomic conditions, the social disorganization (e.g. unemployment) of community [9] and the place where they intersect. Thus, place matters.

Environmental criminology suggests that place not only is logically required, but also influences the likelihood of becoming a crime hotspot through its physical characteristics. Accordingly, place is one of the five necessary and sufficient components that constitute a criminal incident, namely place, time, law, offender and victim [3]. Thus, environmental criminologists are interested in land use, street design, traffic patterns and daily activities of people. However, scholars have virtually ignored other theories (e.g. social disorganization), and bounded their discussion to macro-areas of few cities.

Urban planners and sociologists argue that cities are not a mere artificial construction that group people. A city is a vital process of the people who compose it; and its neighborhoods are the elementary form of cohesion in urban life [15]. One of the seminal books in city planning is Jane Jacobs' The Death and Life of Great American Cities [13]. In this book she introduced the concept of eyeson-the-street, which suggests that safety can be maintained by citizens through urban surveillance. For this to work, some physical qualities need to be present in the neighborhoods (i.e. a mix of residential, commercial and recreational land uses) to guarantee the diversity and continuous presence of people throughout the day. It is thus clear the tight coupling of environmental criminology and urban planning theories.

Traditional approaches on describing crime have failed to provide a clear and broad description of the desirable characteristics the different parts of the city should possess to keep crime events low.

In the present study we seek to fill this gap, formalizing the hypothesis that physical characteristics of the city not only are related to better life conditions [17] and vitality [7], but also greatly influence crime. Thus, we create two types of models. One is focused on *describing* how physical characteristics influence crime in each part of the city. The other *predicts* crime events in a city from the structural features, and answers to the question "can we predict crime from the physical characteristics of the city?". Thanks to several new sources of data and a Negative Binomial model, we model crime through physical and socioeconomic characteristics, but also spatial and routine-activity information.

We find that structural characteristics of the city, namely Jacobs' diversity conditions, are better predictors than socioeconomic status, and that these results are robust across different spatial aggregations. Also, we find that the number of inhabitants' routine movements between the neighborhoods is closely related to crime, with highly-connected points of the city experiencing a higher number of crimes, as suggested by the routine-activity theory [6]. Finally, we observe that the combination of structural, routine and socioeconomic information provides better estimates of crime than each on its own. Together, these observational results suggest that the city structure has a strong connection with crime, and that improving its qualities can discourage criminality.

Thus, our main contributions are: i) we focus on place in a new fashion and show how physical characteristics greatly influence crime events; ii) we built a now-casting model, which is portable from one city to many, able to predict crime counts in a city; iii) we employ new sources of data and combine multiple criminology theories in Bogota.

This paper is organized as follows: in Section 2 we review the literature in this field. In Section 4 we outline the proposed approach and the evaluation process. Finally, we show our results in Section 5, before discussing the implications and limitations in Section 6.

2. Related work

One of the most prolific place that established the hallmark of environmental criminology is Chicago. In the University of this city, sociologists and criminologists started to consider neighborhoods as unit of analysis, both from the social and political (or administrative) perspective.

Ernest Burgess developed a concentric-zone model, based price changes in housing values, to study crime patterns in Chicago [4]. He observed that the distribution of social problems and crime vary in respect to the distances to the center. On the basis of this model, Clifford Shaw extensively researched how young people, juvenile delinquency and adult offenders were distributed in space [8]. He introduced the *spot maps*, *delinquency rate maps*, *radial maps and zone maps* that established a landmark on crime mapping.

In the recent years, many empirical and predictive studies flocked thanks to new methods mainly coming from computer science, and the increasing availability of new sources of data. Graif and Sampson [11] examined the connection of immigration and socioeconomic diversity to homicide. Thanks to a spatial model they found that immigrant concentration is either unrelated or inversely related to homicides, whereas language diversity is negatively correlated to homicides. Verifying how diversity of people influences crime in neighborhoods was the goal of the study of Traunmueller *et al*. [19]. To observe *people dynamics* they used mobile phone records broken by age and gender. They observed that age-diversity

and presence of non-residents are linked to lower criminality. Bogomolov *et al*. [2] used mobile phone data in a similar fashion, and predicted crime *hotspots* thanks to a Random Forest regression.

By only using the ambient population characteristics extracted from mobile phone records they were able to predict with almost 70% of accuracy whether an area would have high or low crime levels in time. The assumption of observing ambient population ignoring the movements of people in the city was tackled by Graif *et al.* [9]. They argued that people are continuously exposed to different neighborhoods, and they proposed a *network of neighborhoods* to describe this. This approach is considered important for understanding how changes of activity spaces can influence crime.

Strikingly, very few scholars considered multiple sources of data and theories in their discussions. Moreover, the importance of place is limited to spatial-autocorrelation or the topology of street patterns (e.g. [16]). Contrairily to this, Wang *et al*. [20] examined crime in Chicago by leveraging census counts as well as new sources of data available on the web, such as crowd-generated Points Of Interest (POI) and taxi flows. POIs were expected to be linked to higher crime incident as they represent suitable targets; taxi flows were considered as proxy for trips made by humans. The incremental now-casting model built by Wang *et al*. showedthat crime rate can be estimated with a relative error of 30% by using demographic information, but it could be also reduced by 5% with the POI data. Finally, he showed that the complexity of the crime in Chicago could be further explained by the taxi flow network, which further improved the results by 5%.

