Addressing Under-Reporting to Enhance Fairness and Accuracy in Mobility-based Crime Prediction

Jiahui Wu
College of Information Studies
University of Maryland
College Park, Maryland, USA
jeffwu@umd.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
enrique.friasmartinez@telefonica.com

Vanessa Frias-Martinez
College of Information Studies
UMIACS
University of Maryland
College Park, Maryland, USA
vfrias@umd.edu

ABSTRACT

Traditionally, historical crimes and socioeconomic data have been used to understand crime in cities and to build crime prediction models. Nevertheless, the increasing availability of mobility data from cell phones to location-based services, has introduced a new family of mobility-based crime prediction models that exploit the relation between mobility patterns and reported crime incidents. One of the major concerns of using reported crime data is under-reporting, which will bias the crime predictions. In this paper, we propose a novel Bayesian Hierarchical model that utilizes domain knowledge about biases in reported crime data to characterize and enhance fairness and accuracy in mobility-based crime predictions. An in-depth feature analysis reveals the influence that various factors might play in crime under-reporting and algorithmic fairness for mobility-based crime predictors.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies → Learning in probabilistic graphical models**; **Supervised learning**.

KEYWORDS

Crime prediction, Under-reporting, Algorithmic fairness

ACM Reference Format:
Jiahui Wu, Enrique Frias-Martinez, and Vanessa Frias-Martinez. 2020. Addressing Under-Reporting to Enhance Fairness and Accuracy in Mobility-based Crime Prediction. In *28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20), November 3–6, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3397536.3422205

1 INTRODUCTION

Historical crime data is of great importance to understand the severity of crimes in society. Countless reports, academic papers, books and news articles rely on reported crime data [27, 38]. This data can be used to, for example, evaluate the effects of programs and policies designed to prevent crime in a city [16]. Crime prediction, on the other hand, is an important topic of research that uses reported crime data to predict future occurrences. For example, historical crime data has been used to predict hotspots so as to assist patrol route planning [28]. Traditionally, historical crimes and socioeconomic data have been used as input to build crime prediction models at various geographical levels *e.g.,* grids, cities, municipalities [8, 27]. Nevertheless, due to the increasing availability of mobility data such as geolocated social media and mobile phone data, a large number of studies have explored the predictive relationship between mobility patterns and reported crimes [3, 39].

There are various theories about the relationship between mobility and crime in urban environments. For example, the *Opportunity makes the Thief* theory claims that the opportunity is the cause of crime [10] *i.e.,* the higher the presence of suitable targets such as people and property, the more crimes could happen; and empirical work has confirmed that theory, showing that there is a super-linear relation between the daily floating population (number of people that has been in a neighborhood) and incidence of property crimes [7]. Mobility patterns not only trace the movement of people, but can also characterize the dynamic spatial structure of the urban environment by detecting urban dense areas, *a.k.a.* hotspots [23]. Urban spatial structure is a critical and well understood concept in environmental criminology and urban quantitative geography, and it has been shown to be correlated to crime incidents [5, 37].

One of the major concerns of using reported crime data in crime prediction is data bias, especially when computational models - built upon such data - could influence future resource allocation *e.g.,* planning police patrol routes [28]. Data bias in this context can be framed under algorithmic fairness whereby crime predictive models can behave differently for disadvantaged groups such as low-income or minorities due to over- or under-representation in the historical crime data [21]. In fact, not all crimes are reported. Sometimes the public does not report crimes that are considered minor [32]; and low-income has been related to higher under-reporting for certain types of crimes [35]. In addition, not all reported crimes end up recorded in the official crime statistics, a decision mostly made by the police [26]. Police may decide not record a report as a crime because of insufficient evidence and/or individual biases [26]. Therefore, the reported crime data statistics that are used in research will naturally be biased, reflecting partial crime incidents mediated by community engagement, police resources and potential police biases towards disadvantaged populations. Although a few papers have looked into the identification of biases in predictive policing tools that exclusively use historical crimes [24], there is no

SIGSPATIAL '20, November 3–6, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8019-5/20/11…$15.00
https://doi.org/10.1145/3397536.3422205

work in the analysis of biases for mobility-based crime prediction models, and more generally, no work in mitigation strategies to enhance fairness without sacrificing accuracy in crime prediction.

In this paper, we propose a Bayesian hierarchical model to identify and mitigate under-reporting issues that could lead to biases and lack of fairness in mobility-based crime incident predictions [29, 31]. Specifically, the predictive model uses mobility-based features to infer the number of *true* crimes, *i.e.,* the actual number of crimes that will occur regardless of whether they will be reported; and use domain knowledge of determinants for under-reporting (*e.g.,* poverty, unemployment rate) to model the reporting rate, *i.e.,* the ratio of the number of reported crimes to true crimes. By distinguishing the number of true crimes and reporting rate from the reported crime data, we will show that our model manages to improve both the accuracy and fairness of the crime prediction. In summary, the main contributions of this paper are:

1) A novel mobility-based crime prediction model that utilizes domain knowledge about biases in reported crime data to improve fairness and accuracy in mobility-based crime predictions. We frame biases within the under-reporting phenomenon whereby the statistics used by the algorithm fail to cover all crimes that actually happened.

2) An in depth analysis of the influence that various features might play in crime under-reporting and algorithmic fairness for mobility-based crime predictors.

2 RELATED WORK

2.1 Crime Prediction with Mobility Patterns

Historical crime data and socioeconomic data are often used in crime prediction models [8]. For example, historical crime hotspots can be used to assess the risk of future crimes [8, 28]. Mohler uses a marked point process to model the dependency between gun crimes and homicides for homicide prediction in cities [28]. Neural networks have also been utilized to model the spatio-temporal patterns in historical crimes for future crime prediction [38]. In addition to historical crimes, census data [20] and points of interest (POI) [39] have also been used to enhance crime prediction. The proliferation of human mobility data, such as mobile phone data, geo-located social media, taxi pick-up/drop-off and check-ins, has allowed for the use of mobility features to predict crime incidents. One of the most common mobility feature used in crime prediction is *footfall* defined as the number of individuals present in a given area at a given time span. Various studies use footfall as a feature to predict future crimes [3, 20]. Bogomolov *et al.* estimate footfall and population diversity such as gender and age from mobile phone data and predict whether a regular grid cell will have a high or low level of crimes in the following month [3]; while Kadar and Pletikosa extracted footfall from check-ins, subway and taxi data, along with other census and POI features, to predict the number of crimes for a given census tract using tree-based machine learning models [20]. *Footfall* can be used to identify urban dense areas, *a.k.a.* hotspots. Hotspots are important in the fields of environmental criminology and urban quantitative geography because they can be used to characterize the dynamic spatial structure of the urban environment, which has been shown to play a role in crime incidence [5, 37]. In addition to the volume of hotspots, urban spatial structure can be quantified via urban sprawl [36] and urban compactness [1]. In this paper, we will focus on mobility-based crime prediction models that exploit the predictive power of the dynamic hotspots and urban spatial structures in cities by analyzing the relationship between hotspots and crime incidents.

2.2 Under-reporting in Crimes Statistics

Concerns about under-reporting in crime data are highly related to the production of the reports themselves. Although crime reporting systems around the world vary a lot, in a simplified way, we can identify two main phases: a crime first needs to be reported to the police by an individual, and it then needs to be recorded as a crime entry into the police database. When crimes are reported, around 80% of them are reported by victims or witnesses, while the police on scene reports about 6% and the rest are reported by offenders, alarm systems or officials other than police, among others [13, 18]. However, there are various reasons why the public might choose not to report a crime. The crime being "too trivial/no loss" used to be the most important reason, but recently "Police could do nothing" has come on top [32]. After an incident is reported, the police decides whether or not to record the incident as a crime event in the database. Various factors can influence the police' decision such as insufficient evidence and/or individual biases [26] As a result, under-reporting in crime is heavily impacted by social disparities. For example, in Kensington, middle-class crime complaints are more likely to be reported and accepted by the police (*i.e.,* high reporting rate and high recorded rate), while the reports from white working-class tend to be rejected (low recorded rate) and racially-mixed communities are less willing to report (low reporting rate) [13].

Therefore, it is critical to address the existing bias in reported crime data so that crime predictions are fair across social groups. Although the reporting and recording of crime incidents are two different phases, in this study, we make no distinction between them as it is almost impossible to obtain such information from local police force. Instead, we simplify and quantify the under-reporting issue of crimes as the reporting rate, which is the ratio of the number of reported crimes in the police database to the number of (unobserved) true crimes that have occurred. This simplification is common in the literature [18].

