

The effect of differential victim crime reporting on predictive policing systems

Nil-Jana Akpınar
nakpinar@stat.cmu.edu
Department of Statistics and Data
Science & Machine Learning
Department
Carnegie Mellon University

Maria De-Arteaga
Information, Risk, and Operations
Management Department
McCombs School of Business
University of Texas at Austin

Alexandra Chouldechova
Heinz College & Department of
Statistics and Data Science
Carnegie Mellon University

ABSTRACT

Police departments around the world have been experimenting with forms of place-based data-driven proactive policing for over two decades. Modern incarnations of such systems are commonly known as hot spot predictive policing. These systems predict where future crime is likely to concentrate such that police can allocate patrols to these areas and deter crime before it occurs. Previous research on fairness in predictive policing has concentrated on the feedback loops which occur when models are trained on discovered crime data, but has limited implications for models trained on victim crime reporting data. We demonstrate how differential victim crime reporting rates across geographical areas can lead to outcome disparities in common crime hot spot prediction models. Our analysis is based on a simulation¹ patterned after district-level victimization and crime reporting survey data for Bogotá, Colombia. Our results suggest that differential crime reporting rates can lead to a displacement of predicted hotspots from high crime but low reporting areas to high or medium crime and high reporting areas. This may lead to misallocations both in the form of over-policing and under-policing.

ACM Reference Format:

Nil-Jana Akpınar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3442188.3445877>

1 INTRODUCTION

Police departments around the world have been experimenting with computer-aided place-based predictive policing systems for over two decades. In a 1998 National Institute of Justice survey, 36% of police agencies employing over 100 sworn officers reported having the computing capability and data infrastructure to digitally generate crime maps [34]. Just a few years later, over 70% of agencies reported using such maps to identify crime hot spots as part of a broader adoption of CompStat approaches to policing [52]. More

modern incarnations of predictive policing date back to 2008, when the Los Angeles Police Department (LAPD) began its explorations of these systems, followed shortly thereafter by efforts such as the New York Police Department's use of tools developed by firms including Azavea, KeyStats and PredPol (2012+). Far from being a US-centric phenomenon, such systems are widely used throughout Europe, the UK, and China [2, 31, 49].

More recently, predictive policing systems have come under scrutiny due to their lack of transparency [53] and concerns that they may lead to further over-policing of minority communities by virtue of being trained on biased or “dirty” data [19, 33, 45]. Critics commonly point to the possibility that such systems may produce dangerous feedback loops, vicious cycles wherein data on recent arrests is used to deploy police in still greater numbers to neighbourhoods where they zealously seek out suspicious activity and conduct even more arrests. Recent work by Lum and Isaac [33] and Ensign et al. [19] has demonstrated both empirically and theoretically how such feedback loops can arise.

Proponents and developers of predictive policing technologies have argued that such analyses are based on models of crime and policing that do not accurately reflect the types of data used as inputs to such systems, nor the types of crime that they seek to predict. The analysis of Lum and Isaac [33], for instance, convincingly demonstrates how using data on drug arrests in Oakland, CA as inputs to a self-exciting point process (SEPP) model of the kind used in PredPol would result in high concentrations of policing in racial and ethnic minority neighbourhoods. Yet PredPol has stated that they do not use data on drug-related offenses (or traffic citations) in generating their predictions, nor do they use data on arrests [42]. Azavea, the creators and former owners of the HunchLab product, likewise note that their models focus on property and violent crimes, and the crime data they use is based on victim reporting rather than arrests [11].

Secondly, proponents and developers have argued that prior studies incorrectly assume that targeted policing strategies lead to an escalation in crime detection and, correspondingly, arrests. However, the adoption of hot spot policing strategies is predicated on an anticipated *deterrence* effect. Studies of the impacts of predictive policing on property and violent crimes and on arrests at targeted locations have produced mixed results. A 2014 analysis of a randomized controlled experiment (RCT) conducted by RAND in Shreveport, Louisiana found no statistical evidence of crime reduction in the prediction-targeted locations compared to control locations [30]. Another RCT conducted in Pittsburgh reported a 34% drop in serious violent crime in “temporary hot spots” and a

¹Code available at <https://github.com/nakpinar/diff-crime-reporting.git>.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FAccT '21, March 3–10, 2021, Virtual Event, Canada

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8309-7/21/03.

<https://doi.org/10.1145/3442188.3445877>

24% drop in “chronic hot spots” [20]. This study found no evidence of crime displacement to nearby locations, and reported that a total of 4 arrests took place during the experiment’s 20,000 hot spot patrols. A peer-reviewed study published by researchers affiliated with PredPol concluded that, while arrests were higher at predicted locations, they were lower or comparable once the counts were adjusted for differences in crime rate [9]. PredPol has reported crime drops ranging from 8-30% depending on the jurisdiction and type of crime [43]. While none of these counterarguments establish (or even claim) that the victim crime reporting data used to inform predictive policing systems is free from bias or leads to unbiased practices, they do point to a need for further investigation in settings that more closely mirror standard practice. Our work presents an initial step in this direction.

In this paper we empirically demonstrate how predictive policing systems trained exclusively on victim crime reporting data (rather than arrest data) may nevertheless suffer from significant biases due to variation in reporting rates. Our analysis is based on a simplified crime simulation patterned after district-level crime statistics for Bogotá, Colombia released by Bogotá’s chamber of commerce, Cámara de Comercio de Bogotá (CCB). We demonstrate that variation in crime reporting rates can lead to significant mis-allocation of police. These findings corroborate the effects of differential victim crime reporting on predictive policing models hypothesized in [12]. We also discuss the limitations of using reporting rates from existing crime victimization surveys to attempt to correct for such biases.

2 BACKGROUND & RELATED WORK

2.1 Feedback loops and other biases in predictive policing

Having already described the work of Lum and Isaac [33], we focus here on [19]. Ensign et al. [19] theoretically characterize why feedback loops occur by modeling arrest-based predictive policing systems via a generalized Pólya urn model. Their analysis also considers a scenario in which both reported and detected crimes (i.e., arrests) are used to update beliefs about existing crime rates. In the latter case they show that if the reported crime rates are an accurate reflection of underlying crime, then feedback loops can be avoided if either (a) underlying crime rates across regions are uniform to begin with; or (b) detected crimes aren’t considered at all. As we will discuss, there is considerable variation in the extent to which reported crimes reflect true underlying crime levels.

Richardson et al. [45] present three case studies where there is evidence that “dirty data” may have biased the targets of predictive policing systems. Their case studies focus primarily on person-based predictive policing systems. In the case of Maricopa County, Arizona, however, the authors report one instance in which biased data may have informed a PredPol system used by the Mesa Police Department and an RTMDx system used by the Glendale Police Department. As the authors note, due to the lack of transparency surrounding what data was used and how, it is difficult to draw definitive conclusions. This, however, does not make the documented patterns of biased practices against Maricopa County’s Latino residents any less concerning.

2.2 Victim crime reporting

Many countries and local governments conduct crime victimization surveys to better understand factors that drive differences in crime reporting rates, and to assess discrepancies between official crime statistics and victimization-based measures of criminal activity. According to the 2018 report released by the Bureau of Justice Statistics, which oversees the annual US National Crime Victimization Survey (NCVS), 61% of aggravated assaults, 63% of robberies, 38% of simple assaults, and only 25% rapes/sexual assaults are reported to police [39]. In this section we briefly overview different sources of disparities in victim crime reporting in the US context. We note that, while our data simulation is based on a 2014 survey conducted in Bogotá—and crime reporting rates are observed to be considerably lower there—a number of our conclusions apply to geography-associated disparities in reporting rates in general. In particular, our analysis indicates that, to the extent that these sources of disparity coincide with geography, we can expect significant under- or over-targeting to result.

The likelihood that a crime is reported to police has been found to be greater for older victims [4, 8, 28, 51] and when the victim is a woman [5]. It is also greater if a third party is present [5], if a weapon is present or the victim is injured [5, 56]. Furthermore, reporting rates tend to increase with the degree to which the victim is of higher socioeconomic status than the offender, which in part accounts for the greater likelihood of white victims reporting crimes perpetrated by black offenders for crimes such as assaults [55]. However, Xie and Lauritsen [55] also observe that black-on-black assaults had by far the highest reporting rate in their study (44%, compared to 25-33% for other racial pairs). This finding of high reporting rates for intra-racial black-on-black crimes was also reported in [1]. In other words, while some might expect reporting rates to be lowest in predominantly black communities, this does not appear to be borne out by the data. Furthermore, the degree of neighborhood socioeconomic disadvantage is not consistently associated with the likelihood of crime reporting [4]. An association has been observed for simple assaults, but not for robbery or aggravated assault. An extensive review of research in victim crime reporting is given in [54].

There are many reasons for why particular incidents may not be reported to police. These include fear of repercussion, victim perceptions that their victimization was ‘trivial’, or might be perceived as such by police, or personal relationships with the offender. Furthermore, there are documented examples of police actively discouraging victims from filing complaints in order to deflate serious crime statistics [45].

2.3 Predictive policing models

The literature on predictive policing has considered a range of different modeling approaches for spatio-temporal crime forecasting and hot spot selection [21]. To the best of our knowledge, only a small subset of models have been deployed and evaluated in practice.