3. Description of the data

3.1 Spatial data

Bogotá, the capital city of Colombia, is divided into 20 localities, each of which contains between 1 to 12 zonal planning units, known as UPZ in Spanish. In total, there are 113 UPZ. Each UPZ, in turn, is divided into neighborhoods, which are themselves composed of a set of blocks.

In addition to the UPZ scheme, Bogotá also has spatial divisions defined by the national census. The designations used by the census are (from largest to smallest): urban sector, urban section, and block. A sector is a cartographic census division, roughly equivalent to a neighborhood (especially for large cities), and comprising between 1 and 9 sections. Each section is composed of approximately 20 contiguous blocks, all falling within the same sector. Finally, a block is a lot of land, built or unbuilt, bounded by public paths, roads, crosswalks, etc. Blocks may also be bounded by a natural feature such as a river, stream or channel, as long as it is permanent and easy to locate in the field.

Our main source of spatial data for this analysis is the Capital District's Spatial Data Infrastructure dataset, known as Infraestructura de Datos Espaciales del Distrito Capital (IDECA) in Spanish. Its function is to facilitate the access to geographic information about Bogot´a and support its social, economic, and environmental development.

The IDECA dataset used for this study is a compilation of 10 spatial datasets which cover the following topics: buildings, lots, blocks, localities, land use, points of interest, strata, transport nodes, cycling trails, and road network.

3.2 Socioeconomic and crime data

The main sources for socioeconomic and crime data are the National Statistics Office of Colombia, known as Departamento Administrativo Nacional de Estadística (DANE) in Spanish, and the National Police of Colombia. We use the population census and multipurpose survey from DANE and crime data from the National Police.

3.2.1 Census data

We used census data from the last census held in Colombia in 2005, and the population projections made by DANE for 2015. The Secretary of Planning of Bogotá, using DANE's projections, issued projections of the 2015 population of Bogotá by UPZ. From the census, we have access to socioeconomic data at the block level. In Bogotá, this census was conducted in 1,931,372 households, distributed across 37,473 blocks, and includes 6,778,691 people. The projections estimate a rise in Bogotá's population to 7,878,783 people in 2015.

3.2.2 Multipurpose Survey

The Multipurpose Survey, known as Encuesta Multipropósito (EM) in Spanish, was performed in 2014 by DANE, and financed by the Secretary of Planning of Bogotá. Its objective was to obtain statistical information on social, economic, and environmental aspects of urban households and residents of Bogotá. DANE uses the EM to derive income data, multidimensional poverty indixes, and subjective poverty indexes, among others. The information from this survey is statistically representative at a locality level (which is less granular than the UPZ).

3.2.3 SISBEN Survey

The Identification System of Potential Welfare Recipients, known as Sistema de Identificación de Potenciales

Beneficiarios de Programas Sociales (SISBEN) in Spanish, was created with the purpose of reducing the cost of targeting social benefits receivers and keeping track of these in the whole country.

The SISBEN survey is the tool through which individuals are categorized as recipients of social aid. It contains a set of variables related to durable goods consumption, human capital endowment and current income. The 2012 survey contains 3,545,789 observations distributed through different Colombian departments; and 123 variables, most of them related to socioeconomic conditions.

It will be possible to localize the amount of beneficiaries for each block in Bogotá using the 2012 survey, and thus generate a more granular characterization of socioeconomic status in different zones of Bogotá. The Non-Disclosure Agreement that allows using this data will be signed soon.

3.2.4 OD network

The Origin-Destination network is extracted from mobile phone data, automatically collected for billing purposes. We use data from the largest telecommunications operator in the area. For each area (UPZ) a and b we aggregate the number of people that move from a to b in a typical day. This results in a matrix were for each UPZ we have the information about the routine activity of people, which is related to both the presence of people and the risk of victimization.

3.2.5 Crime data

The criminal cases dataset includes geo-located and timestamped records of reported crime in Bogotá. It consists of 27,863 criminal cases for homicide and theft (burglaries of commercial property, burglaries of houses, and robberies) for 2014. Specifically, the dataset includes the category and subcategory of the crime, the longitude, latitude, and address of where the crime was reported to have occurred, and the responsible police department.