2.3 Algorithmic Fairness

There exists a plethora of computational algorithms making decisions with high societal impact such as loan requests, crime prediction or criminal sentencing. As a result, algorithmic fairness or the design of algorithms that treat social groups similarly, becomes a critical component of any predictive approach. Algorithmic fairness, especially the most commonly used notion of group or statistical fairness, is based on the notion of protected or sensitive attributes, such as gender and race (minority and non-minority). A protected attribute usually represents a population sub-group that has historically suffered from discrimination and therefore some form of (approximate) parity or non-discrimination regulation in the predictive algorithm is desired for these protected groups [9]. Fairness is a complex concept and there are different and sometimes conflicting definitions and thus a variety of fairness metrics [34]. Although the definitions of fairness vary, it has been empirically shown that there is usually a trade-off between the accuracy and fairness of prediction, *i.e.,* improvement in fairness is generally at the expense of the algorithmic accuracy [2]. In this paper, we propose a novel algorithm to correct under-reporting in crime data while controlling for fairness across protected attributes and accuracy. By properly incorporating domain knowledge about potential under-reporting - which is a source of data bias - we will show that we can improve both fairness and crime prediction accuracy for mobility-based crime prediction algorithms.

3 METHOD

Mobility-based crime prediction can be framed as a regression problem: given a region of interest i, *e.g.,* a city, a set of mobility-based features $\boldsymbol{u_i}$ characterizing the dynamic spatial structure of i extracted from past mobility data and a set of determinants $\boldsymbol{s_i}$ that characterize under-reporting in i, predict the number of future crimes z_i in that region, *i.e.,*

$$z_i = F(\boldsymbol{u_i}, \boldsymbol{s_i}), \tag{1}$$

where F is the predictor to be trained. F can represent under-reporting-unaware models, *i.e.,* models that do not address under-reporting and use crime data as is, such as generic machine learning models; we hypothesize these models will make biased crime prediction due to the inherent bias in the reported crime data. F can also represent under-reporting-aware models, such as our proposed Bayesian model that explicitly models the under-reporting issue so as to mitigate bias.

In this section, we will introduce three major components of our proposed method: 1) The construction of mobility-based features $\boldsymbol{u_i}$ based on Call Detail Records (CDR) data; 2) The Bayesian hierarchical model for mobility-based crime prediction that addresses the under-reporting issue using a set of under-reporting determinants $\boldsymbol{s_i}$; 3) The process of fairness and accuracy evaluation for crime prediction. Table A.1 in the Appendix presents a notation summary.

3.1 Mobility-based Hotspots Features

As stated in the Related Work, we will focus on mobility-based crime prediction models that exploit the predictive power of the dynamic hotspots and urban spatial structures in cities by analyzing the relationship between hotspots and crime incidents.

3.1.1 From Mobility Data to Urban Hotspots. The mobility data used in this paper are Call Detail Records (CDR). CDR are a commonly used mobile phone data collected by telecommunication networks for billing purposes. CDR provide - among other features - spatio-temporal data about individual mobility behaviors. CDR locations are represented as the (latitude, longitude) pairs of the cellular towers that mobile phones are using when making phone calls or sending texts. The spatial coverage of cellular towers is often approximated via Voronoi tessellation. For each Voronoi polygon we can compute the hourly footfall, defined as the average number of hourly unique users present at a given polygon (see Figure A.1(a) in Appendix).

Due the irregularity of Voronoi tessellation, we interpolate footfall from Voronoi polygons to regular grids, with the assumption that footfall within a Voronoi polygon is uniformly distributed over space. That is, the footfall for a grid within a Voronoi polygon is proportional to the overlap between grid and Voronoi polygon (see Figure A.1(b)). In order to detect urban hotspots, we follow a similar approach to [23]: 1) for each hour of the day, we apply the Loubar method to the hourly footfall of each grid so as to detect the upper bound of the number of hourly hotspots (Figure A.1(c)); 2) the grids that are detected as hotspots over the 24 hours of the day are identified as permanent hotspots (Figure A.1(d) in Appendix). The permanent hotspots represent the most important centers of dense activity in the urban environment and are the ones that we will use to predict crime incidents.

3.1.2 Urban Hotspot Features. In this paper, we will explore three types of urban hotspot features that have been traditionally used in related literature for hotspot analysis and urban spatial structure: scale, sprawl and compactness. The definition and calculation of features are explained as follows:

(1) Hotspot Scale quantified in terms of number of grids in a city that are detected as hotspots (*NHS*) and the total geographical area covered by the hotspots detected (*AHS*).

(2) Urban sprawl characterizes a type of metropolitan decentralization or sub-urbanization where a large percentage of a city's residential and/or business activity takes place outside of its central location [36]. We use the following indices to quantify the degree of urban sprawl:

- Compacity coefficient (*COMP*) [23] measures the sprawl of the detected hotspots over a city, with smaller *COMP* values associated to less dispersed hotspots with respect to the size of the city. Let A be the geographic area of the city of interest, hs be the set of hotspots, $|hs|$ be the number of hotspots and $d_{j,k}$ the distance between the centroids of hotspot j and k.

$$\text{COMP} = \frac{\text{D}_{hs}}{\sqrt{A}},\ \text{D}_{hs} = \frac{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} d_{j,k}}{|hs|\,(|hs|-1)/2} \tag{2}$$

- Mass Compacity coefficient (*MCOMP*) is a modified compacity coefficient that weights the distance between hotspots by the population of each grid, and measures the average distance between individuals located within the detected hotspots. The smaller MCOMP is, the less dispersed the hotspots are with respect to the size and population of the city. Let p_j be the population in grid j.

$$\text{MCOMP} = \frac{\text{MD}_{hs}}{\sqrt{A}},\ \text{MD}_{hs} = \frac{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} d_{j,k}\, p_j\, p_k}{\sum_{j=1}^{|hs|}\sum_{k=j+1}^{|hs|} p_j\, p_k} \tag{3}$$

(3) Urban compactness The major difference between urban compactness and urban sprawl indices is that sprawl is always measured with respect to the size of a city *e.g.,* the indices are normalized by the square root of the geographical area, while compactness is based on the assumption that the most compact form of a shape is a circle [1]. Therefore, compactness indices measure compactness in terms of geometrical properties, and are thus normalized by the reference circle *e.g.,* an equal-area or equal-perimeter circle. Urban compactness indices range from 0 to 1, with 1 representing the exact continuous cirle. We consider the following four indices that are commonly used in hotspot measurement literature:

- Cohesion (*COHE*) [1] is the ratio of the average distance-squared among all points in the reference circle and the average distance-squared among all points in the hotspot areas. Large cohesion means people in hotspot areas are very close to each other. Let r be the points of hs in the rasterized format, $|r|$ be the number of points and $dr_{j,k}$ be the distance between the j- and k-th point.

$$\text{COHE} = \frac{\text{AHS}/\pi}{\frac{2}{|r|(|r|-1)} \sum_{j=1}^{|r|} \sum_{k=j+1}^{|r|} dr_{j,k}^2} \tag{4}$$

- Proximity (*PROX*) [1] is the ratio of the average distance from all points in the reference circle to its centre and the average distance to the geometry center of the hotspot areas. The proximity index focuses on the distance between points from the geometry center instead of the point-wise distance in the cohesion index. Let g be the center of gravity of hs and dg_i be the distance between the j-th point and the center g.

$$\text{PROX} = \frac{\frac{2}{3}\sqrt{\text{AHS}/\pi}}{\frac{1}{|p|} \sum_{j=1}^{|p|} dg_j} \tag{5}$$

- Normalized moment of inertia (*NMI*) [22] is based on the dispersion of points from the center of its shape. It involves the calculation of the second moment of an area about a point, also known as the moment of inertia (MI). The MI is then normalized by the MI of the reference circle, hence normalized moment of inertia.
- Normalized mass moment of inertia (*NMMI*) [22] takes into account the mass distribution of a shape. The previous three compactness indices consider only the geometric shape *i.e.,* each point in the shape is equally important in the compactness. Nevertheless, in our case, each hotspot might have a different estimated population or mass, and they can still be compact - even though their geometry shape is not - by having the majority of the population concentrate around the mass center. The reference circle in *NMMI* is no longer an equal-area circle, but a circle with equal-effective-area. The mathematical derivation for the calculation of *NMI* and *NMMI* can be found in [22].