PredPol, one of the largest vendors of predictive policing software in the US, has been one of few companies to publish modeling details of their hot spot prediction algorithm [36–38]. The PredPol algorithm relies on a Self-Exciting Spatio-Temporal Point Process (SEPP) model that uses the location and time of historical incidents

to predict the spatio-temporal distribution of future crime within a city. Hot spot predictions for subsequent time steps can be obtained by evaluating the predicted crime distribution on a grid of cells overlaying the city. This model, which has its roots in seismology, separates crime occurrences into “background crime” and “offspring crime” with the rationale that, similar to earthquakes which often trigger close-by aftershock earthquakes, crime tends to form clusters in time and space with burglars returning to the same areas or gang conflicts leading to retaliatory violence [37].

While the SEPP method models both the space and time distribution explicitly, many other common approaches focus on one component at a time. For instance, one straightforward approach is to apply time-series analysis to forecast crime counts in small pre-defined spatial units such as individual segments of streets or grid cells. In a field experiment with the Pittsburgh Bureau of Police, Fitzpatrick et al. [20] used a within-cell moving average of crime counts in order to predict chronic hot spots and a within-cell multi-layer perceptron on lagged crime count features to predict temporary crime flare-ups. The authors report that the relatively simple moving average model alone was able to capture more crime on average than other models including SEPP models. The Shreveport Police Department in Louisiana conducted experiments with a logistic regression model in 2012 [30]. In addition to different lagged crime counts, predictors also included the number of juvenile arrests in the past six month and the presence of residents on probation and parole in each of the 400-by-400-foot grid cells. Other methods focus on the spatial distribution of crime and aggregate the temporal component. Spatial kernel density estimates (KDE's) and risk terrain modelling, which involves risk factors beyond crime rates, are used to help identify chronic hot spots but generally require visual inspection if spatial discretization is to be avoided [21, 23]. A number of these models including SEPP's [17], KDE's, moving-average type models, and other approaches [3, 18] have previously been applied to historical crime data from Bogotá. Barreras et al. [3] found, for instance, that KDE models performed the best in their analysis.

In this study, we focus on SEPP models for crime hot spot prediction as they appear to be one of the most widely used and analysed type of model, a trend driven in part by PredPol's popularity and the descriptions of their models publicly available in peer-reviewed literature. For comparison purposes, we also consider a moving average model as analysed in [20]. Both models are based only on the location and time of previous crimes, which makes them particularly accessible to police departments.

3 METHODOLOGY

3.1 Self-Exciting Spatio-Temporal Point Processes

Self-Exciting Spatio-Temporal Point Processes (SEPP) are a commonly used class of models for applications in which the rate of events depends on nearby past events, e.g. modeling of earthquakes or the spread of infectious diseases. In the purely temporal case, this class of models is also known as Hawkes processes. We give a short introduction to SEPP, the specifications used in predictive policing and the model used in this study. A more detailed review of SEPP can be found in [44].

SEPPs separate events into two types: background events and offspring events. Background events are generally assumed to occur independently across space and time according to a Poisson point process. Each event can then cause offspring events in its vicinity according to a triggering function decaying in space and time. The rate of events at locations $(x, y) \in X \times Y \subseteq \mathbb{R}^2$ and times $t \in [0, T]$ is characterized by the conditional intensity, defined as

$$\lambda(x, y, t | \mathcal{H}_t) = \mu(x, y) + \sum_{\{k: t_k < t\}} g(t - t_k, x - x_k, y - y_k), \quad (1)$$

where $\mathcal{H}_t = \{(x_i, y_i, t_i)\}_{i=1}^n$ is the history of events up to time t which we will omit for simplification of notation. The background intensity $\mu(x, y)$ is often assumed to be time-independent while the triggering function $g(t - t_k, x - x_k, y - y_k)$ is generally chosen to be separable in time and space for computational simplicity. For each event (x_k, y_k, t_k) , the number of offspring events follows a Poisson distribution with mean

$$m = \int_{X \times Y} \int_T g(t, x, y) dt dx dy.$$

If properly normalized, $g(t - t_k, x - x_k, y - y_k)$ induces the probability distribution of the locations and times of these events. After model fitting, the SEPP can be used to predict the locations and times of future events. Assume we want to predict the number of events $N_{A,t}$ within an area $A \subseteq X \times Y$ at a given time $t = t'$. This prediction can be obtained by computing the integral

$$\widehat{N_{A,t'}} = \int_A \lambda(x, y, t | \mathcal{H}_{t'}, t = t') dx dy. \quad (2)$$

SEPP models were first applied to crime data for hot spot prediction by Mohler et al. [37]. Initially, the authors suggested non-parametric estimation of μ and g based on only background or offspring crimes respectively which requires a computationally expensive iterative stochastic declustering procedure. In subsequent work, Mohler [36] introduced a parametric approach that uses all data to estimate the background intensity with kernel density estimation and assumes a triggering function that is exponential in time and Gaussian in space. The benefit of this parametric approach is that model parameters can be estimated with a less expensive Expectation-Maximization procedure. In field experiments with the Los Angeles Police Department and the Kent Police Department, United Kingdom, [38] forgo a complicated spatial model by fitting a cell-wise constant background intensity and a triggering function only exponential in time.

In this work, we draw on a fully parametric SEPP model that is inspired by the simulations conducted in [37]. We assume a scaled Gaussian background intensity, defined as

$$\mu(x, y) = \frac{\bar{\mu}}{2\pi(15)^2} \exp\left(-\frac{x^2}{2(15)^2}\right) \exp\left(-\frac{y^2}{2(15)^2}\right), \quad (3)$$

where the spatial deviation is chosen purposefully large to ensure support on the whole city map. Our triggering function is similar to the proposed parametric functions and takes the form

$$g(t, x, y) = \theta \omega \exp(-\omega t) \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \exp\left(-\frac{y^2}{2\sigma_y^2}\right). \quad (4)$$

Choosing a fully parametric model allows us to analyze a best-case scenario of the bias introduced by differential crime reporting rates

as similar models can be used for data simulation and model fitting keeping error introduced by model misspecification at a minimum. In addition, the model choice enables efficient computation of the prediction integrals in Equation 2. For crime hot spot prediction, city maps are generally split into small areas by imposing a grid with fixed cell lengths. To predict the number of crimes within a cell at time t , integration over the estimated intensity function is necessary which can be computationally expensive depending on the exact model choice. To the best of our knowledge, the model we use is similar to the model employed by PredPol's commercial hot spot prediction software.

3.2 Expectation-Maximization Procedure

The parameters of the SEPP model in Equation 1-4 are estimated using maximum likelihood. As an analytical solution is intractable, [50] introduced an Expectation-Maximization (EM) algorithm that maximizes the log-likelihood. Assuming we know the branching structure of the data set $\{(x_i, y_i, t_i)\}_{i=1}^n$, i.e. which events were triggered by which previous events and which events come from the background process, we introduce a latent variable u_i which equals j if crime i was triggered by crime j and 0 if it was sampled from the background process. Given these latent quantities, the complete-data log-likelihood of the parameter vector $\Theta = (\bar{\mu}, \theta, \omega, \sigma_x, \sigma_y)$ can be written as

$$\begin{aligned} l(\Theta) = & \sum_{i=1}^n \mathbb{I}(u_i = 0) \log(\mu(x_i, y_i)) \\ & + \sum_{i=1}^n \sum_{j=1}^n \mathbb{I}(u_i = j) \log(g(t_i - t_j, x_i - x_j, y_i - y_j)) \\ & - \int_{X \times Y} \int_T \lambda(x, y, t) dt d(x, y), \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Given a data set of crime events, the EM algorithm provides an iterative procedure of estimating the triggering probabilities u_i and the parameters Θ . In the E step, we estimate the triggering probabilities based on current parameter values as

$$P(u_i = j) = \begin{cases} \frac{g(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(x_i, y_i, t_i)} & \text{if } t_j < t_i, \\ 0 & \text{else.} \end{cases}$$

and $P(u_i = 0) = \mu(x_i, y_i) / \lambda(x_i, y_i, t_i)$. These latent values can then be plugged into the expected complete-data log-likelihood which gives

$$\begin{aligned} \mathbb{E}[l(\Theta)] = & \sum_{i=1}^n P(u_i = 0) \log(\mu(x_i, y_i)) \\ & + \sum_{i=1}^n \sum_{j=1}^n P(u_i = j) \log(g(t_i - t_j, x_i - x_j, y_i - y_j)) \\ & - \int_{X \times Y} \int_T \lambda(x, y, t) dt d(x, y). \end{aligned}$$

In the M step, we maximize the expected log-likelihood with respect to Θ and return to the E step with the new parameter estimates. This procedure is repeated until the parameter values converge.

3.3 Bogotá Victimization and Reporting Survey

Victimization rates—i.e. the fraction of the population who has been victim of a crime within a given time window—and victim crime reporting rates—i.e. the fraction of crime victims who has reported the offense to the police—can generally not be assessed based on only police data but require large-scale surveys. Often, these surveys are not conducted or published with a high-enough spatial resolution to give a sense of differences at a local level. For instance, the US Bureau of Justice Statistics conducts a bi-annual National Crime Victimization Survey with around 95,000 households and publishes rates of victimization and crime reporting on a national level and aggregated by urban, suburban and rural areas [39].

In order to study the effect of differential victim crime reporting on predictive policing systems, which are generally limited to a single city, a higher spatial resolution of victimization and crime reporting rates is required. Fine-grained data sets like this are rare and, based on availability, we draw on district-level data from Bogotá, Colombia collected by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá (CCB).