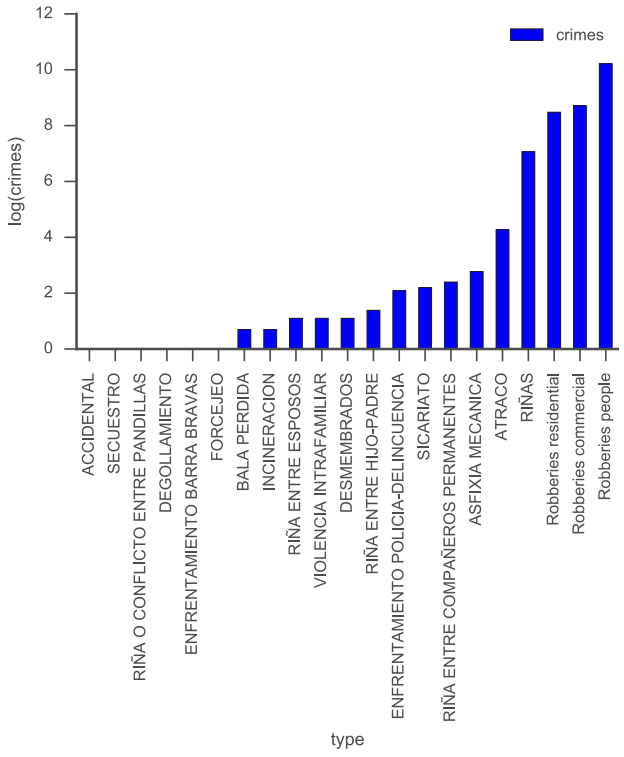

Figure 1: Crimes per type in Bogotá.

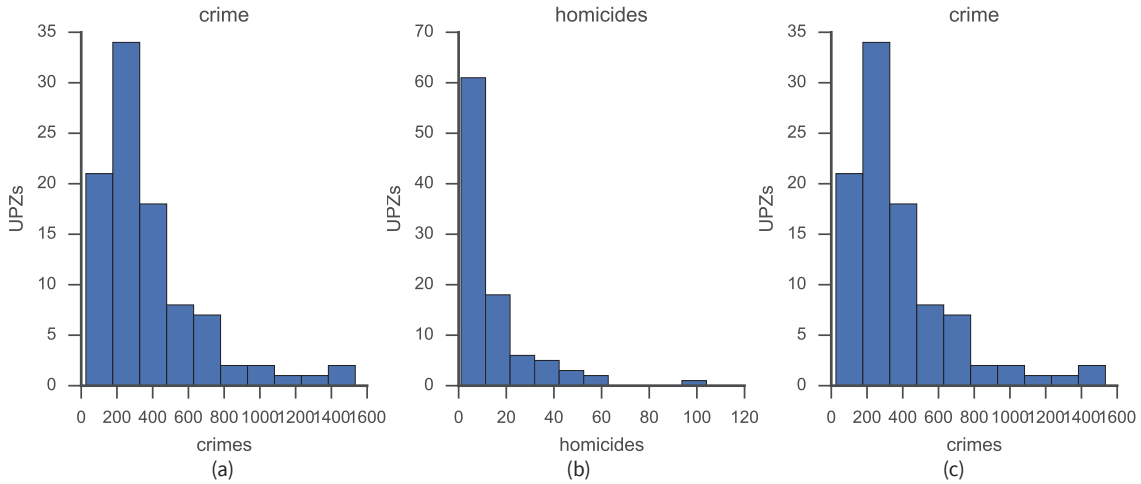

Figure 2: a) Crime distribution in the UPZ; b) Violent crime distribution in the UPZ; c) Robberies distribution in the UPZ.

4. Methods

Crime is considered a rare occurring phenomenon: a small proportion of people are victimized, with also few unreported and undiscovered crime. Crime offenses are not distributed evenly in space; they tend to cluster in parts of the city that exhibits similar characteristics, favorable to the spread of crime. Rare events, expressed through count variables, have been widely modeled on Ordinary Least Squares (OLS) through a logarithmic and square root transformation of the response variable. However, this introduces extra estimation bias, homoschedability assumptions and impossible predictions [14]. Moreover, spatial auto-correlation is rarely accounted for. For this reason, we use a Negative Binomial regression (NB) for count variables, and an eigenvector spatial filter (ESF) to account for spatial auto-correlation.

In this paper we want to *describe* a city but also to be able to *predict* it. Differently from the common meaning, here we use the word "predict" to mean the estimation of a property not directly observed (nowcasting).

Thus, we first employ a descriptive model for Bogota where it is possible to understand the interactions of each component of the model to describe crime events. Then, we create a predictive model validated with a 5-fold Cross-validation with 1000 repetitions (to avoid overfit). This allows to answer to the question "can I predict crime events from the characteristics of the city?".

4.1 A regression model for rare events

The Poisson regression model, a class of Generalized linear models (GLM), is particularly attractive to model count response variable. However, its restrictive assumption to have identical mean and variance is violated with many real-world situations. The NB introduces an extra parameter k to the Poisson model with parameter λ_i , where the variability of λ_i has a gamma distribution with mean μi and index ν. It follows that $Y_i \sim NB(k, \mu_i/k)$ and:

$$E(Y_i) = \kappa\frac{\mu_i}{\kappa} = \mu_i \qquad (1)$$

$$Var(Y_i) = \mu_i\frac{\mu_i^2}{\kappa} \qquad (2)$$

with k accounting for the extra variability with a quadratic function on μ_i. and as As $k \to \infty$ the distribution of Y_i converges to a Poisson random variable. With a log-link function on μ_i, the NB model can be written as:

$$Ln[E(Y_i)] = Ln[E(i)] + \beta_0 + \sum_{k=1}^{n} X_k B_k \qquad (3)$$

where $X_0, X_1, \ldots, X_n$ are the covariates, $\beta_0, \beta_1, \ldots, \beta_n$ the regression parameters, and $LN[E(i)]$ is the offset variable and its role is to control for size differences across the units.