3.2 Bayesian Model for Under-Reported Crimes (BURC)

As explained in previous sections, the problem of under-reporting in crime data is an important source of potential bias in mobility-based crime prediction algorithms that might affect protected groups. In this paper, we develop a Bayesian hierarchical model to mitigate under-reported crime incidents by inferring two variables (1) the unobserved "true" crime incidents *i.e.*, all the crimes that have occurred regardless of whether they have been reported and recorded; and (2) the reporting rate *i.e.*, the ratio of true crimes being reported in the crime data. The core of this Bayesian model is that we assume the crime incidents are generated following a Poisson distribution given the urban spatial structure features and that the reporting rate is dependent on the determinants of under-reporting through a logistic link function.

For a given city i, let y_i be the volume of true crime incidents (hidden variable), λ_i be the average incident occurring rate for true crime incidents for the Poisson distribution, z_i be the volume of reported crime incidents, and π_i be the reporting rate (hidden variable). We model the generative process of crime incidents as follows:

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{6}$$

$$z_i|\lambda_i, \pi_i \sim \text{Poisson}(\pi_i \lambda_i) \tag{7}$$

$$log(\lambda_i) = \alpha_0 + \sum_{k=1}^{K} \alpha_k u_i^{(k)} \tag{8}$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j s_i^{(j)} \tag{9}$$

The volume of true crimes y_i follows the Poisson distribution given the average occurring rate λ_i; and the volume of reported crimes z_i also follows a Poisson distribution but the occurring rate is $\pi_i\lambda_i$, discounted by the reporting rate π_i. The λ_i is modeled by the logarithmic link function to ensure $\lambda_i > 0$ and the π_i is modeled by the logistic link function to ensure $\pi_i \in (0, 1)$. $\boldsymbol{u}_i = (u_i^{(1)}, ..., u_i^{(K)})^T$ is the feature vector for city i and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_K)^T$ are the coefficients to model the true crimes occurring rate. $\boldsymbol{s}_i = (s_i^{(1)}, ..., s_i^{(J)})^T$ is the feature vector for city i and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)^T$ are the coefficients to model the reporting rate of reported crimes. In our study, the feature vectors $\boldsymbol{u}_i$ for the true crimes occurring rate λ_i are the urban hotspot features that characterize the dynamic spatial structure of a city and which has been related to crime incidents [37]. The feature vectors $\boldsymbol{s}_i$ are determined by domain knowledge about the determinants for the reporting rate of the types of crimes of interest. For example, studies have shown that poverty rate [35] and unemployment rate [25] can decrease the likelihood of property crime incidents, such as burglaries, being reported. Therefore the feature vector $\boldsymbol{s}_i$ for the under-reporting process for property crimes would contain poverty rate (PR) and unemployment rate (UR) for each municipality. Similarly, gender, age and marital status of the victims [18], as well as the percentage of female-headed households with children, poverty rate (PR) and foreign born population rate (FR) of census tracts [33] have been shown to influence violent crime reporting behavior. Therefore, these factors would be the $\boldsymbol{s}_i$ determinants for violent crime prediction. Section 4.1 explains the specific features we use for our model in detail.

By treating the volume of reported crime incidents as observed variables and the volume of true crime incidents and reporting rate as hidden variables, this model manages to separate the bias in the crime reporting process from the volume of true crimes. In Section 4.2 we will show that this model can more accurately infer crime volumes while making more fair predictions.

3.3 Fairness and Accuracy Evaluation

As mentioned in the related work, fairness evaluation is often based on the notion of protected attributes such as gender, race or income levels. Although there are various definitions of fairness, its main objective is to achieve some form of (approximate) parity across the various groups defined by the protected attribute *e.g., female vs. male, low-income vs. high-income.* In this paper we will consider two protected attributes that have been observed to receive unfair

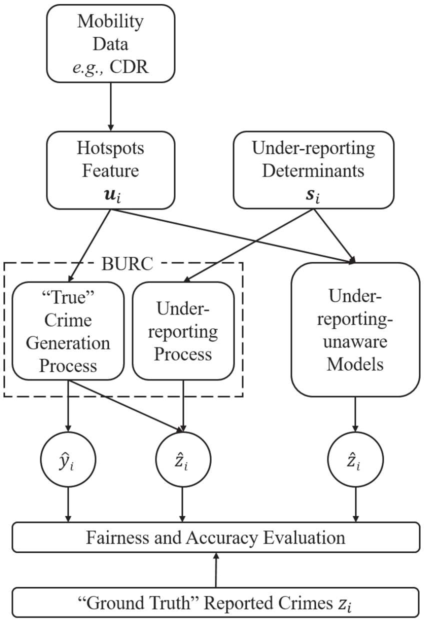

Mobility Data *e.g.,* CDR
Hotspots Feature $\boldsymbol{u}_i$
Under-reporting Determinants $\boldsymbol{s}_i$
BURC
"True" Crime Generation Process
Under-reporting Process
Under-reporting-unaware Models
$\hat{y}_i$
$\hat{z}_i$
$\hat{z}_i$
Fairness and Accuracy Evaluation
"Ground Truth" Reported Crimes z_i

Figure 1: The framework of our study. BURC is our proposed Bayesian hierarchical model and its implementation details are in Fig. A.4. $\hat{z}_i$ is the predicted number of reported crimes by different models and $\hat{y}_i$ is the predicted number of "true" crimes by BURC.

treatment and suffer from discrimination in the criminal justice system: income and race [17].

In our crime prediction problem setting, we are evaluating fairness for a regression problem. A common choice of fairness metric for regression problems given a binary protected attribute is the mean difference *i.e.,* the difference between average prediction values in the positive group *e.g.* female, and the average prediction values in the negative group *e.g.,* male [6]. The mean difference is a real number with a value of zero signifying no attribute effect or dependency. The larger the absolute value of mean difference is, the less fair the predictions are for a given protected attribute. Given that we will work with non-binary protected attributes (income and race) *i.e.* attributes defining more than two population groups, we generalize the definition of mean difference for binary groups into multiple groups. We compute the mean difference for each group as a "1 vs all" setting *i.e.,* the MD_i for group i is computed as the difference between average prediction for group i and average prediction for other groups. In addition, we will compute the group error *i.e.,* the RMSE between predicted and ground truth reported crimes within each group, to show the fairness in terms of performance difference across different protected groups.

To assess the impact of addressing under-reporting in crime data, we will use the proposed BURC model to infer the volumes of true crimes; and we will evaluate the fairness and the accuracy of the BURC predictions via mean difference across protected groups (see Figure 1). Finally, these results will be compared against a set of baseline classifiers that use the reported crime data without any under-reporting treatment. Given that the volumes of true crimes have a different scale than the reported crimes, and given that the mean difference we use to measure fairness is scale dependent, we will use the mean difference normalized by the average of reported crimes and true crimes to allow comparison between models. Similarly, since different protected groups will have different scales for average prediction, we will also normalize RMSE by the group average of the prediction to show the relative group error.

4 EXPERIMENTS

To assess whether addressing under-reporting in crime data can improve the fairness and accuracy of crime prediction models, we focus on crime and mobility data from $1,379$ municipalities in Mexico. 90% of the municipalities have population less than $80,000$ and geographic area less than $2,000$ km^2 while the largest population is $1,815,786$ and the largest area is $53,256$ km^2. In our study, we consider two types of crime: property crimes and violent crimes across municipalities in Mexico.

4.1 Experiment Setting

4.1.1 Data. We have used four data sources:

1. Mobility data is extracted from aggregated and anonymized Call Detail Records (CDR) from October 2009 to June 2010 across all $1,379$ municipalities in Mexico (see Fig. A.2 in Appendix). No individual data has been used, only aggregated statistics at the cell tower level. As described in Section 3.1, CDR data is used to extract footfall and hotspot features.

2. Reported crime statistics are obtained from Mexico's *Secretary General of National Public Security* (SESNSP) [30]. We have retrieved property and violent crime data from 2011 for the $1,379$ municipalities under study. Property crimes in this study mainly include thefts, thefts from vehicles and burglaries, while violent crimes include robbery, sexual offense, homicide, battery, assault and kidnapping. These annual volumes of reported property crime or violent crime are used as the observed variables $z_i, i = 1, ..., 1379$ in the BURC model. The range of number of reported property (violent) crimes in these municipalities is [0, 17655] ([0, 28329]), the average is 265 (522) and the standard deviation is 1091 (1868). The volumes of reported property (violent) crime for 90% of the municipalities are less than 450 (900). Therefore there is a large variation in the number of crimes across municipalities.