The bi-annual CCB crime perception and victimization survey includes approximately 10,000 randomly selected participants from all socio-economic statuses and all 19 urban districts of Bogotá. Among other questions, participants are asked to indicate whether they have been the victims of a crime in the present calendar year and, if yes, whether they have reported the crime to the police. Results of the surveys are available on the CCB website and are used to inform the definition and adjustment of the city's public policies [13]. Not all of the published reports stratify results by districts. For our experiments, we use victimization and victim crime reporting rates stratified by district based on the survey that covers the first half of 2014 [14]. Districts, population sizes and rates are depicted in Figure 1. Both the crime victimization rates and the victim crime reporting rates vary significantly between different districts with victimization rates between 5 % and 18 % and victim crime reporting rate from 13 % to 33 %. Although the range of both rates can be expected to vary significantly between different cities and countries, this data allows us to analyze the impact of differential crime reporting on predictive policing in a realistic scenario.

3.4 Synthetic Data Generation

We simulate location and time of reported and unreported crime incidents in Bogotá districts according to the victimization and victim crime reporting rates displayed in Figure 1. In order to minimize possible errors due to model misspecification and instead concentrate on the effect of differential reporting rates, we sample data directly from a high-intensity SEPP and subsample according to each district's victimization rate. The background intensity of the SEPP is a sum over bivariate Gaussian distributions centered at 14 locations spread out evenly on the Bogotá map. Each background crime triggers offspring according to a triggering function that is Gaussian in space and exponential in time coinciding with the model we are fitting (see Equation 1-4). Since the data will be used to predict hot spots on a fixed grid, we impose a grid of $1 \text{ km} \times 1 \text{ km}$ cells on the Bogotá map as depicted in Figure 1. District membership of each cell is decided based on its center and each

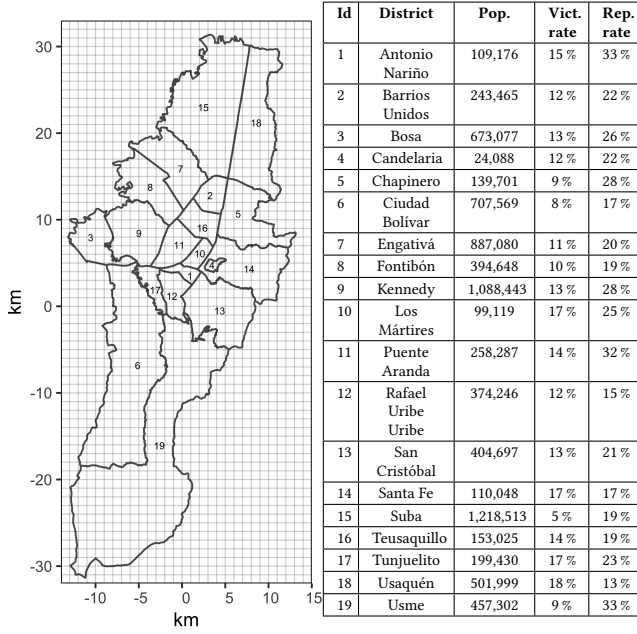


Figure 1: Bogotá district map with division into 1 km×1 km grid cells for hot spot prediction and survey-based victimization and victim crime reporting rates. Districts differ notably in size, population numbers, and rates.

point is attributed to the district of the cell it falls into. We discretize the time component into daily units and simulate crime data for 2,190 timesteps (6 years) as follows:

- (1) Sample a set of candidate points $C = \{(x_i, y_i, t_i)\}_{i=1}^N$ from λ and discard all points that fall outside of the city bounds or time horizon.
- (2) For each district d and data within its bounds $C_d \subseteq C$, we subsample $n_d \sim \text{Bin}(|C_d|, p_d)$ of the points to form the true crime data set \mathcal{D} , where

$$p_d = \frac{\text{population}(d) \cdot \text{victimization rate}(d) \cdot 12}{|C_d|}.$$

- (3) To get a data set of only reported crime, we subsample $n_d \sim \text{Bin}(|\mathcal{D}_d|, q_d)$ crimes for each district d where $\mathcal{D}_d \subseteq \mathcal{D}$ is the set of crimes falling into the given district and

$$q_d = \text{victim crime reporting rate}(d).$$

We implicitly assume that each person is victim of at most one crime which leads to the time scaling factor $2190/(365/2) = 12$ in step 2 as the CCB survey provides rates of victimization for a half-year period. In addition, district population counts are scaled by $1/40$ to speed up the run time of the whole simulation. The described sampling procedure for the true data \mathcal{D} ensures that crime is sampled according to population size and victimization rates but remains distributed according to a thinned SEPP that can be accessed for evaluation of the ground truth conditional intensity. Since $\mathcal{D} \sim p_d \lambda$, the true expected number of crimes in a subarea

A_d of district d in time $t = t'$ can be computed as

$$\mathbb{E}[N_{A_d, t'}] = \int_{A_d} p_d \lambda(x, y, t | \mathcal{H}_{t'}, t = t') d(x, y),$$

where $\mathcal{H}_{t'} = \{(x_i, y_i, t_i) \in C : t_i < t'\}$.

Figure 6 depicts a summary of the sampled number of crimes per district, the number of crimes expected according to above integral and the number of crimes as implied by the CCB survey showing that the synthetic data set has the desired rates of victimization for each of the districts.

4 RESULTS

4.1 Hot spot prediction procedure

We fit SEPP models (see Equation 1-4) on the full and reported crime data by discarding the data from the first 500 simulated time steps and training on the subsequent 1,500 days (≈ 4 years) of sampled incidents. Ignoring the first 500 time steps omits the period in which the data generating SEPP is converging to its equilibrium rate and provides a data set that more closely resembles the crime data over fixed time windows we would expect to see in practice. In addition, the time range of approximately 4 years is reasonably close to real crime data sets and falls well within the range of 2-5 years that is suggested by PredPol specifically [41].

The fitted models are used to predict crime intensities on a day-to-day basis for 189 evaluation days where, after each time step, the data for the time step is observed and added to the estimated intensity function for future predictions. On each prediction day, we compute the models' intensity integrals in each of the $1 \text{ km} \times 1 \text{ km}$ Bogotá grid cells. These integrals correspond to the absolute predicted crimes per cell and are subsequently used for hot spot selection. Since police are generally only able to patrol small fractions of a city effectively, we select the top 50 cells with highest predicted crime as hot spots which corresponds to approximately 5.7 % of the city's area. Results are aggregated over 50 simulation runs where each simulation samples a new crime data set.

4.2 Equity between districts

4.2.1 Relative number of predicted hot spots. We now discuss the equity of hot spot selection at a district-level. We start by examining how the number of predicted hot spots compares to the number of true hot spots per prediction day in each district. In the case where police are deployed in accordance with the model's predictions, this directly corresponds to the degree of police presence per district relative to a best-case hot spot policing program in which the true crime distribution is known.

Figures 2 and 3 depict the relative hot spot counts for a subset of districts over all evaluation time steps and simulation runs. For Figure 2, we set the relative count to 1 for cases in which the district has zero true hot spots and the model correctly predicts zero hot spots and exclude cases with zero true but non-zero predicted hot spots. We see that the SEPP model that was trained on all crime data, i.e. reported and unreported, performs well at selecting the correct number of hot spots uniformly over all districts (S1). This observation is unsurprising given that the fitted model closely resembles the data generating model.

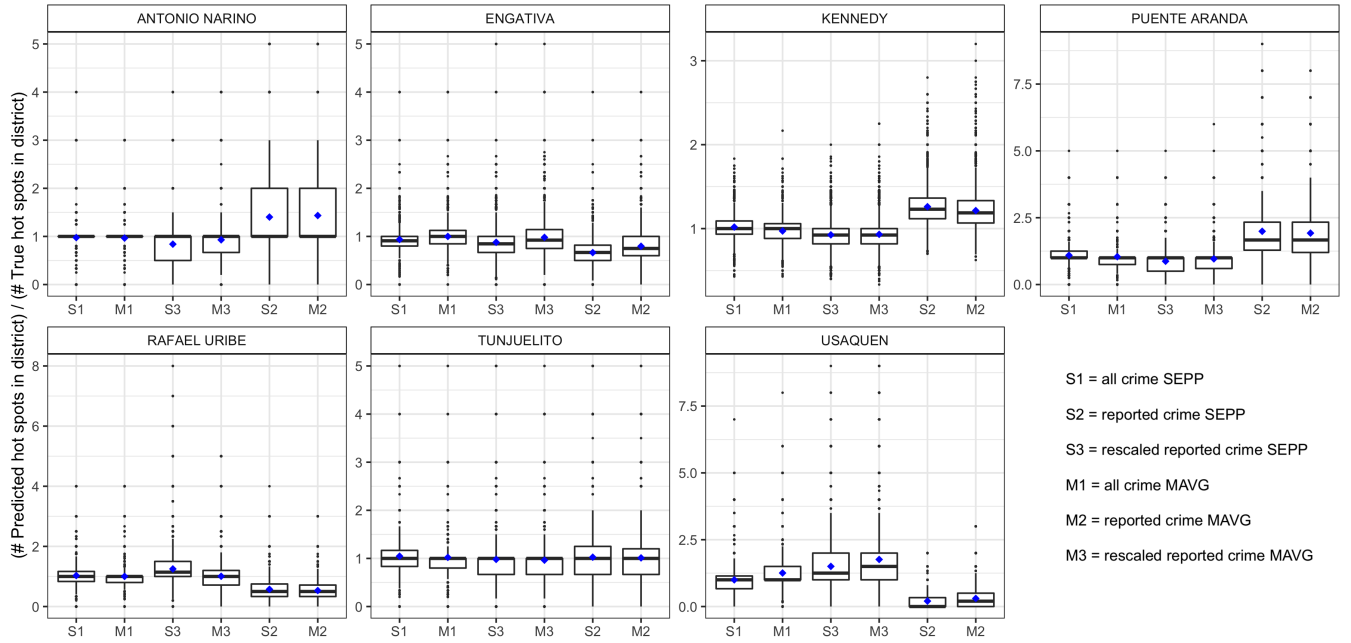


Figure 2: Relative number of predicted crime hot spots for a selection of Bogotá districts. Each data point represents a district-specific fraction at a given evaluation time step (189 days) in a given simulation run (50 runs). A total of 50 hot spots are selected at each time step. See Figure 7 for relative predicted hot spot counts for all districts.