4.1.1 Spatial auto-correlation

Models dealing with spatial data analysis require to test for spatial auto-correlation. Positive (negative) spatial auto-correlation refers to the attitude of nearby attributes to have similar (dissimilar) values. There are numerous quantitative methods to measure and deal with spatial auto-correlation. The eigenvector spatial filter (ESF) [12] introduces a set of independent variables that account for the spatial relationship of the variables. These variables are a subset of the eigenvectors extracted from the numerator of the Moran's I coefficient [5]:

$$(I - \frac{11^T}{n})W(I - \frac{11^T}{n}) \qquad (4)$$

where I is a $(n \times n)$ identity matrix, 1 is a $n \times 1$ vector of ones, and W is a generic $(n \times n)$ distance matrix. This matrix can we either be defined with geographical distance or flow of people and freight.

To model the proximity of each region, we define the W distance matrix as the inverse squared distance separating the observations:

$$w_{i,j} = \begin{cases} 0 & if i = j \\ 1/d_{i,j}^{\gamma} & \text{otherwise} \end{cases} \qquad (5)$$

where $d(i, j)$ is the Euclidean distance between i and j, and i is γ a penalization parameter. We set γ=2 to place a greater weight on close observations and leave a marginal role to distant observations. The matrix is variance stabilized.

The n eigenvectors describe the full-range of uncorrelated spatial patterns; but employing all n eigenvectors in a regression framework is not desirable for reasons of model parsimony. Thus, we first filter the eigenvectors with low (and opposite) Moran's I ($MI_E/MI_{max} \geq 0.25$). Then, from this subset we select the smallest subset of eigenvectors $\{E_1, E_2, \ldots, E_p\}$ each time adding, in a step-wise fashion, the eigenvector that reduces the most the Akaike Information Criterion (AIC) of the model. The process stops when the AIC does not decrease anymore. The step-wise process does not guarantee to select the *best* eigenvectors, but it is a very simple and fast method to select the orthogonal eigenvectors to add. For further details on eigenvector selection and implementation strategies see Tiefelsdorf *et al.* [18]. The final subset of candidate eigenvectors represents the spatial filter for the variable analyzed. Thus, the aspatial NB model defined in Equation (3) takes the form:

$$Ln[E(Y_i)] = Ln[E(i)] + \beta_0 + \sum_{k=1}^{n} X_k B_k + \sum_{j=1}^{p} E_j B_{n+j} \qquad (6)$$

4.2 Spatial aggregation

Bogota is regionally recognized for the important strides it has made in reducing violence related to organized crime in recent years. To understand the multitude of aspects that influence crime in a neighborhood, we first have to define it. A neighborhood is a geographical unit composed by people who usually interact with each-other, and sharing common goals. This spatial community has a loose definition [13] and it has to be "big enough (in population) to swing weight in the city as a whole, but small enough so that street neighbourhoods were not lost or ignored". From this description we selected the Unidades de Planeamiento Zonal (UPZ) as a valid spatial aggregation for neighborhoods. The function of a UPZ is to help in the planning in the development of urban norms in the city. There are 113 UPZ and their average population is 63,720, with average area of 3.7 in2009.

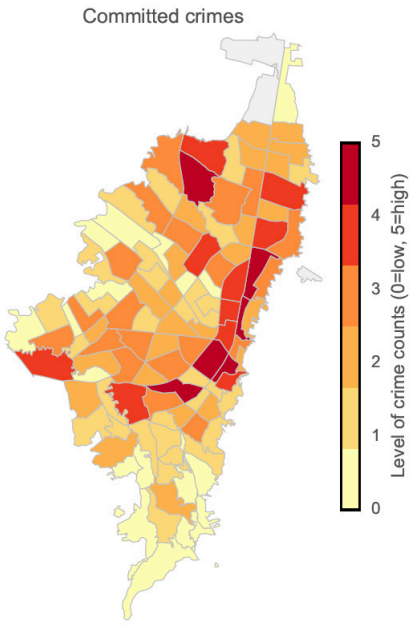

Figure 3: Crime counts in Bogota for each UPZ unit.

4.3 Measure of crime

Crime counts represent a measure to map and understand, in absolute terms, where the crime takes place. Nonetheless, crime rates over residential population are usually preferred as they assess the risk of people to be victimized in a particular location [1]. In a NB model, a rate variable is defined with an offset, which is a variable that is forced to have a coefficient of 1 in the model. This is particular useful to create risk maps, but we think it is a too restrictive constraint. To achieve a greater flexibility we prefer to add population as covariate and estimate its coefficient in the model. A population coefficient (B_p) greater than one means that spatial units with more population have higher rate of criminality. On the contrary, units with less population have fewer crimes per inhabitants when $B_p < 1$.