3. Determinants of under-reporting in BURC (as described in Section 3.2) include poverty rate (PR), unemployment rate (UR), adult rate (AR), the percentage of people who are never married (never married rate, NMR), male to female ratio (M/F), male-headed to female-headed household ratio (M/FHH) and the percentage of population born in other municipalities (foreign-born rate, FR). Poverty rates are obtained from Mexico's *National Council for the Evaluation of Social Development Policy* (CONEVAL) [12] and the other indicators are obtained from the 2010 Population Census [19]. PR and UR are used in the BURC model as the domain knowledge features $\boldsymbol{s}_i$ to characterize the reporting rate of property crimes *i.e.,* factors that affect the percentage of crime incidents being reported; while AR, NMR, M/F, M/FHH, FR and PR are used as $\boldsymbol{s}_i$ in the violent crime model.

4. Protected attributes for fairness evaluation include average income and statistics of indigenous population from CONEVAL [12]. The average income is a real-value attribute. We have divided

income into quartiles of average income, and assigned an income group label to each municipality: from *IcQ1* (lowest average income) to *IcQ4* (highest average income). On the other hand, the census identifies 4 types of municipalities determined by the presence of indigenous population (IP): *IP1* characterizes municipalities without indigenous population (there are 5 such municipalities in our dataset); *IP2* are municipalities with less than 40% of the population being indigenous and the indigenous population being less than 5000 (955 municipalities); *IP3* characterizes municipalities with less than 40% of the population being indigenous and the indigenous population being 5000 or more (213 municipalities); and *IP4* that represents municipalities with more than 40% of the population being indigenous (206 municipalities). These four types of indigenous municipalities, from *IP1* to *IP4*, characterize the increasing presence of indigenous population in a municipality.

4.1.2 BURC settings. The BURC model is implemented using NIMBLE in R [14] and the posterior distribution is inferred by Markov chain Monte Carlo (MCMC) sampling. The basis of MCMC sampling is that when the Markov chain converges, the samples generated by MCMC sampling are the joint posterior distribution of the Bayesian model. The burn-in period is $80,000$ iterations where samplings from MCMC are discarded before the Markov chains converge to the posterior distribution. After the burn-in period, another $80,000$ iterations are used to generate posterior samples with thinning intervals of 40. Four independent chains are used to sample and examine the convergence of the model.

The prior distribution for α and β in the BURC model is computed as follows: α_0 is defined by a normal distribution N(4,2) to be conservative when making large crime volume predictions *i.e.,* the probability of $\alpha_0 > 8$ (number of true crimes > 2981 given all $\boldsymbol{u}$ features equal to 1) is 2.5%. The prior distribution for β_0 is defined as N(-2, 0.5), because the national survey of victimization in Mexico (ENVIPE) suggests that the under-reporting rate of all crimes is around 88% in 2010 [15] and the inverse logit of -2 is 0.12. The prior distributions for other coefficients, α_k and β_j are defined with N(0,100) which are relatively non-informative priors.

After assessing the convergence of the MCMC sampler, we use the mean point estimate of the parameters to make predictions for each municipality i. Specifically, we compute: 1) the predicted volume of true crimes, which is the expected value of the Poisson distribution for the generation of true crimes, and which is estimated as $\hat{y}_i = \hat{\lambda}_i$; 2) the predicted reporting rate, which is estimated as $\hat{\pi}_i$; and 3) the predicted volume of reported crimes, which is the expected value of the Poisson distribution for the generation of reported crimes, and which is estimated as $\hat{z}_i = \hat{\pi}_i \hat{\lambda}_i$. The process of implementing BURC using NIMBLE is summarized in Fig. A.4.

4.1.3 Evaluation. The evaluation focuses on understanding if addressing the under-reporting issue in mobility-based crime predictors improves the fairness and accuracy of the predictive models. To achieve that, we will analyze fairness and accuracy of the proposed BURC model against a battery of three baselines, which are commonly used machine learning models for regression: Random Forest (RF), Bagging (BAG) and XGBoost (XGB). All baselines use random search hyperparameter tuning with a validation set from the training data to select the best hyperparameters. The feature vectors used to train the baselines are a concatenation of hotspot features ($\boldsymbol{u}_i$) and domain knowledge features *e.g.,* unemployment, poverty rates or gender ($\boldsymbol{s}_i$) so that baselines have access to the same information as our proposed BURC model.

	Metric	RF	BAG	XGB	BURC
Property Crimes	RMSE	763.1	803.8	810.8	**601.2**
	MAE	198.4	210.5	211.5	**180.4**
	Correlation	0.73	0.68	0.67	**0.82**
Violent Crimes	RMSE	1301.4	1294.5	1306.0	**1160.0**
	MAE	404.7	406.1	404.9	**346.1**
	Correlation	0.73	0.74	0.73	**0.81**

Table 1: Average cross validation performance for baselines and BURC model. BURC model has much lower error and higher correlations than the baselines.

In our experiment, we use 5-fold cross validation to split the data into training and testing sets: the $1,379$ municipalities are randomly split into 5 folds and in each experiment, 1 fold is used as testing set for evaluation and the 4 remaining folds are used for training models. Model performance and model fairness are reported as averages across all 5 runs. To evaluate the performance of the mobility-based crime prediction models, we use the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the correlation between predicted and *ground truth* volumes of reported crimes. Compared to MAE, RMSE penalizes models that have large errors. Fairness, on the other hand, is measured by the mean difference and group errors as described in Section 3.3. Next, we present our main results.

4.2 Results

4.2.1 Convergence of BURC. We assess the convergence of the BURC model by autocorrelation tests and Gelman–Rubin convergence diagnostic [4]. Samples from MCMC samplers are not independent *i.e.*, the current sample being drawn is dependent on the previous sample, and thus there is autocorrelation among the posterior samples. Autocorrelation tests compute the autocorrelation with lag k, which is defined to be the correlation between the samples k steps apart. If the MCMC sampler has converged and reached the stationary distribution, the autocorrelation value should be small as k increases and 0 means samples are independent with samples after k iterations [11]. In the reported property crime experiment, the autocorrelation drops as k increases and eventually converges around 0 after 50 iterations (see Figure A.3 in the Appendix). Similar behavior was observed for the violent crime model. Gelman–Rubin convergence diagnostic requires multiple Markov chains with different starting points and assesses the convergence by computing the potential scale reduction factor (PSRF) based on between-chain and within-chain variance. If the MCMC sampler converges, the PSRF is close to 1 [4]. The PSRFs of all coefficients in BURC in both types of crime experiment are less than 1.01 suggesting our model converges well in both experiment and the samples from this sampler can be used to estimate the posteriors.

4.2.2 Performance of Reported Crime Prediction. In this section, we compare the BURC model performance against the baselines.

	IcQ1	IcQ2	IcQ3	IcQ4	AbsSum
$\boldsymbol{z}$	-1.30	-1.16	-0.89	3.45	6.80
$\hat{\boldsymbol{z}}_{RF}$	-1.20	-0.99	-0.60	2.84	5.63
$\hat{\boldsymbol{z}}_{BAG}$	-1.16	-0.98	-0.56	2.76	5.46
$\hat{\boldsymbol{z}}_{XGB}$	-1.21	-1.05	-0.66	2.98	5.90
$\hat{\boldsymbol{z}}_{BURC}$	-1.28	-1.14	-0.71	3.22	6.34
$\hat{\boldsymbol{y}}_{BURC}$	**-0.87**	**-0.62**	**-0.32**	**1.84**	**3.66**

Table 2: MD for protected attribute income group in property crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all and each of the income groups.

	IP1	IP2	IP3	IP4	AbsSum
$\boldsymbol{z}$	-0.60	-2.04	4.40	-1.07	8.10
$\hat{\boldsymbol{z}}_{RF}$	-0.56	-1.14	2.80	-0.94	5.44
$\hat{\boldsymbol{z}}_{BAG}$	-0.53	**-0.98**	2.51	-0.92	4.94
$\hat{\boldsymbol{z}}_{XGB}$	-0.54	-1.11	2.80	-0.99	5.44
$\hat{\boldsymbol{z}}_{BURC}$	**-0.47**	-1.36	3.27	-1.08	6.17
$\hat{\boldsymbol{y}}_{BURC}$	-0.52	-1.10	**2.37**	**-0.59**	**4.58**

Table 3: MD for protected attribute indigenous group in property crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{\boldsymbol{y}}_{BURC}$) are fairer than baselines across all groups and are in favor of IP3 and IP4 which have more presence of indigenous population.

Table 1 summarizes the experimental results. For property crime prediction, the best correlation between actual and predicted crime incidents for the baseline models is 0.73 (Random Forest) while that value increases to 0.82 for our proposed BURC model. As for violent crime prediction, the performance of three baselines is similar and our BURC model still has the highest correlation. This result shows the effectiveness of using urban hotspots features to predict future crime incidents; but more importantly, it also demonstrates that by explicitly modeling under-reporting in crime data, our BURC model can perform better than common machine learning models. In addition to higher correlation, BURC reduces the RMSE and MAE by 21.2% and 9% for property crimes prediction and by 10.4% and 14.4% for violent crimes.