In contrast, the SEPP model that was trained on only reported crime data (S2) is found to have differential performance across districts. Although in some districts, e.g. in Tunjuelito, the relative hot spot counts of the two models appear to be similar, the model with under-reporting on average overestimates the number of hot spots in districts such as Antonio Nariño, Puente Aranda or Kennedy, while underestimating the number of hot spots in districts such as Usaquén, Rafael Uribe Uribe or Engativá. The direction of the introduced error aligns with the victim crime reporting rates of the respective districts as compared to a Bogotá-wide average with fewer of the true hot spots detected in low reporting areas and instead overly many hot spots predicted in high reporting areas. In Usaquén, which with 13 % has the lowest victim crime reporting rate among all districts, only 20.4 % of the number of true hot spots are predicted on average. Meanwhile in Kennedy, which has a comparatively high reporting rate of 28 %, the model on average predicts 126.1 % the number of true hot spots.

Thus far, we have disregarded cases in which none of the true hot spots fall into a given district but the prediction model selects one or more cells. Figure 3 gives a summary of the fraction of cases with no true hot spots, further confirming the observed displacement effect of hot spot predictions. In Usaquén, the number of times crime hot spots are predicted when none of the true top 50 crime hot spots lie in the district is over twice as high in the full data SEPP compared to the reported crime SEPP. The same fraction increases more than threefold in the high-reporting district Antonio Nariño, and almost twofold in Puente Aranda. Notably, Figure 3 also shows that the displacement effect both impacts districts that almost always have

areas with highly concentrated crime and districts that do not. This phenomenon is a function of victimization rates, population sizes and the size of districts.

Finally, average absolute numbers of over- or underpredicted hot spots are displayed in Figure 9. Although comparison of relative counts ensures that districts of different sizes are evaluated similarly, in some cases we might be interested in the number of grid cells affected by the introduced bias as they roughly relate to the number of impacted individuals. For example, we see that the displacement of predicted hot spots based on differential victim crime reporting rates leads to on average 3.3 too many hot spots predicted in Kennedy while only 0.64 too many cells in Antonio Nariño are selected on average.

Overall, differential reporting rates across districts seem to lead to differentially well-measured aggregate crime levels which distorts the distribution of hot spots. If the police follows the model’s recommendations, the consequence would be an unfair allocation of police patrols where areas with low victim crime reporting rates are met with artificially decreased police presence while areas with higher reporting rates are chronically over-policed.

4.2.2 Crime threshold for hot spot selection. Calculating relative counts of predicted hot spots gives insights into how much under- or over-policing we can expect per district. A natural way of comparing between districts is to look at the true crime rates required for a cell to be selected as a hot spot. If this threshold is much lower for some districts than for others, the consequence could be more average police presence in these districts despite similar or even higher crime levels in other areas.

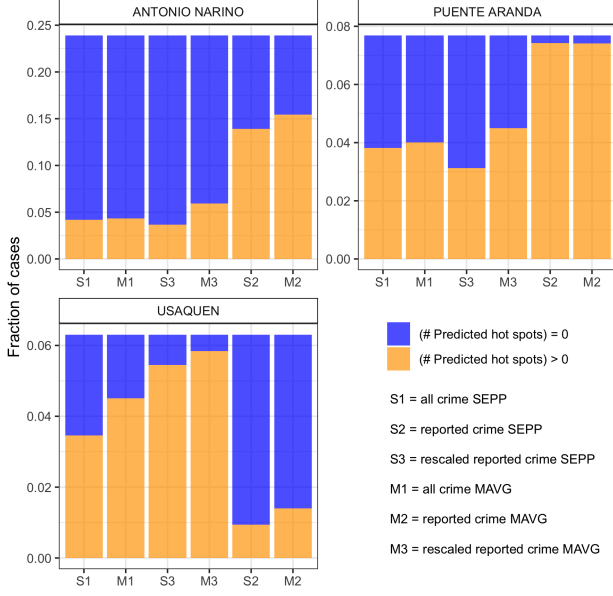


Figure 3: Fraction of prediction time steps with no true hot spots in districts. We separate instances into cases with predicted and no predicted hot spots. See Figure 8 for a version with all districts.

Figure 4 shows that the predicted crime rates implied by the reporting data SEPP model present a differentially well-adjusted approximation of true crime rates. Comparing the normalized maps in the Figure, the reported crime SEPP appears to overestimate the relative concentration of crimes in the high-reporting regions Kennedy and Antonio Nariño, and underestimate the relative concentration of crimes in low-reporting districts such as Rafael Uribe Uribe and Usaquén. Moreover, crime rate prediction seems to perform poorly in areas with little true crime. While the ground truth shows clear differences between crime intensities in areas such as Ciudad Bolívar and Usme, the model predictions in these districts appear to be almost indistinguishable.

In order to measure equity of model predictions between districts, we consider the minimum true crime rate that leads to a predicted hot spot at each prediction step and summarize the results in Figure 5. Since exact crime counts vary over time and this metric omits steps with no predicted hot spots falling into the respective district, the average thresholds have some variability even for full data models. However for districts that are regularly predicted to have hot spots, the full data SEPP model (S1) exhibits very similar hot spot prediction threshold of around 0.5 expected crimes per cell and time step where the low threshold is explained by the population scaling we conducted while simulating Bogotá crime data. In contrast, the model trained on only reported crime data results in varying thresholds even across districts which are regularly predicted to have hot spots. The district-wide average threshold of 0.45 true expected crimes per cell is increased in areas with low crime reporting, e.g. to a rate of 0.73 true crimes on average in Rafael Uribe Uribe and 0.62 in Usaquén. At the higher

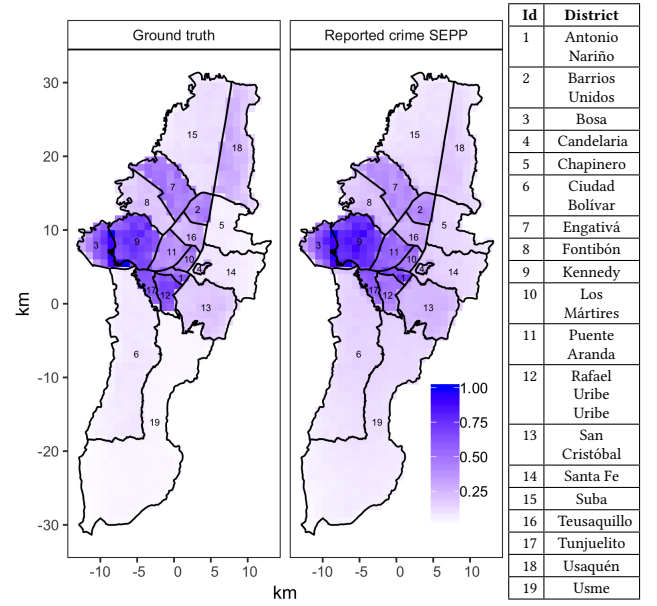


Figure 4: Normalized average crime in each cell. The left side depicts the average over true intensity integrals, while the right side uses predictions from the SEPP model trained on only reported crime data. In both cases, we normalize by dividing by the respective maximum average prediction value.

end of victim crime reporting rates, grid cells in Puente Aranda on average only require a rate of 0.32 true crimes and cells in Kennedy only 0.27 to be selected as a crime hot spots. More concretely, this means that on average the minimum true crime rate that leads to a predicted hot spot in Rafael Uribe Uribe is 2.7 times the minimum crime rate required in Kennedy. In order to rule out the possibility that Kennedy's threshold is artificially high because all of the cells in the district are regularly selected as hot spot, we examine the absolute predicted hot spot counts and find that at no time step more than 72.97 % of Kennedy is selected as hot spot area with a mean of 48.18 %.

These findings imply that crime hot spot prediction in real-world settings with differently sized regions and differential victimization and crime reporting rates can have noticeably biased outcomes that lead to over-policing of some areas of a city while others have higher levels of crime.

4.3 Scaling by victim crime reporting rates

It is not unusual for police and predictive policing algorithms to leverage data sets beyond registered crime incidents [22, 32, 46]. In the case presented here, one could imagine pairing the reported crime data with the survey data to attempt to correct the bias introduced by differential crime under-reporting. Intuitively, this entails rescaling the predicted crime rates according to the reporting rates. Of course, in most cases exact crime reporting rates are unknown to the police. However, as we discuss in this section, even in cases where the reporting rates are known, this rescaling does not necessarily recover the true crime distribution at the finest level.