4.4 Daily routines

As aforementioned, daily routines are supposed to influence the presence of offenders, victims and guardians in a place. Moreover, frequent trips between two places are supposed to influence each other's crime.

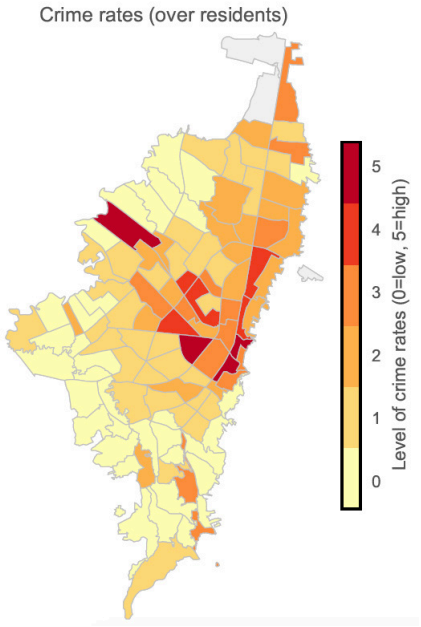

Figure 4: Crime rate ($|crimes|/(\beta_p|population|)$) Bogota for each UPZ unit.

Wang *et al*. [20] used the taxi flow network with a spatial lag regression. We instead use the same ESF method used to take into account the spatialautocorrelation. Firstly we compute the total number of trips made in a typical day between spatial units. Then, we define the symmetric weight matrix W_t from the Mobility network dataset, by applying this transformation to the original mobility network W_m:

$$W_t = W_m^T W_m \tag{7}$$

The eigenvectors inserted in the NB model are supposed to be a proxy for the mobility network dependencies between spatial units.

4.5 Jane Jacobs' diversity

As aforementioned, our hypothesis is that Jane Jacobs' diversity conditions have an impact on crime measures. Thus, in this section we describe the metrics, priorly validated [7, 17], that operationalize the Jacobs' theory.

Land use mix. A common way of quantifying the mixture of land uses is through average Shannon entropy. The average entropy, which we here call Land Use Mix (LUM), is defined as:

$$\mathrm{LUM}_{3L,i} = -\sum_{j \in 3L} \frac{P_{i,j} \log(P_{i,j})}{\log(|3L|)} \tag{8}$$

where $P_{i,j}$ is the percentage of square meters having land use j in unit i, and $3L$ = { residential, commercial and istitutional, park and recreational} represents the land uses considered. The LUM ranges between 0, wherein the unit is composed by only one land use (e.g. residential), and 1, where in developed area is equally shared among the n land-uses. The problem with this index is that it depends on the way land uses are grouped together, and no distinction is made on the order of land uses. Thus, an entropy of 0.75 could either mean high land use mix with a major role of residential land use, or high land use mix with a major role of parks. To better represent the different outcome we also employ a second entropy measure, based on the distinction between residential and non-residential land uses:

$$\mathrm{LUM}_{rnr,i} = -\sum_{j \in rnr} \frac{P_{i,j} \log(P_{i,j})}{\log(|rnr|)} \tag{9}$$

where rnr = {residential, non-residential, }.

Jacobs argued for mixing primary uses so that people are on the street at different times of the day. To characterize spatial use in terms of activities, we determine whether each place is used daily (e.g., convenience stores, restaurants, sport facilities) or not. Based on that, we define the average accessibility of the buildings in a spatial unit i as:

$$\mathrm{A}_i = \frac{1}{|B_i|} \sum_{j \in B_i} \mathrm{dist}(j, \mathrm{closest}(j, D))^{-1} \tag{10}$$

where D is the set of places that are used on a daily bases (e.g. convenience and grocery stores), dist(a, b) is the Euclidian distance between a and b, closest(a, C) is a function that finds the closest item in set C from point a, and B_i is the set of buildings in unit i.

Consistent with the methodology used by the website developers to calculate Walk Score[1], we define the weighted *walkability* score as:

$$\mathrm{walk}_i = \frac{1}{|B_i|} \sum_{c \in C} w_c \sum_{b \in B_i} \mathrm{wdist}(b, \mathrm{closest}(b, \mathrm{POI}_c))^{-1} \tag{11}$$

where C = {Grocery, Food, NightLife, Shops, Cultural}, POI_c is the set of POIs of category c, and w_c is the weight of importance to POIs, which is 3 for Food, Nightlife and Grocery POIs, 2 for Shops and 1 for others. wdist(a, b) is a linear function from 1, when the dist(a, b) <= 500 m, to 0 when the dist(a, b) >= 2500 m.