4.2.3 Fairness: Mean Difference. In this section, we evaluate the fairness of the BURC and baseline models for two protected attributes, income and presence of indigenous groups, using the mean difference (MD) described in Section 3.3. We mostly discuss the MD results for property crime prediction, since results for violent crime follow a similar trend. Results for the latter can be found in the Appendix (Tables A.2 and A.3). Tables 2 and 3 summarize the normalized MD for both protected attributes in the property crime prediction. The first 4 columns in each Table represent the mean difference MD_i between the average volume of crime incidents for group i and all other groups *e.g.,* column *IcQ1* represents the mean difference between the average volume of crimes for group *IcQ1* and the average of crime volumes in municipalities that are not in group *IcQ1.* The last column *AbsSum* is the sum of the absolute mean difference from all four columns, and we use it to evaluate the overall fairness across different groups. On the other hand, the row $\boldsymbol{z}$ represents the mean difference of the actual reported crimes in the testing set (ground truth) and evaluates the fairness in the data itself; the row $\hat{\boldsymbol{z}}_{model}$ represents the MD of the predicted reported crimes for the three baseline models and for our proposed BURC model without under-reporting correction (model in formula 7, Section 3.2); and $\hat{\boldsymbol{y}}_{BURC}$ represents the MD of the predicted volumes of true crimes *i.e.,* volumes of crimes post under-reporting correction as computed by BURC (model in formula 6, Section 3.2).

Tables 2 and 3 show that the BURC model addressing under-reporting ($\hat{\boldsymbol{y}}_{BURC}$) has fairer crime predictions across all groups *i.e.,* AbsSum MD is the lowest for both income and presence of indigenous groups. These results highlight that by correcting the under-reporting, the BURC model does a better job at providing fairer predictions across groups. Looking in depth into each protected attribute, we see that BURC provides the lowest mean difference (highest fairness) across all income groups: from low (IcQ1) to high (IcQ4) average income. However, although BURC has the lowest sum of absolute MD across all four types of indigenous population presence, we observe that the BURC model provides fairest predictions only for groups IP3 and IP4, which are the groups with the largest indigenous population, and those who have traditionally suffered more from biased predictions; while other models that do not correct for under-reporting provide slightly fairer predictions for groups IP1 and IP2, which are those with the lowest percentages of indigenous population and that represent groups that have been traditionally associated to lower biases by prediction models. Similar experiments with violent crimes revealed that BURC achieved the highest fairness for all IP groups except for IP1 (see Table A.3 in Appendix). These results also show that by correcting for under-reported crime rates, we are slightly positively discriminating in favor of disadvantaged municipalities with mid to high volumes of indigenous people.

Based on results for the mean difference for both property and violent crimes, we have the following high-level observations for both protected attributes: 1) the ground truth reported crimes ($\boldsymbol{z}$) show high bias both in terms of income and indigenous groups, as the sum of absolute MD is large; 2) the predictions from all the models have lower MD for each group than the ground truth, suggesting that using urban hotspot features for crime prediction decreases the bias (increases fairness) when compared to the *ground truth*; 3) although BURC's prediction for reported crimes ($\hat{\boldsymbol{z}}_{BURC}$) is less fair in terms of MD, the advantage of BURC is that it can predict the true crimes including those failed to be recorded in the crime statistics. The inferred true crimes ($\hat{\boldsymbol{y}}_{BURC}$) reduce the AbsSum almost by half compared with the reported crimes and is much fairer than the baselines. This suggests that modeling under-reporting can improve the prediction accuracy and fairness at the same time, because no fairness regularization is added to limit the accuracy of the prediction for the observed reported crimes ($\hat{\boldsymbol{z}}_{BURC}$). $\hat{\boldsymbol{z}}_{BURC}$ has a distribution more similar to the ground truth ($\boldsymbol{z}$) than the baselines, thus also suffers from data bias in reported crimes, albeit lower than $\boldsymbol{z}$. However, by modeling under-reporting through $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{y}}_{BURC}$ manages to mitigate bias and increase fairness. The fact that $\hat{\boldsymbol{y}}_{BURC}$ is much less biased suggests we can use $\hat{\boldsymbol{y}}_{BURC}$ to guide crime-related decision making.

	IcQ1	IcQ2	IcQ3	IcQ4
RF	50.0 (7.4)	96.1 (3.0)	197.0 (2.5)	1507.3 (1.6)
BAG	102.2 (14.8)	105.1 (3.5)	212.9 (2.8)	1579.2 (1.7)
XGB	69.7 (10.1)	87.1 (2.8)	**159.9 (2.0)**	1602.9 (1.7)
BURC	**26.4 (3.9)**	**73.3 (2.5)**	180.6 (2.3)	**1154.3 (1.2)**

Table 4: The absolute group RMSE (relative RMSE to the group average in the brackets) for income groups. BURC has more balance performance across all income groups and reduces relative errors in the low income group substantially.

	IP1	IP2	IP3	IP4
RF	**14.5 (5.2)**	**275.4 (2.9)**	1824.0 (1.5)	89.3 (4.0)
BAG	27.7 (10.7)	309.8 (3.3)	1901.3 (1.5)	151.9 (6.5)
XGB	21.5 (11.1)	337.4 (3.4)	1901.1 (1.6)	93.9 (4.3)
BURC	59.6 (19.5)	363.4 (3.6)	**1277.5 (1.0)**	**46.4 (2.0)**

Table 5: The absolute group RMSE (relative RMSE to the group average in the brackets) for indigenous groups. BURC reduces substantially the prediction errors for IP3 and IP4, *i.e.,* municipalities with large indigenous population.

4.2.4 Fairness: Group Error. In this section, we evaluate the fairness in terms of group error. With this metric, we aim to find the best model with balanced performance for each protected attribute. Together with the mean difference, these two metrics will allow us to identify the best model in terms of performance (lowest error) and fairness. Here we use RMSE to measure performance as the error of the predictions. However, as shown in Tables 2 and 3, different groups for a protected attribute have different scales for the number of reported crimes *e.g.,* the average of reported crimes in group *IcQ1* is different from the average in group *IcQ4*. Therefore, we calculate not only the absolute RMSE but also the relative RMSE - normalized by the group average of reported crimes - as shown in Tables 4 and 5. We only discuss property crime prediction results. Violent crime analyses have a similar outcome and are presented in the Appendix (see Tables A.4 and A.5).

Based on the results for group errors as well as the mean difference, we make the following observations: 1) In terms of absolute errors, BURC substantially reduces the large errors observed in the baselines *e.g.,* the RMSE for *IcQ4* is reduced from 1507.33 to 1154.32 or *IP3* is reduced from 1823.94 to 1277.39; this error reduction allows BURC to make more balanced predictions across different groups, thus increasing accuracy and fairness. This also explains why BURC decreases the RMSE by 21%, a much larger improvement than MAE, as mentioned in Section 4.2.2; 2) In terms of relative errors, the prediction errors are distributed more evenly over the income groups when compared to other baselines, and BURC substantially reduces the relative errors in the lowest income group; 3) Although BURC does not achieve the lowest group errors for all groups, BURC consistently makes good predictions for disadvantaged groups, such as municipalities with low income or municipalities with high percentages of indigenous population. This is meaningful because BURC provides higher confidence in that disadvantaged groups are not unfairly treated in the prediction; 4) BURC performs similarly both in terms of mean difference and group error *i.e.,* BURC has good scores for almost all income groups and for municipalities with large indigenous population, confirming that by addressing under-reporting both performance and fairness can be improved.

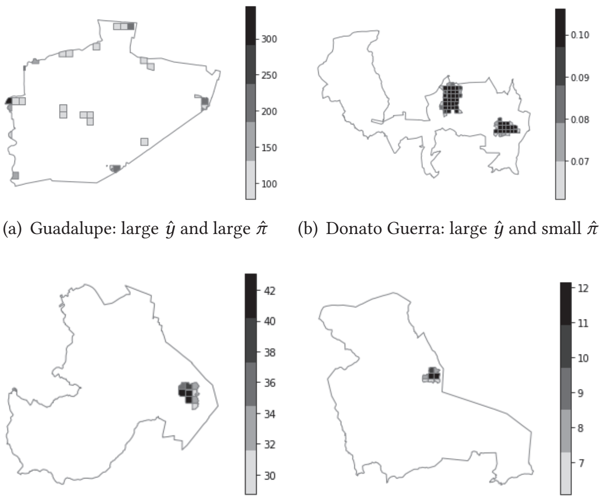
5 INSIGHTS ABOUT CRIME OCCURRENCE AND UNDER-REPORTING

In this section, we aim to quantify the influence of different mobility and socio-economic features on the true crime occurring rates and reporting rate in BURC. This analysis will reveal insights that could be used to 1) understand better the relationship between crimes and mobility patterns so as to improve safety in cities, and 2) evaluate the role that demographic and socio-economic data including poverty rate, unemployment or gender play in under-reporting so as to inform policies to encourage reporting.