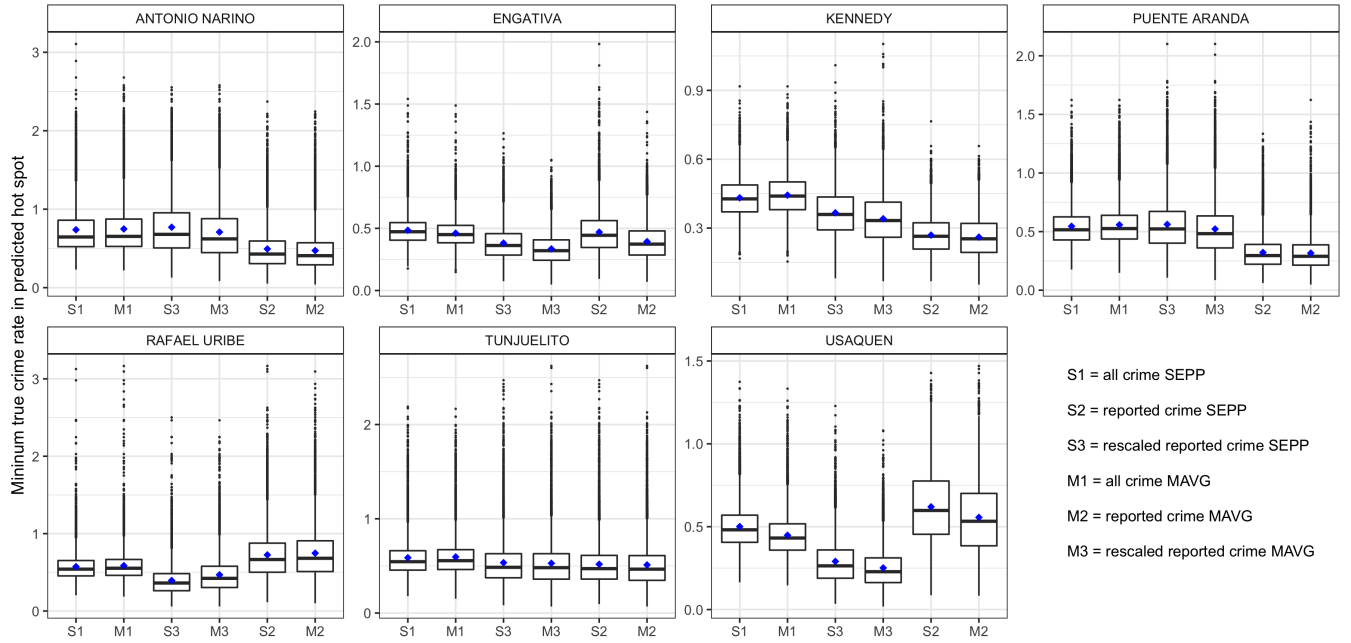


Figure 5: True crime thresholds for hot spot selection in a set of Bogotá districts. Each point corresponds to an evaluation time step (189 days) and a simulation run (50 runs). See Figure 10 for all districts.

We explore the rescaling approach as an additional model in our hot spot prediction simulation by taking the integrated intensities in grid cells supplied by the reporting data SEPP and dividing them by the victim crime reporting rate of the respective district. After rescaling, we select the cells with the top 50 highest predictions as hot spots analogous to the other models. The relative predicted hot spot counts of the rescaled model (S3) are displayed along the other models in Figure 2. Across the displayed districts, the mean relative number of predicted hot spots is just as close or closer to the number predicted by the full data model (S1) than the reporting data based predictions (S2). This indicates that the rescaling strategy successfully reduces outcome disparities. However, this conclusion is called into question when examining the implied minimum true crime rate for hot spot selection shown in Figure 5. For example in Usaquén, the rescaled model implies a visibly lower average true crime threshold for hot spot selection than the full data model, and in Engativá the difference between the full data and rescaled models appears to be larger than the difference between the full and reporting data models.

The conflict between the equity measures is observed because the relative predicted hot spot counts are an aggregate metric over all cells and not sensitive to which cells are selected, in contrast to the minimum true crime threshold for hot spot selection. Rescaling of the reporting data SEPP predictions increases predictions in all cells of a district by the same factor without accounting for how much crime was unobserved in each of the cells. As a consequence, the rescaled model selects an approximately correct number of hot spots in many of the districts while the exact cells might not coincide with the true hot spots. In order to recover the cell-wise true crime

distribution, a cell-by-cell rate of crime reporting would be required, which presupposes separate representative surveys in hundreds of micro-areas. While incorporating victimization survey information does help to reduce disparities to some extent, it evidently does not suffice in order to fully debias the prediction system.

4.4 Comparison to a moving average model

In this section we study the behavior of a simple moving average (MAVG) prediction model to assess whether our findings hold more generally outside of the SEPP prediction model setting. MAVG prediction models are fitted analogously to the SEPP models on the full and reported crime data sets. Despite being perhaps the simplest possible prediction model, MAVG's have been found to perform particularly well for detecting long-term hot spots [20] in real world data.

For our application, we aggregate crime in the same $1\text{ km} \times 1\text{ km}$ grid cells previously used and fit a within-cell MAVG model to predict the daily crime counts on the same training data sets as before. To obviate the need for tie-breaking that would arise if using simple averaging, we instead employ an exponentially-weighted MAVG model. The same model parameter is estimated for the entire Bogotá grid by searching over a linear scale of bandwidths for the exponential smoothing kernel and selecting the parameter that induces minimal average error with lagged prediction on the training data set. The models are updated on a daily basis by incorporating the crime counts of the previous day.

The performance of the full data MAVG model (M1), the reporting data MAVG model (M2), and the rescaled MAVG model (M3) are depicted in Figure 2, 3 and 5 alongside the corresponding SEPP

models. We observe that the MAVG models generally perform similarly to their respective SEPP counterparts. The full MAVG model (M1), for instance, performs on par with the full data SEPP model. Likewise, the reporting data MAVG (M2) induces similar outcome disparities in relative hot spot counts and minimum true intensities as the SEPP trained on victim crime reporting data, and the rescaled MAVG model (M3) struggles to correct the finer resolution bias similarly to the rescaled SEPP model.

At first glance these similarities might be surprising, especially because the true data was simulated from a SEPP. However both the SEPP and the MAVG model follow similar modeling ideas. While the MAVG model forgoes the spatial modeling component by discretizing into grid cells prior to prediction, whereas the SEPP has a continuous underlying intensity that is later integrated over grid cells, both methods model the time between events with an exponential function. In addition, both models make predictions based on a weighted average of previous nearby events and the weights can be fairly similar if we assume that the spatial deviation of the triggering function is small in comparison to the size of the grid cell such that most offspring crimes fall into the same cell as their parent. This assumption is often justified as the criminology literature tends to describe crime hot spots as micro areas of only a few blocks or street segments with high concentration of crime [21]. Indeed, in their randomized controlled field trials, the researchers affiliated to PredPol omit the spatial component of the SEPP altogether and discretize crimes into cells before modeling [38].

5 DISCUSSION

Our analysis demonstrates how predictive policing systems exclusively trained on victim crime reporting data can lead to spatially biased outcomes due to geographic heterogeneity in crime reporting rates. This in turn can result in over-policing of certain communities while others remain under-served by police.

Our findings are based on synthetic crime data simulated according to district-level victimization and victim crime reporting rates published by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá. We empirically evaluate the equity of predictions across districts of a hot spot prediction algorithm similar to the models used by PredPol. Our findings suggest that districts with low crime reporting rates have fewer of their crime hot spots detected by the algorithm. Conversely, districts with high crime reporting rates are found to have a higher concentration of predicted hot spots than the true crime levels would justify. Moreover, the effective true level of crime required for the model to predict a hot spot is found to vary by more than a factor of two across the districts.

We explore if known victim crime reporting rates can be used to debias hot spot predictions by scaling crime expectations appropriately. The results suggest that this is unsuccessful when reporting rates are known at a district level but hot spots are predicted at a smaller individual cell level since noise introduced by individually thinned crimes is propagated to the rescaled predictions which makes singling out of specific cells in comparison to other cells in the same district difficult.

Prior work has focused on feedback loops and the potential harms of arrest data-based predictive policing systems [19, 33]. Yet, in practice, predictive policing systems are based on data from

victim crime reports [11]. Our work presents an initial step toward understanding the effect of bias in victim crime reporting data on predictive policing systems. Our analysis demonstrates the importance of considering reporting rate variation when assessing predictive policing systems for potential harms and disparate impacts.

5.1 Limitations

5.1.1 Crime location vs. survey location. Victimization surveys generally provide us with information on crime reporting based on where people live, not based on where crimes occur. On a small scale like a single city, this spatial disparity makes it hard to take survey-based information into account for police allocation. While this limitation of how survey data is collected does not invalidate our findings—in our simulations we treat the reporting rates as reflecting rates of reporting for crimes *occurring* in the given district—it does present a challenge for using such survey data to de-bias predictions in practice. In order for survey data to be useful for this purpose, questions need to ask not only where respondents reside, but also where the victimization(s) occurred.

5.1.2 Static reporting rates and potential deterrence effects. Thus far, we do not take the effects of the actual interventions in the form of patrolled hot spots into account. We hypothesize that both victimization rates and victim crime reporting rates can be susceptible to police presence and a model that jointly describes the interplay of crime, reporting rates and police deployment is required for a more complete picture. One component currently omitted is a deterrence effect of policing. Failing to consider such effects could result in the reallocation of police patrols away from neighbourhoods where they are having the intended deterrence effect, precisely because reported crime rates would be lower when police are successful in deterring crime.

5.2 Implications and Generalizability

5.2.1 Relationship to socioeconomic advantage. Research on victim crime reporting shows that the decision to report a crime is influenced by the severity of crime [e.g. 25, 26], victim characteristics [e.g. 29, 47], and contextual factors such as neighborhood characteristics [e.g. 4, 48, 57]. While we lack information on victim and crime characteristics in the survey data, we are able to speak to a number of socio-technical implications of our results.