Small blocks. Small blocks are believed to support stationary activities and provide opportunities for short-term and low-intensity contacts, easing into interactions with other people in a relaxed and relatively undemanding way. We compute the average block area among the set B_i of blocks in unit i as:

$$\text{Blocks area}_i = \frac{1}{|B_i|} \sum_{b \in B_i} \text{area}(b) \qquad (12)$$

In addition, we compute the average distance between each building and the nearest street, a proxy for the concept *eyes on the street*, which suggest that the safety of neighborhoods can be maintained through continued surveillance of their inhabitants For each unit i and the set S of streets, it is defined as:

$$\textit{Eyes on the street}_i = \frac{1}{|B_i|} \sum_{b \in B_i} \text{dist}(b, \text{closest}(b, S))^{-1} \qquad (13)$$

Buildings. Jacobs stressed the importance of having diverse buildings in a district to create vitality. Diverse buildings allow the mix of different socioeconomic groups, as well as the tendency of an easier accommodation of creative people and small enterprises.

Colombia has a fiscal policy, called stratum, that classifies buildings in different regimes of tax payments for utilities and rents. Stratum is based on the external physical characteristics of the building, and it reflects the quality of life of residents with a six-level classification from 1 (lower low) to 6 (high). For this reason, we computed the heterogeneity of a unit i as:

$$\text{Strata}_{\sigma_i} = \sqrt{\frac{1}{|H_i|} \sum_{b \in H_i} (\text{strata}_b - \overline{\text{Strata}_i})^2} \qquad (14)$$

where H_i is the set of houses belonging in district i.

Concentration. Jacobs' fourth and final condition is about having concentration of both buildings and people. First, we determine population density measures by dividing the number of people by the unit's net area.

$$\text{Population density}_i = \frac{|\text{Population}_i|}{\text{area}_i} \qquad (15)$$

Then we compute the floor-area density per each unit i as:

$$\text{Buildings density}_i = \frac{|\text{Buildings}_i|}{\text{area}_i} \qquad (16)$$

[1] http://www.walkscore.com

Border Vacuums. Border vacuums are places that act as physical obstacles to pedestrian activity. For instance, parks can be a hub of pedestrian activity, if efficiently managed [13], but they could also be deplorable places in which criminality flourishes (especially at night). Thus, we compute the average closeness of each building from the nearest park as:

$$\text{Closeness to LP}_i = (\frac{1}{|B_i|}\sum_{j \in B_i} \text{dist}(j, \text{closest}(j, LP)))^{-1} \tag{17}$$

where dist(j, closest(j, LP) is the distance between block j and its closest large park.

4.6 Covariates

The number of committed crimes is mainly influenced by the number of residents, their social disorganization and routine activity. Social disorganization is the inability of the neighborhood to maintain effective social control. Social disorganization is higher in deprived areas, social heterogeneous units and in places with high unemployment rate, which is also a proxy for motivated offenders. We define the social heterogeneity through unemployment rate:

$$\text{unemployment}_i = \frac{|\text{unemployed residents}_i|}{|\text{residents}_i|} \tag{18}$$

and a proxy of income heterogeneity through the weighted standard deviation of property values:

$$\text{social heterogeneity}_i = \sqrt{\frac{\sum_{b=1}^{n} w_b^2}{\left(\sum_{b=1}^{n} w_b\right)^2}\sigma^2} \tag{19}$$

where x_b is the property value of a block in spatial unit i and x_b is the residential population count of block b.

The number of residents is calculated as:

$$\text{population}_i = |\text{residents}_i| \tag{20}$$

4.7 Evaluation

The R^2 statistic is an intuitive interpretation of the proportion total variation in outcome that is accounted for by a OLS model. Concerning GLMs there is no directly analogous R^2 measure. For this reason, GLMs models are usually evaluated through their AIC, Pseudo-R^2 and Root Mean Squared Error (RMSE). One of the most interesting Pseudo-R^2 measures is the McFadder Pseudo-R^2, which:

$$\text{McFadder Pseudo-}R^2 = 1 - \frac{\log \widehat{L}(M_{full})}{\log \widehat{L}(M_{intercept})} \tag{21}$$

where $\hat{L}(M_{full})$ is the log likelihood of the full model and $\hat{L}(M_{intercept})$ is the log likelihood of the null model. It is worth to remember that this is not a true measure of fit, because it only compares the log likelihood of the full model with the one of the null model.

In the predictive model, evaluated through the K-fold Cross-validation (K=5), we create multiple models that use a subset of the features. Thus, the subsets are:

- Socioeconomic: demographic and social disorganization variables;
- City: Jane Jacobs' diversity variables;
- Dynamics: daily routines variables;

and the combinations of them. This allows to understand the contribution of each subset to the description of the crime events in a city.

5. Results

Our contribution in this paper is twofold. At first we focus on describing the relation of the various factors with crime. Then, we built a predictive model to understand how the results can be generalize in different cities.

5.1 Descriptive model

From the results (in Table 2) we can observe the β coefficients of features to understand the importance of each variable, holding the others as constant. The most important variables to describe crime are building density, population density and closeness to daily-use buildings. Particularly, the concentration variables are very important in describing crime, with building density that is positively correlated with crime (0.498). By contrast, the higher the population density is, the less crime events there are.