Coefficient	Feature	Property Crime	Violent Crime
$\boldsymbol{\alpha}$	α_0 (intercept)	0.78 (0.03)	5.89 (0.03)
	log(NHS)	0.93 (0.01)	0.08 (0.01)
	log(AHS)	0.45 (0.01)	1.01 (0.01)
	log(COMP)	0.49 (0.02)	1.05 (0.02)
	log(MCOMP)	-1.49 (0.02)	-1.83 (0.02)
	log(COHE)	-3.89 (0.51)	-12.49 (0.39)
	log(PROX)	-0.71 (0.01)	-0.09 (0.01)
	log(NMI)	5.67 (0.51)	13.55 (0.39)
	log(NMMI)	-2.12 (0.01)	-1.58 (0.01)
$\boldsymbol{\beta}$	β_0 (intercept)	2.71 (0.04)	-27.69 (0.19)
	log(UR)	0.18 (0.01)	/
	log(PR)	-1.59 (0.01)	-0.28 (0.00)
	log(AR)	/	2.08 (0.04)
	log(NMR)	/	4.90 (0.03)
	log(M/F)	/	-0.61 (0.04)
	log(M/FHH)	/	-1.40 (0.01)
	log(FR)	/	0.47 (0.00)

Table 6: Mean and standard deviation (Std) for posterior distribution of the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the link function for corresponding features. "/" means the corresponding feature is not used in the prediction model.

For that purpose, we fit the proposed BURC model with all the reported crime statistics for: (1) property crimes and (2) violent crimes for the 1,379 municipalities. The distribution of the mean point estimate for the reporting rate for property and violent crimes across all municipalities, reveals a prevalent under-reporting issue with 94% of municipalities having less than 10% of violent crimes being reported (see Figure A.5 in Appendix). These results are consistent with the findings of the ENVIPE survey in Mexico where under-reporting rates were reported to be around 90% from 2010 to 2014 [15]. To understand the role that mobility and socio-economic features play on the true crime occurring rates and reporting rate in BURC, we compute the mean point estimate of the coefficients $\boldsymbol{\alpha}$ in the log link function and $\boldsymbol{\beta}$ in the logistics link function, respectively. Table 6 shows the coefficients for both property and violent crimes models.



(a) Guadalupe: large $\hat{y}$ and large $\hat{\pi}$ (b) Donato Guerra: large $\hat{y}$ and small $\hat{\pi}$

(c) Emiliano Zapata: small $\hat{y}$ and large $\hat{\pi}$ (d) Ziracuaretiro: small $\hat{y}$ and small $\hat{\pi}$

Figure 2: Permanent hotspot distribution in four sample municipalities to show the diverse spatial structure with one or multiple activity centers. The legends represent the footfall per hotspot. Varying levels of predicted volumes of true crimes $\hat{y}$ and reporting rate $\hat{\pi}$ per municipality are reported in Table 7.

municipality	(a)	(b)	(c)	(d)
z	4579	46	9	20
$\hat{z}_{BURC}$	4563.41	65.45	30.96	6.01
$\hat{y}_{BURC}$	4738.33	1176.45	50.32	64.30
$\hat{\pi}$	0.96	0.06	0.62	0.09
NHS	27	60	9	8
AHS	5.42	13.39	1.92	1.43
COMP	0.75	1.20	0.55	0.32
MCOMP	0.79	1.18	0.53	0.31
COHE	0.03	0.13	0.80	0.88
PROX	0.18	0.36	0.92	0.95
NMI	0.03	0.13	0.80	0.88
NMMI	0.02	0.13	0.80	0.88
PR	0.85	41.47	5.30	25.49
UR	4.79	7.52	8.74	2.39

Table 7: The ground truth volumes of reported property crimes z, predicted reported crimes $\hat{z}_{BURC}$, predicted true crimes $\hat{y}_{BURC}$, predicted reporting rate $\hat{\pi}$, urban hotspots features and reporting rate determinants, poverty rate (PR) and unemployment rate (UR), for the examples in Figure 2.

5.1 True Crime Rates Analysis

For the true crime occurring rates, we have one intercept term, α_0, and eight coefficients corresponding to the eight urban hotspot features (expressed in log scale). $\alpha_0 = 0.78$ in the property crime model represents the setting when all urban hotspot features take the value of 1 and for which the true crime occurring rate is 2.18 (α_0 is in the log link function and $exp(0.78) = 2.18$). Similar interpretation applies to the violent crime model. Positive (Negative) coefficients mean that larger (smaller) feature values are associated to the larger (smaller) true crime occurring rates. The coefficients for NHS and AHS are positive, which means that the more hotspots detected *i.e.,* the more active people move around in the municipality, the more crimes there are. For the urban sprawl features, whether or not to weigh the distance between two hotspots by population density has different effects on the crime occurring rate. MCOMP has a negative coefficient and the scale is larger than the coefficient for COMP, suggesting that if the population is more spread out relative to the size of the municipality, the crime incident numbers will be smaller. In fact, having the population more spread out translated into low population density, which means that the potential targets for property crimes are sparse. Note that the maximum value for the urban compactness features (COHE, PROX, NMI and NMMI) is 1 which represents the most compact form *i.e.,* the reference circle. The negative coefficient for COHE, PROX and NMMI suggests that the minimum crime occurring rate is achieved in the most compact form; as the hotspots become less compact with respect to the equal-area reference circle, the crime occurring rate increases.

As an example to delve into these relationships, Figure 2 and Table 7 show the distribution of permanent hotspots and the variables in the BURC property crime models for four example municipalities in our dataset: Guadalupe (a), Donato Guerra (b), Emiliano Zapata (c) and Ziracuaretiro (d), respectively. These four municipalities have different levels of true crime occurring rate and reporting rate, as shown in Table 7. Recall that the predicted volume of reported crimes is the expected value of the Poisson distribution and therefore could be smaller than the ground truth reported crime; and that the reporting rate is the ratio of predicted reported crimes $\hat{z}$ over predicted true crimes $\hat{y}$. Looking at Table 7 and comparing Figure 2(a) and 2(b) with Figure 2(c) and 2(d), we observe that municipalities with high volumes of true crimes ((a) and (b) have 4738.33 and 1176.45 as shown in Table 7) tend to have disperse spatial structure *i.e.*, have multiple activity centers; while municipalities with low volumes of true crimes ((c) and (d) have 50.32 and 64.30 as shown in Table 7) tend to be more compact *i.e.*, only one activity center is identified.

5.2 Reporting Rates

For the reporting rates, Table 6 shows that we have one intercept term, β_0, two coefficients for property crimes and six for violent crimes corresponding to different socioeconomic determinants in log scale. $\beta_0 = 2.71$ in the property crime model reflects that when the PR and UR are 1%, the reporting rate is 93.8% (β_0 is in the logistics link function and inverse logit of 2.71 is 0.938). Similar interpretation applies to the violent crime model. In previous studies about under-reporting of property crimes, when studied independently, higher poverty and unemployment levels are associated to higher under-reporting [25, 35]. Here we model the PR and UR together and our results show that the scale and direction of influence on the reporting rate is different. In our model, poverty rate has a much larger influence on the reporting rate than the unemployment rate. The larger the poverty rate is, the smaller the reporting rate is, as reflected previously in the literature [35]. On the other hand, unemployment rate has a small and positive coefficient, meaning that controlling the influence from the poverty rate, unemployment rate only has a small effect on reporting rate with larger unemployment rate corresponding to slightly higher reporting rate. However, this coefficient is extremely small to draw any conclusions. For violent crimes, the direction of influence of these determinants are mostly consistent with the findings in the literature [18, 33]. Going back to the examples in Figure 2 and Table 7, when we compare columns (a), (c) with (b), (d) in Table 2, we can observe that the poverty rate is much lower when the reporting rate is large than when it is small.

6 CONCLUSIONS

Reported crime data are an important basis for computational models that predict future crimes. Such models could potentially assist city agencies on better resource allocation to mitigate crime. Nevertheless, one of the most important sources of bias in such data is under-reporting, which can affect the quality of the final predictions. By leveraging the domain knowledge about possible determinants for the under-reporting of crimes *e.g.*, poverty rate has influence on the under-reporting of property and violent crimes, we developed a novel Bayesian model that explicitly addresses and corrects under-reporting issues. The experiments and evaluation show that our proposed model not only improves substantially the accuracy of mobility-based crime predictors, but that also provides fair predictions that balance performance across protected groups.