Prior research finds links between severe socioeconomic neighborhood disadvantage and lower reporting rates for simple assault incidents [4]. Goudriaan et al. [25] obtain similar results analysing crime incidents from the Netherlands paired with the Dutch Police Population Monitor survey. Some studies describe a more indirect effect of socioeconomic status on the likelihood of reporting. For example, [7] find that victims who are involved in illegal behavior are less likely to report violent acts against them to the police, and this effect is particularly pronounced in disadvantaged neighborhoods. The findings of [47] suggest that prior police-initiated contact with law enforcement has a negative impact on the reporting of future crime that is amplified for African Americans and poorer individuals. The authors of [10] study the help-seeking behavior of women who experience intimate partner violence. The study finds that, for the lowest income women, the severity of violence does not predict

whether law enforcement is contacted. With increasing income the severity becomes more indicative of victim crime reporting.

In the Bogotá survey data, there appears to be no clear association between reporting rates and socioeconomic advantage at the district level. Ciudad Bolívar, a district with large urban slums that is home to some of the most socioeconomically disadvantaged residents of Bogotá, has a reporting rate of 17%. In line with previous research, this lies well below the average reporting rate across districts of 22.7%. However, Usaquén, the district with the lowest reporting rate of 13%, is also one of the wealthiest districts in Bogotá. We hypothesize that this is in part explained by the spatial clustering of specific crime types. In particular, Usaquén experiences a greater proportion of residential burglary and theft than other districts [15, 22]. Given that victim crime reporting rates vary based on perceived severity [54], this might contribute to a decreased victim crime reporting rate. Additionally, this may also be influenced by intra-district heterogeneity of wealth, as socioeconomically disadvantaged neighborhoods such as El Códito are also located in this district.

There is no simple relationship between socioeconomic level of districts and the geographical disparities induced by the hot spot prediction algorithm. This is in part driven by the observed complexity in the relationship between socioeconomic status and crime reporting at the district level. For example, areas that we project to be over-policed under hot spot policing include the middle class district Antonio Nariño, the lower middle or working class district Puente Aranda and the working and lower class district Kennedy. Areas observed to be under-predicted likewise include districts inhabited by upper, middle, working and lower class residents. Thus our findings *do not* indicate that variation in crime reporting rates systematically disadvantages Bogotá's districts in a manner that falls along socioeconomic lines.

5.2.2 Relationship to demographics. Demographic factors such as age [4, 8, 28, 51], gender [5] and race [1, 55] can play a role in victim crime reporting. Desmond et al. [16] examine the change in Milwaukee's crime reporting rates after public broadcast of police violence against an unarmed black man. They find that, particularly in black neighbourhoods, residents were far less likely to report crime to police following the incident. Ultimately, race and ethnicity are often correlated with socioeconomic status and location, which makes it difficult to identify the direct relationships between demographic variables and victim crime reporting rates [46].

Due to data limitations, we are unable to provide an indepth discussion of the relationship between race, ethnicity and predictive disparities in the Bogotá context, as we do not have access to demographic information on the victimization survey participants. A discussion of the Bogotá-specific interplay of race, ethnicity, crime and policing, and how it might generalize to other contexts, thus remains beyond the scope of this paper.

5.2.3 Generalizability to other jurisdictions. Crime reporting decisions also operate in a macrolevel context encompassing specific cities, local governments or whole nations [54]. Gutierrez and Kirk [27] find that, within the US, metropolitan areas with greater proportions of foreign-born or non-US citizens have decreased crime reporting rates, and the results of [35] suggest that cities with more female police officers have higher rates of victim crime reporting for

violent crimes against women. Goudriaan et al. [24] analyze data from 16 Western industrialized countries and find that differences in crime reporting rates are not entirely explained by crime types, individual and local contexts, but vary with nation-level factors such as the perceived competence of the police at large.

Since our study is exclusively based on survey data from Bogotá, specific findings do not necessarily generalize to other geographies. In particular, while our analysis did not find evidence of a simple relationship between socioeconomic factors at a district level and predictive disparities, results would likely be different in regions—or at resolutions—where socioeconomic factors are more directly associated with reporting rates.

Although the specific spatial distribution and societal implications of the predictive disparities are likely to vary between different jurisdictions, our results suggest that some form of outcome disparity can be expected if victim crime reporting rates have sufficient spatial variation. Such spatial variation is relatively commonplace and can be expected if, for example, the city has some amount of socioeconomic segregation since crime reporting rates vary with neighborhood disadvantage [4, 7, 54].

5.2.4 Combining data sources and debiasing. Predictive policing algorithms rely on crime data collected by law enforcement that has repeatedly been found to be flawed, biased or in other ways 'dirty' [45]. Much of the attention has focused on biases in police-initiated and particularly arrest data, for instance racial bias in drug related arrests or traffic stops in the US [6, 40]. PredPol acknowledges some of these biases and publicly states that no drug-related offenses, traffic stops or arrest data are used in their prediction system [42]. Yet there is a lack of transparency as to how the data types that are 'too biased' to be included were identified, to what extent other data sources are biased, and which types of biases were considered. To the best of our knowledge, there has been no consideration of reporting biases although their link to socioeconomic, demographic and cultural factors as described in earlier sections is known. Motivated by this problem, we show that, even when predictive policing algorithms only operate on victim crime reporting data, and thereby attenuate the effects of biased police arrest practices, differential victim crime reporting rates can lead to geographically biased prediction outcomes.

In addition to police data, some predictive policing systems incorporate contextual data from other sources [22, 46]. For example, HunchLab combines public reports of crime and requests of police assistance with data including weather patterns, geographical features, schedules of major events and even moon phases [46]. In the setting of the hot spot prediction analyzed in this study, one could imagine the proposal to account for the bias introduced by differential reporting rates by scaling model outputs by the survey-based geographically stratified reporting rates. However, our preliminary experimentation suggests that, although in some cases bias can be decreased, a complete mitigation is not possible if the surveyed victim crime reporting rates do not have sufficient spatial resolution. For successful debiasing, we would require close-to-optimal estimates of victim crime reporting rates at the grid cell level, which is impossible to obtain in practice. Ultimately, it is unclear if debiasing victim crime reporting data is any easier than the unsuccessful previous efforts of mitigating bias introduced by arrest data.