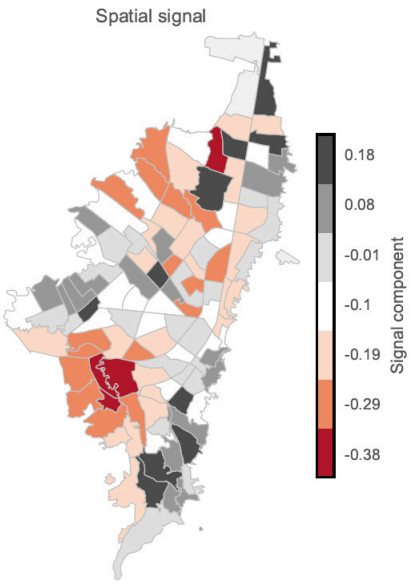

Figure 5: Stochastic signal component B_eE representing the spatial auto-correlation for each UPZ in Bogota.

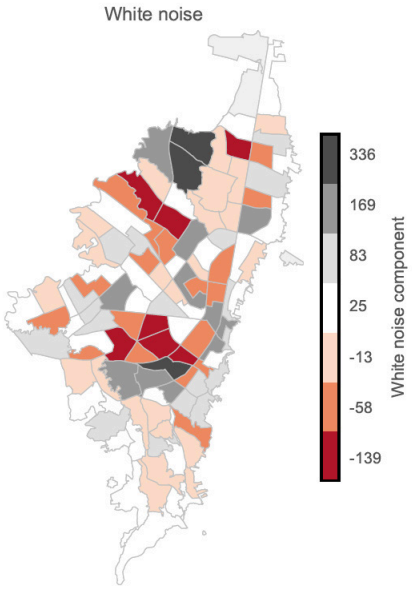

Figure 6: White noise component $y - (XB + B_eE)$ representing the undetected number of crimes for each UPZ in Bogota.

	β coefficient	std. error	p-value	95% CI
Land use				
Land use mix (LUM_3L)(8)	-0.078	0.029	**	[-0.135, -0.021]
Land use mix (LUM_{rnr})(9)	-0.098	0.034	**	[-0.166, 0.032]
Closeness daily buildingsl(10)	0.212	0.035	***	[0.144, 0.281]
Walkability score (11)	0.146	0.060	*	[0.029, 0.263]
Small blocks				
Block area (12)	-0.033	0.054		[-0.140, 0.073]
Eyes on the street (13)	0.270	0.055	***	[0.163, 0.378]
Buildings				
Strata diversity^s(14)	0.070	0.030	*	[0.011, 0.129]
Concentration			***	
Population density (15)	-0.363	0.045	***	[-0.451, -0.276]
Building density (16)	0.498	0.054	***	[0.393, 0.603]
Vacuums				
Closeness parks^l(17)	0.031	0.039		[-0.045, 0.108]
Covariates				
Unemployment (18)	0.110	0.036	**	[0.039, 0.181]
Population (20)	0.6647	0.058	***	[0.550, 0.779]
Social heterogeneity (19)	0.159	0.041	***	[0.078, 0.240]
OD eigenvectors (Sec. 4.4)	-0.184	0.042	***	[-0.267, -0.101]
Spatial eigenvectors	2			
McFadder Pseudo-R^2 †	0.145			
RMSE	92.81			
Moran's I (p-value)	0.02 (0.42)			

† This is not a true measure of fit, and not bounded to 1. It indicates the degree to which the model parameters improve upon the prediction of the null model.

Table 1: Negative Binomial regression model that describes the number of crimes in each spatial unit.

	S	C	D	S+D	C+D	S+C	FULL
McFadder Pseudo-R^2 †	0.077	0.113	0.085	0.106	0.120	0.141	**0.143**
RMSE	231.93	145.04	312.70	181.76	133.36	143.35	**127.76**

† This is not a true measure of fit, and not bounded to 1. It indicates the degree to which the model parameters improve upon the prediction of the null model.

Table 2: Negative Binomial regression models that predict the number of crime in each spatial unit. The results are average across 1000 iterations of a 5-fold Cross-validation. S: demographic and social disorganization variables only; C: Jane Jacobs' diversity variables only; D: daily routine variables only; S+D: demographic, social disorganization and daily routine variables; C+D: Jane Jacobs' diversity and routine variables; S+C: demographic,

Small blocks are also very related to crime, as the distance of buildings from the nearest street has a positive correlation with crime events (0.270). This is in accordance to the *eyes on the street* theory of Jane Jacobs, that generates a virtuous loop which, in turn, increases public safety.

Covariates coming from criminology literature are also significative. We found that Social heterogeneity and unemployment are positively correlated with crime. Thus, higher deprivation and disorganization might stop the mechanisms by which residents themselves achieve guardianship and public order.

Finally, we find that social isolation increase crime, as it supposedly limits neighborhood possibilities and social capital [10]. Thus, we see that highly-connected points of the city experiencing lower number of crimes.