Acknowledgements: This work was funded by the National Science Foundation under Grant NSF-1750102.

REFERENCES

[1] Shlomo Angel, Jason Parent, and Daniel L Civco. 2010. Ten compactness properties of circles: measuring shape in geography. *The Canadian Geographer / Le Géographe canadien* 54, 4 (Dec. 2010), 441–461.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. (June 2017). arXiv:cs.LG/1706.02409

[3] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2015. Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics. *Big Data* 3, 3 (Sept. 2015), 148–158. https://doi.org/10.1089/big.2014.0054

[4] Stephen P Brooks and Andrew Gelman. 1998. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* 7, 4 (1998), 434–455.

[5] Elizabeth Burton. 2000. The Compact City: Just or Just Compact? A Preliminary Analysis. *Urban Stud.* 37, 11 (Oct. 2000), 1969–2006.

[6] T Calders, A Karim, F Kamiran, W Ali, and X Zhang. 2013. Controlling Attribute Effect in Linear Regression. In *2013 IEEE 13th International Conference on Data Mining*. 71–80. https://doi.org/10.1109/ICDM.2013.114

[7] Carlos Caminha, Vasco Furtado, Tarcisio H C Pequeno, Caio Ponte, Hygor P M Melo, Erneson A Oliveira, and José S Andrade, Jr. 2017. Human mobility in large cities as a proxy for crime. *PLoS One* 12, 2 (Feb. 2017), e0171609.

[8] Charlie Catlett, Eugenio Cesario, Domenico Talia, and Andrea Vinci. 2019. Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive Mob. Comput.* 53 (Feb. 2019), 62–74. https://doi.org/10.1016/j.pmcj.2019.01.003

[9] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. (Oct. 2018). arXiv:cs.LG/1810.08810

[10] Ronald V Clarke. 2012. Opportunity makes the thief. Really? And so what? *Crime Science* 1, 1 (Dec. 2012), 3. https://doi.org/10.1186/2193-7680-1-3

[11] Shay Cohen. 2019. *Bayesian Analysis in Natural Language Processing* (second edition ed.). Morgan & Claypool, California.

[12] CONEVAL. 2010. MEDICIÓN DE LA POBREZA. https://www.coneval.org.mx/Medicion/MP/Paginas/Medicion-de-la-pobreza-municipal-2010.aspx. Accessed: 2020-4-17.

[13] R N Davidson. 1981. *Crime and Environment.* St. Martin's Press.

[14] Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. 2017. Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *J. Comput. Graph. Stat.* 26, 2 (April 2017), 403–413.

[15] ENVIPE. 2014. National Survey of Victimization and Perception of Public Security. https://www.inegi.org.mx/programas/envipe/2014/. Accessed: 2020-4-17.

[16] John A Eterno, Arvind Verma, and Eli B Silverman. 2016. Police Manipulations of Crime Reporting: Insiders' Revelations. *Justice Q.* 33, 5 (July 2016), 811–835.

[17] Gabriel Ferreyra-Orozco. 2012. Race, Ethnicity, Crime and Criminal Justice in Mexico. In *Race, Ethnicity, Crime and Criminal Justice in the Americas*, Anita Kalunta-Crumpton (Ed.). Palgrave Macmillan UK, London, 169–191.

[18] Timothy C Hart and Callie Marie Rennison. 2003. *Reporting crime to the police, 1992-2000.* Technical Report.

[19] INEGI. 2010. 2010 Census of Population and Housing Units. https://www.coneval.org.mx/Medicion/MP/Paginas/Medicion-de-la-pobreza-municipal-2010.aspx. Accessed: 2020-4-17.

[20] Cristina Kadar and Irena Pletikosa. 2018. Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science* 7, 1 (July 2018), 26.

[21] Keith Kirkpatrick. 2017. It's Not the Algorithm, It's the Data. *Commun. ACM* 60, 2 (Jan. 2017), 21–23. https://doi.org/10.1145/3022181

[22] Wenwen Li, Tingyong Chen, Elizabeth A Wentz, and Chao Fan. 2014. NMMI: A Mass Compactness Measure for Spatial Pattern Analysis of Areal Features. *Ann. Assoc. Am. Geogr.* 104, 6 (Nov. 2014), 1116–1133.

[23] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. 2014. From mobile phone data to the spatial structure of cities. *Sci. Rep.* 4 (June 2014), 5276.

[24] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (Oct. 2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[25] Ziggy MacDonald. 2000. The impact of under-reporting on the relationship between unemployment and property crime. *Appl. Econ. Lett.* 7, 10 (Oct. 2000), 659–663. https://doi.org/10.1080/135048500415978

[26] Gail Mason and Rachael Stanic. 2019. Reporting and recording bias crime in New South Wales. *Current Issues in Criminal Justice* 31, 2 (April 2019), 164–180. https://doi.org/10.1080/10345329.2019.1594920

[27] George O Mohler. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* 30, 3 (July 2014), 491–497.

[28] G O Mohler, M B Short, Sean Malinowski, Mark Johnson, G E Tita, Andrea L Bertozzi, and P J Brantingham. 2015. Randomized Controlled Field Trials of Predictive Policing. *J. Am. Stat. Assoc.* 110, 512 (Oct. 2015), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

[29] Elías Moreno and Javier Girón. 1998. Estimating with incomplete count data A Bayesian approach. *J. Stat. Plan. Inference* 66, 1 (Jan. 1998), 147–159.

[30] SESNSP. 2011. Datos Abiertos de Incidencia Delictiva. https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva. Accessed: 2020-4-17.

[31] Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. 2019. A Hierarchical Framework for Correcting Under-Reporting in Count Data. *J. Am. Stat. Assoc.* (March 2019), 1–17.

[32] Roger Tarling and Katie Morris. 2010. Reporting Crime to the Police. *Br. J. Criminol.* 50, 3 (May 2010), 474–490. https://doi.org/10.1093/bjc/azq011

[33] Sean P Varano, Joseph A Schafer, Jeffrey Michael Cancino, and Marc L Swatt. 2009. Constructing crime: Neighborhood characteristics and police recording behavior. *J. Crim. Justice* 37, 6 (Nov. 2009), 553–563.

[34] S Verma and J Rubin. 2018. Fairness Definitions Explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. ieeexplore.ieee.org, 1–7. https://doi.org/10.23919/FAIRWARE.2018.8452913

[35] Barbara D Warner. 1997. Community characteristics and the recording of crime: Police recording of citizens' complaints of burglary and assault. *Justice Q.* 14, 4 (Dec. 1997), 631–650.

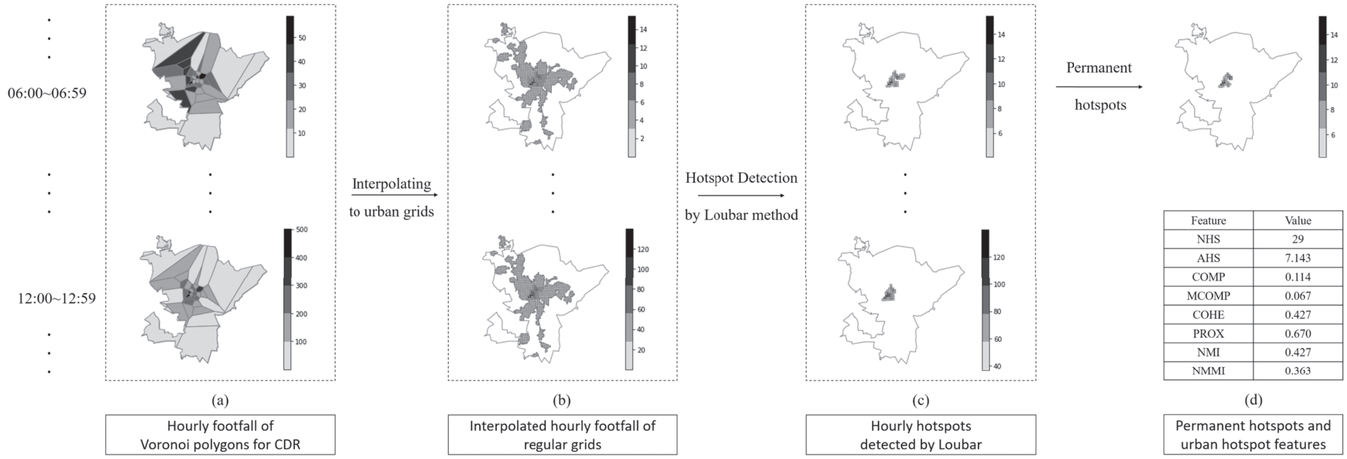
[36] Robert W Wassmer. 2000. Urban sprawl in a US metropolitan area: ways to measure and a comparison of the Sacramento area to similar metropolitan areas in California and the US. *CSUS Public Policy and Administration Working Paper* 2000-03 (2000).