REFERENCES

- [1] Edem F Avakame, James J Fyfe, and Candace McCoy. 1999. "Did you call the police? What did they do?" An empirical assessment of Black's theory of mobilization of law. *Justice Quarterly* 16, 4 (1999), 765–792.
- [2] Alexander Babuta and Marion Oswald. 2020. *Data Analytics and Algorithms in Policing in England and Wales*. Royal United Services Institute, London, UK.
- [3] Francisco Barreras, Alvaro Riascos, and Mónica Ribero. 2016. Comparison of different crime prediction models in Bogotá. (2016).
- [4] Eric P Baumer. 2002. Neighborhood disadvantage and police notification by victims of violence. *Criminology* 40, 3 (2002), 579–616.
- [5] Eric P Baumer and Janet L Lauritsen. 2010. Reporting crime to the police, 1973–2005: A multivariate analysis of long-term trends in the National Crime Survey (NCS) and National Crime Victimization Survey (NCVS). *Criminology* 48, 1 (2010), 131–185.
- [6] Katherine Beckett, Kris Nyrop, and Lori Pfingst. 2006. Race, Drugs, and Policing: Understanding Disparities in Drug Delivery Arrests. *Criminology* 44, 1 (2006), 105–137.
- [7] Mark T. Berg, Lee Ann Slocum, and Rolf Loeber. 2011. Illegal Behavior, Neighborhood Context, and Police Reporting by Victims of Violence. *Journal of Research in Crime and Delinquency* 50, 1 (2011), 75–103.
- [8] Stacey J Bosick, Callie Marie Rennison, Angela R Gover, and Mary Dodge. 2012. Reporting violence to the police: Predictors through the life course. *Journal of Criminal Justice* 40, 6 (2012), 441–451.
- [9] P Jeffrey Brantingham, Matthew Valasik, and George O Mohler. 2018. Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and public policy* 5, 1 (2018), 1–6.
- [10] Lauren Bennett Cattaneo and Heidi L. M. DeLovich. 2008. The role of socioeconomic status in helpseeking from hotlines, shelters, and police among a national sample of women experiencing intimate partner violence. *American Journal of Orthopsychiatry* 78, 4 (2008), 413–422.
- [11] Robert Cheetham. 2019. Why we sold HunchLab. <https://www.azavea.com/blog/2019/01/23/why-we-sold-hunchlab/>. Azavea. [Online; accessed 1/20/2021].
- [12] Maria Cuellar and Maria De-Arteaga. 2020. Algoritmos y crímenes. <https://www.semana.com/opinion/articulo/prevenccion-de-delitos-columnistas-maria-de-artea-gonzalez-y-maria-cuellar-correa/608182/>
- [13] Cámara de Comercio de Bogotá. [n.d.]. Encuesta de Percepción y Victimización. <https://www.ccb.org.co/Transformar-Bogota/Seguridad-y-Justicia/Encuesta-de-Percepcion-y-Victimizacion>. [Online; accessed 10/4/20].
- [14] Cámara de Comercio de Bogotá. 2014. *Encuesta de Percepción y Victimización - Primer semestre de 2014 (Chapinero)*. [Presentation].
- [15] Cámara de Comercio de Bogotá. 2015. *Observatorio de Seguridad en Bogotá: Balance de Seguridad en Bogotá - 2014*. Vol. 48.
- [16] Matthew Desmond, Andrew V. Papachristos, and David S. Kirk. 2016. Police Violence and Citizen Crime Reporting in the Black Community. *American Sociological Review* 81, 5 (2016), 857–876.
- [17] Mateo Dulce, Simón Ramírez-Amaya, and Álvaro Riascos. 2018. Efficient allocation of law enforcement resources using predictive police patrolling. *arXiv preprint arXiv:1811.12880* (2018).
- [18] Mateo Dulce Rubio et al. 2018. *Predicting criminal behavior with Levy flights using real data from Bogotá*. Master's thesis. Uniandes.
- [19] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*.
- [20] Dylan Fitzpatrick, Wilpen Gorr, and Daniel B. Neill. 2018. Hot-Spot-Based Predictive Policing in Pittsburgh: A Controlled Field Experiment. *Preprint* (2018). http://halley.exp.sis.pitt.edu/comet/presentColloquium.do?col_id=16153 [Online; accessed 1/20/21].
- [21] Dylan J. Fitzpatrick, Wilpen L. Gorr, and Daniel B. Neill. 2019. Keeping Score: Predictive Analytics in Policing. *Annual Review of Criminology* 2, 1 (2019), 473–491.
- [22] Alejandro Giménez-Santana, Joel M. Caplan, and Grant Drawve. 2018. Risk Terrain Modeling and Socio-Economic Stratification: Identifying Risky Places for Violent Crime Victimization in Bogotá, Colombia. *European Journal on Criminal Policy and Research* 24, 4 (2018), 417–431.
- [23] Wilpen L. Gorr and Yonglei Lee. 2014. Early Warning System for Temporary Crime Hot Spots. *Journal of Quantitative Criminology* 31, 1 (2014), 25–47.
- [24] Heike Goudriaan, James P. Lynch, and Paul Nieuwebeerta. 2004. Reporting to the police in western nations: A theoretical analysis of the effects of social context. *Justice Quarterly* 21, 4 (2004), 933–969.
- [25] Heike Goudriaan, Karin Wittebrood, and Paul Nieuwebeerta. 2006. Neighbourhood Characteristics and Reporting Crime. *The British Journal of Criminology* 46, 4 (2006), 719–742.
- [26] Martin S. Greenberg and R. Barry Ruback. 1992. *After the Crime*. Springer US.
- [27] Carmen M. Gutierrez and David S. Kirk. 2015. Silence Speaks: The Relationship between Immigration and the Underreporting of Crime. *Crime & Delinquency* 63, 8 (2015), 926–950.
- [28] Patricia Y Hashima and David Finkelhor. 1999. Violent victimization of youth versus adults in the National Crime Victimization Survey. *Journal of Interpersonal Violence* 14, 8 (1999), 799–820.
- [29] Keith L. Hullenaar and R. Barry Ruback. 2020. Gender Interaction Effects on Reporting Assaults to the Police. *Journal of Interpersonal Violence* (2020).
- [30] Priscilla Hunt, Jessica Saunders, and John S. Hollywood. 2014. *Evaluation of the Shreveport Predictive Policing Experiment*. RAND Corporation, Santa Monica, CA.
- [31] Fieke Jansen. 2018. *Data Driven Policing in the Context of Europe*. DATAJUSTICE project.
- [32] Brian Jordan Jefferson. 2017. Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race. *Annals of the American Association of Geographers* 108, 1 (2017), 1–16.
- [33] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.
- [34] Cynthia A. Malamian and Nancy Gladys La Vigne. 1999. *The use of computerized crime mapping by law enforcement: Survey results*. US Department of Justice, Office of Justice Programs, National Institute of Justice.
- [35] Amalia R. Miller and Carmit Segal. 2018. Do Female Officers Improve Law Enforcement Quality? Effects on Crime Reporting and Domestic Violence. *The Review of Economic Studies* 86, 5 (2018), 2220–2247.
- [36] George Mohler. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting* 30, 3 (2014), 491–497.
- [37] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. 2011. Self-Exciting Point Process Modeling of Crime. *J. Amer. Statist. Assoc.* 106, 493 (2011), 100–108.
- [38] 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. Amer. Statist. Assoc.* 110, 512 (2015), 1399–1411.
- [39] Rachel E. Morgan and Barbara A. Oudekerk. 2019. *Criminal victimization, 2018*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.
- [40] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jensen, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* 4, 7 (2020), 736–745.
- [41] PredPol. [n.d.]. <https://www.predpol.com/law-enforcement/#predPolicing> [Online; accessed 10/7/20].
- [42] PredPol. 2017. Machine Learning and Policing. <https://blog.predpol.com/machine-learning-and-policing>. [Online; accessed 1/20/21].
- [43] PredPol. 2017. Proven Results of our Predictive Policing Software. <https://www.predpol.com/results/>. [Online; accessed 1/20/21].
- [44] Alex Reinhart. 2018. A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications. *Statist. Sci.* 33, 3 (08 2018), 299–318.
- [45] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems and Justice. *New York University Law Review Online* 192 (2019).
- [46] Aaron Shapiro. 2017. Reform predictive policing. *Nature* 541, 7638 (2017), 458–460.
- [47] Lee Ann Slocum. 2017. The Effect of Prior Police Contact on Victimization Reporting: Results from the Police–Public Contact and National Crime Victimization Surveys. *Journal of Quantitative Criminology* 34, 2 (2017), 535–589.
- [48] Lee Ann Slocum, Terrance J. Taylor, Bradley T. Brick, and Finn-Aage Esbensen. 2010. Neighborhood structural characteristic, individual-level attitudes, and youths' crime reporting intentions. *Criminology* 48, 4 (2010), 1063–1100.
- [49] Daniel Sprick. 2020. Predictive Policing in China: An Authoritarian Dream of Public Security. *Nordic Journal of Law and Social Research* (2020).
- [50] Alejandro Veen and Frederic P. Schoenberg. 2008. Estimation of Space–Time Branching Process Models in Seismology Using an EM–Type Algorithm. *J. Amer. Statist. Assoc.* 103, 482 (June 2008), 614–624.
- [51] Adam M. Watkins. 2005. Examining the disparity between juvenile and adult victims in notifying the police: A study of mediating variables. *Journal of research in Crime and Delinquency* 42, 3 (2005), 333–353.
- [52] David Weisburd, Rosann Greenspan, Stephen Mastrofski, James J. Willis, Police Foundation, and United States of America. 2008. Compstat and organizational change: A national assessment. *National Institute of Justice* (2008).
- [53] Ali Winston. 2018. Palantir has secretly been using New Orleans to test its predictive policing technology. <https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd> The Verge. [Online accessed; 1/20/21].
- [54] Min Xie and Eric P. Baumer. 2019. Crime Victims' Decisions to Call the Police: Past Research and New Directions. *Annual Review of Criminology* 2, 1 (2019), 217–240.
- [55] Min Xie and Janet L. Lauritsen. 2012. Racial context and crime reporting: A test of Black's stratification hypothesis. *Journal of quantitative criminology* 28, 2 (2012), 265–293.

- [56] Min Xie, Greg Pogarsky, James P Lynch, and David McDowall. 2006. Prior police contact and subsequent victim reporting: Results from the NCVS. *Justice quarterly* 23, 4 (2006), 481–501.
- [57] Lening Zhang, Steven F. Messner, and Jianhong Liu. 2007. An exploration of the determinants of reporting crime to the police in the city of Tianjin, China. *Criminology* 45, 4 (2007), 959–984.