5.2 Predictive model

Our preliminary findings (see Table 2) indicate that structural characteristics of the city, namely Jacobs' diversity conditions, are a better predictor of the target variables (the number of homicides and robberies) than socioeconomic conditions such as unemployment and deprivation. Mobility networks, and thus the routine activity theory, improve the prediction of the model by 15%. The combination of structural and socioeconomic information provides better predictions than each on its own.

6. Discussion

In this paper we modeled, for the first time, multiple aspects of urban life to describe and predict crime. We have done so by operationalizing the Jane Jacobs theory to describe the urban fabric, the social disorganization and the routine activity variables from criminology. We can now discuss some implications of our work.

Descriptive maps. Maps are invaluable tools in criminology to understand where the problems are, and to evaluate initiatives for crime prevention. Thus, our framework allows policy makers to visually analyze crime rates (Figure 2), crime counts (Figure 1) and, most importantly, spatial auto-correlations (Figure 3) and heterogeneous effects (Figure 4).

Factors for crime. With our descriptive model we have shown that it is possible to have a static description of what happens in the city, and how the multitude of complex features of city life come together. Therefore, now more than ever, it is important to control for the many relations that come together, especially as a consequence of urban fabric and mobility. Urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited by policy makers to reduce crime.

Generalization. Our predictive model that the combination of structural and socioeconomic information, and mobility, provides better predictions than each on its own. However, it is striking that the urban fabric has such an important role in the prediction. Thus, we think it is critical to consider it in new models and analysis in different cities in the world.

This work is not without any limitation. First, we don't analyze crime over time. Then, cities are not to be considered island unto themselves, as they are embedded in a country-wide complex system of social interactions. Routine of residents exposes them to different cities, conditions and possibilities on a daily basis. Thus, we think that in the next future it is important to consider also these factors.

Our findings are apt to crime control and prevention action plans for Colombian cities. This study provides valuable insights for local governments so that they can base urban management decisions on empirical evidence on the deterrents of crime.

7. References

[1] Martin A Andresen. Crime measures and the spatial analysis of criminal activity. *British Journal of criminology*, 46(2):258–285, 2006.

[2] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the street: *Classifying* crime hotspots using aggregated anonymized data on people dynamics. Big Data, 3(3):148–158, 2015.

[3] Paul J. Brantingham. *Environmental Criminology*. Waveland Press, 1991.

[4] Ernest Watson Burgess. *The growth of the city: an introduction to a research project*. Ardent Media, 1967.

[5] Andrew Cliff and Keith Ord. Spatial autocorrelation. Technical report, 1973.

[6] Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.

[7] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*, pages 413–423. International World Wide Web Conferences Steering Committee, 2016.

[8] Neva R Deardorff and Clifford R Shaw. Delinquency areas: A study of the geographic distribution of school truants, juvenile delinquents, and adult offenders in chicago, 1930.

[9] Corina Graif, Andrew S Gladfelter, and Stephen A Matthews. Urban poverty and neighborhood effects on crime: Incorporating spatial and network perspectives. *Sociology Compass*, 8(9):1140–1155, 2014.

[10] Corina Graif, Alina Lungeanu, and Alyssa M Yetter. Neighborhood isolation in chicago: Violent crime effects on structural isolation and homophily in inter-neighborhood commuting networks. *Social Networks*, 2017.

[11] Corina Graif and Robert J Sampson. Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. *Homicide studies*, 2009.

[12] Daniel A Griffith. *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer Science & Business Media, 2013.

[13] Jane Jacobs. *The death and life of great American cities*. Vintage, 1961.

[14] Robert B O'hara and D Johan Kotze. Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122, 2010.

[15] Robert E Park. The city: Suggestions for the investigation of human behavior in the city environment. *The American Journal of Sociology*, 20(5):577–612, 1915.

[16] Gabriel Rosser, Toby Davies, Kate J Bowers, Shane D Johnson, and Tao Cheng. Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, pages 1–26, 2016.

[17] Hyungun Sung and Sugie Lee. Residential built environment and walking activity: Empirical evidence of jane jacobs' urban vitality. *Transportation Research Part D: Transport and Environment*, 41:318–329, 2015.

[18] Michael Tiefelsdorf and Daniel A Griffith. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A*, 39(5):1193–1221, 2007.

[19] Martin Traunmueller, Giovanni Quattrone, and Licia Capra. Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics*, pages 396–411. Springer, 2014.

[20] Hongjian Wang, Zhenhui Li, Daniel Kifer, and Corina Graif. Crime rate inference with big data. In KDD, 2016.

Papiers de Recherche de l'AFD

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ilss'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l'AFD n'en exclut aucune autre.

L'Agence Française de Développement (AFD), institution financière publique qui met en oeuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l'AFD a consacré 8,1 milliards d'euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy

defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

AFD, 5 rue Roland Barthes
75598 Paris Cedex 12, France
ResearchPapers@afd.fr
ISSN 2492 - 2846

BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018