[37] Sarah White, Tobin Yehle, Hugo Serrano, Marcos Oliveira, and Ronaldo Menezes. 2014. The spatial structure of crime in urban environments. In *International Conference on Social Informatics*. Springer, 102–111.

[38] Xian Wu, Chao Huang, Chuxu Zhang, and Nitesh V Chawla. 2020. Hierarchically Structured Transformer Networks for Fine-Grained Spatial Event Forecasting. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2320–2330.

[39] Xiangyu Zhao and Jiliang Tang. 2017. Modeling Temporal-Spatial Correlations for Crime Prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 497–506.

A APPENDIX



Feature	Value
NHS	29
AHS	7.143
COMP	0.114
MCOMP	0.067
COHE	0.427
PROX	0.670
NMI	0.427
NMMI	0.363

Figure A.1: Extracting urban hotspot features from CDR data.

Variable	Definition	Variable	Definition
A_i	the geographic area of region i	hs_i	the set of hotspots
$d_{j,k}$	the distance between centroids of hotspot j and k	p_j	the population in hotspot j
r_i	the points of hs_i in the rasterized format	$dr_{j,k}$	the distance between j- and k-th point
g_i	the center of gravity of hs_i	dg_j	the distance between the j-th rasterized point and g_i
$\boldsymbol{u}_i$	A set of mobility-based features	$\boldsymbol{s}_i$	A set of determinants of under-reporting behavior
y_i	Annual number of truth crimes	z_i	Annual number of reported crimes
λ_i	Occurring rate of Poisson distribution that models y_i	π_i	Under-reporting rate, the expected ratio of z_i to y_i

Table A.1: Notations for a region of interest i, e.g., a city.

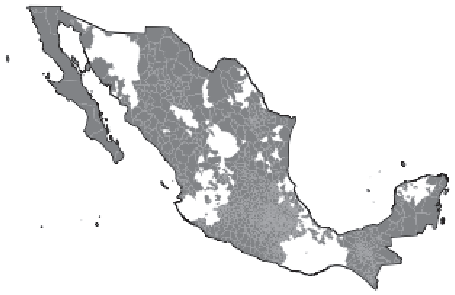


Figure A.2: Municipalities in Mexico studied in this paper, colored in grey.

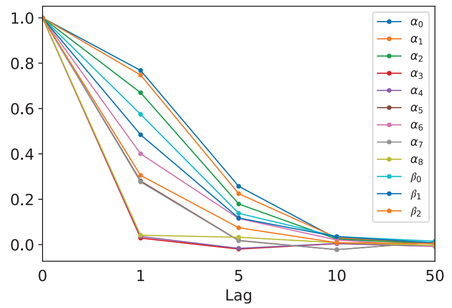


Figure A.3: Lag-k autocorrelation for the coefficients in BURC for the reported property crime experiment. The autocorrelation for all coefficients drops to zero with lag larger than 10.

	IcQ1	IcQ2	IcQ3	IcQ4	AbsSum
z	-1.28	-1.09	-0.76	3.20	6.32
$\hat{z}_{RF}$	-1.16	-0.96	-0.54	2.70	5.36
$\hat{z}_{BAG}$	-1.11	-0.94	-0.54	2.65	5.24
$\hat{z}_{XGB}$	-1.16	-0.99	-0.55	2.76	5.46
$\hat{z}_{BURC}$	-1.22	-1.01	-0.69	2.98	5.90
$\hat{y}_{BURC}$	**-0.77**	**-0.51**	**-0.19**	**1.47**	**2.94**

Table A.2: MD for protected attribute income group in violent crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{y}_{BURC}$) are fairer than baselines across all and each of the income groups.

	IP1	IP2	IP3	IP4	AbsSum
z	-0.60	-1.95	4.22	-1.02	7.79
$\hat{z}_{RF}$	-0.45	-1.12	2.73	-0.90	5.20
$\hat{z}_{BAG}$	**-0.43**	-1.13	2.69	-0.85	5.10
$\hat{z}_{XGB}$	-0.49	-1.27	2.96	-0.89	5.60
$\hat{z}_{BURC}$	-0.57	-1.61	3.62	-1.02	6.82
$\hat{y}_{BURC}$	-0.46	**-0.85**	**1.84**	**-0.47**	**3.63**

Table A.3: MD for protected attribute indigenous group in violent crime prediction. The predicted volumes of true crimes with under-reporting correction ($\hat{y}_{BURC}$) are fairer than baselines across all groups and are in favor of IP1, IP3 and IP4 which have more presence of indigenous population.

	IcQ1	IcQ2	IcQ3	IcQ4
RF	102.0 (4.7)	**183.8 (2.3)**	390.0 (2.0)	2560.4 (1.4)
BAG	205.8 (9.6)	195.0 (2.5)	393.0 (2.1)	2547.7 (1.4)
XGB	168.6 (8.0)	226.2 (2.9)	378.6 (2.0)	2568.9 (1.4)
BURC	**98.2 (4.7)**	194.7 (2.4)	**331.9 (1.7)**	**2305.4 (1.4)**

Table A.4: The absolute group RMSE (relative RMSE to the group average in the brackets) for income groups in violent crime prediction.

	IP1	IP2	IP3	IP4
RF	109.5 (33.9)	**502.5 (2.5)**	3128.0 (1.3)	189.7 (3.0)
BAG	127.5 (31.9)	522.1 (2.6)	3078.0 (1.3)	302.9 (4.7)
XGB	211.8 (93.6)	518.2 (2.6)	3131.6 (1.3)	255.3 (4.0)
BURC	**22.9 (9.8)**	518.4 (2.7)	**2719.8 (1.2)**	**120.2 (1.8)**

Table A.5: The absolute group RMSE (relative RMSE to the group average in the brackets) for indigenous groups in violent crime prediction.

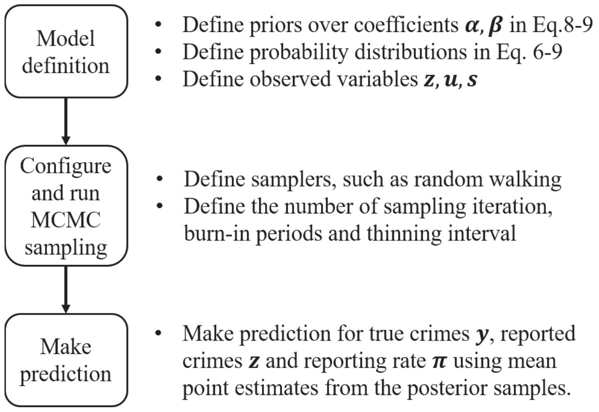

Model definition
• Define priors over coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in Eq.8-9
• Define probability distributions in Eq. 6-9
• Define observed variables $\boldsymbol{z}, \boldsymbol{u}, \boldsymbol{s}$
Configure and run MCMC sampling
• Define samplers, such as random walking
• Define the number of sampling iteration, burn-in periods and thinning interval
Make prediction
• Make prediction for true crimes $\boldsymbol{y}$, reported crimes $\boldsymbol{z}$ and reporting rate $\boldsymbol{\pi}$ using mean point estimates from the posterior samples.

Figure A.4: The process of implementing BURC using NIMBLE.

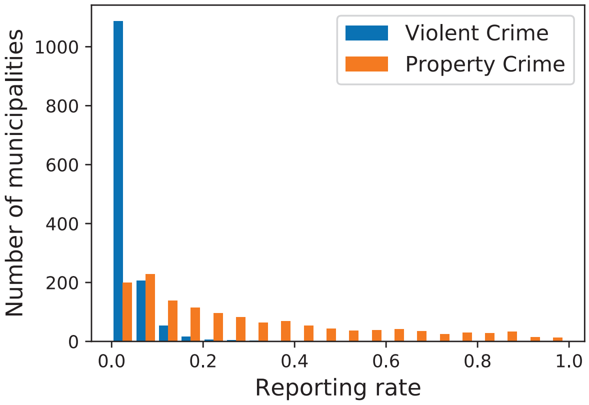


Figure A.5: Distribution of the reporting rate for violent crimes and property crimes across all municipalities. Violent crimes have more serious under-reporting issue.