A SUPPLEMENTARY FIGURES

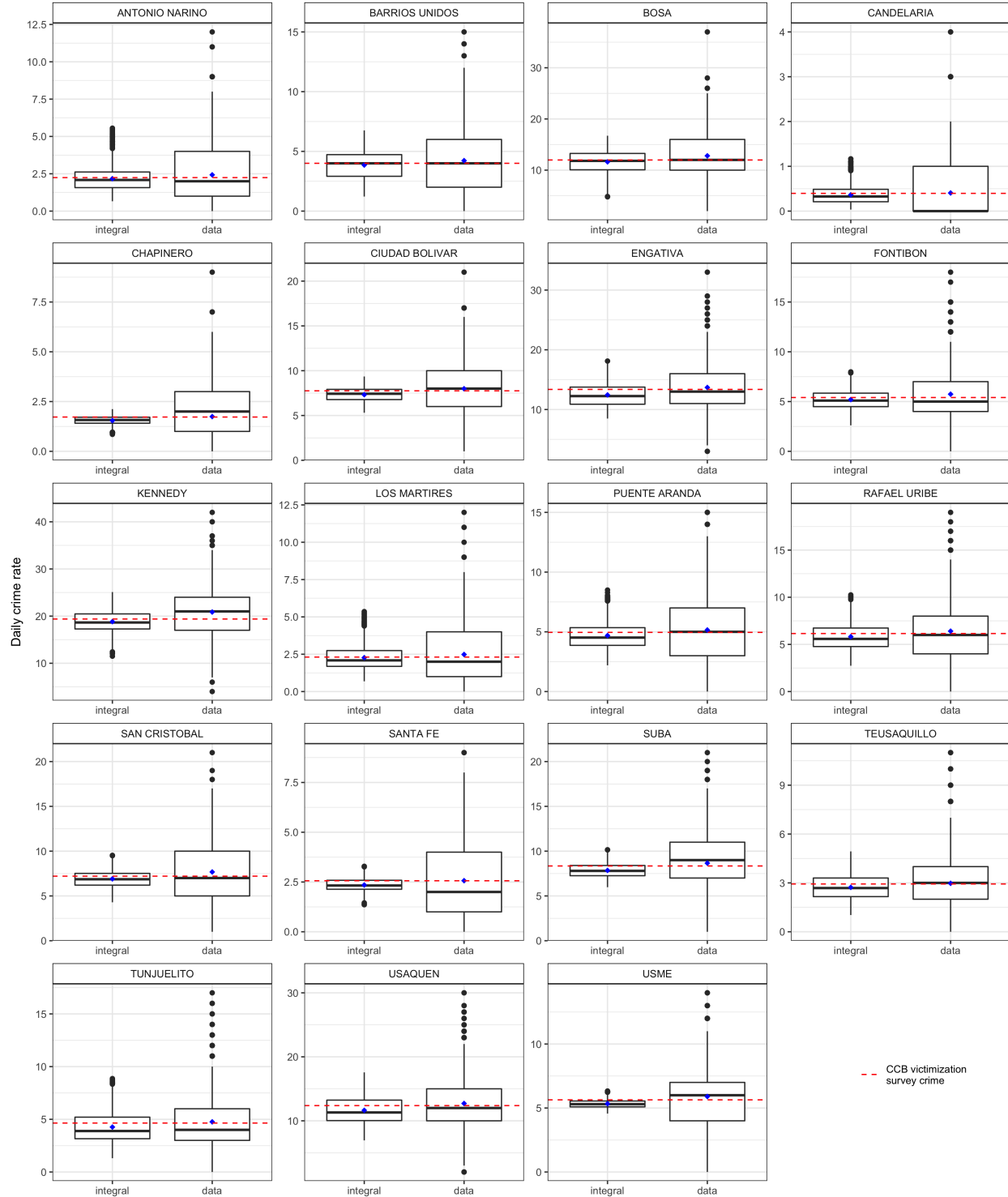


Figure 6: District-wise sanity check of synthetic crime data over all 2,190 time steps. The average daily counts of simulated data align well with the rates obtained by integration of the data generating thinned SEPP and the desired rates implied by the CCB victimization survey.

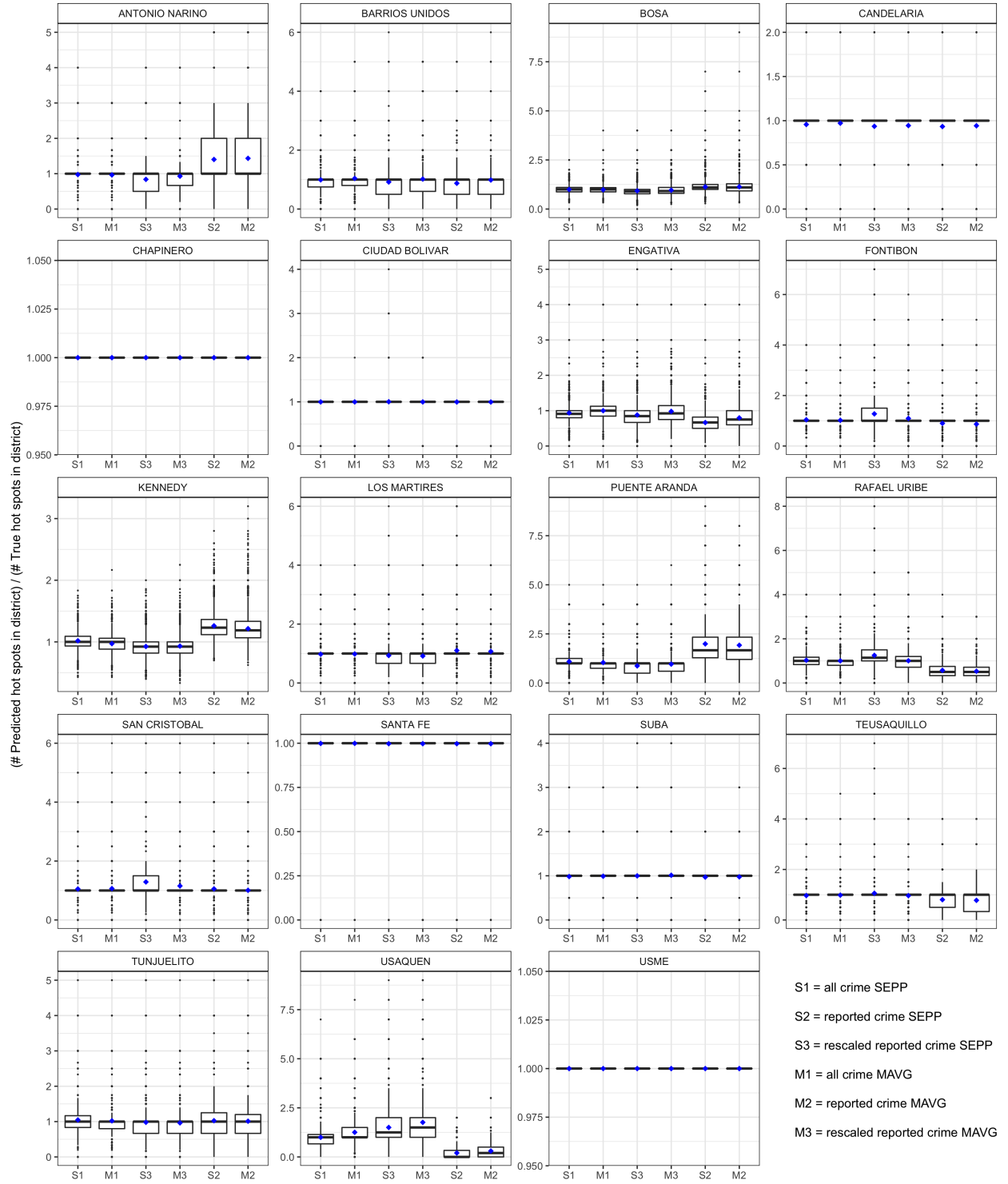


Figure 7: Relative number of predicted crime hot spots in Bogotá districts. Each data point represents a district-specific fraction at a given evaluation time step (189 days) in a given simulation run (50 runs). A total of 50 hot spots are selected at each time step. If both the true and predicted number of hot spots is zero, we set the relative count to one. Cases for which the number of predicted hot spots is non-zero but no true hot spots are available are excluded for visualization.

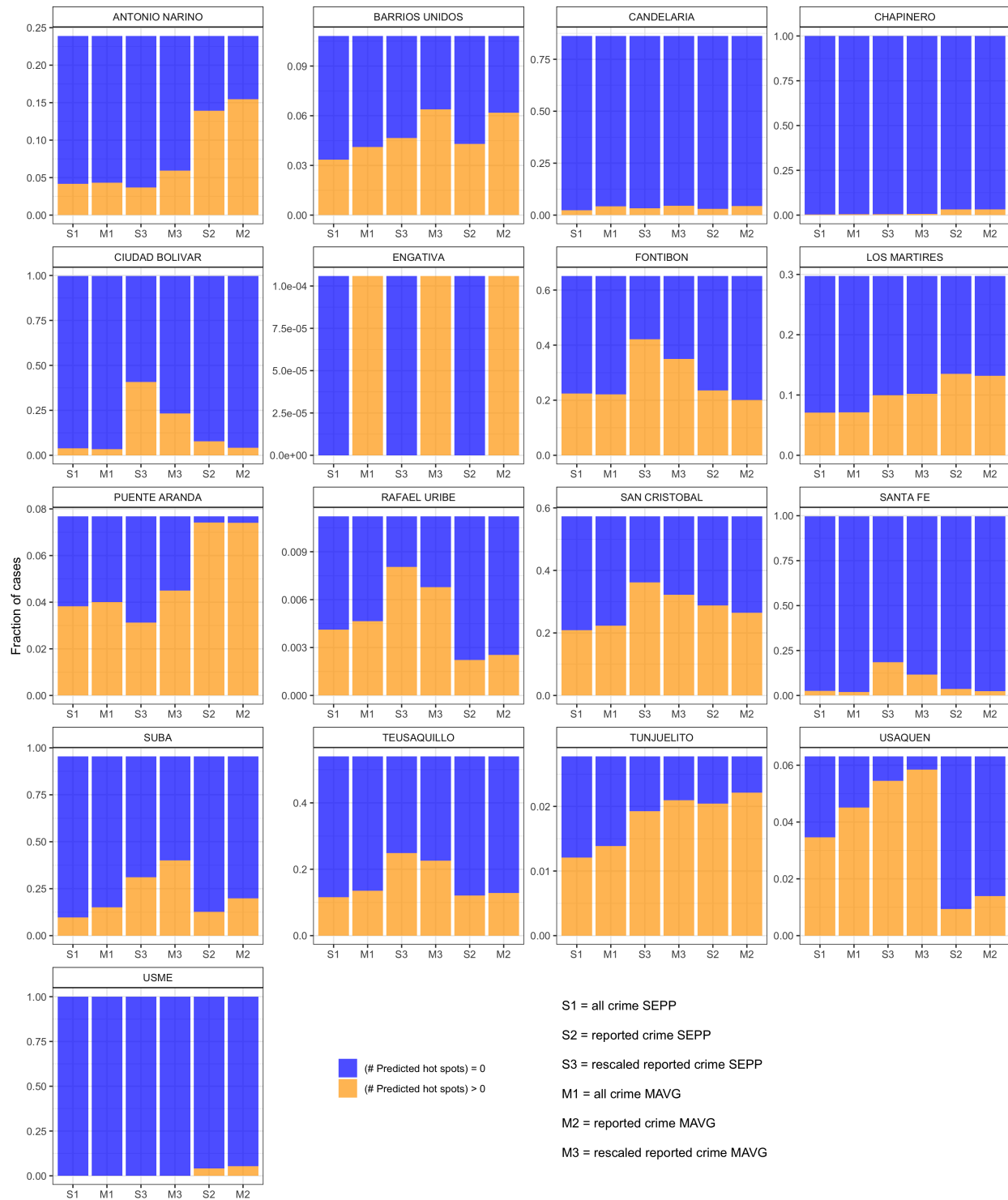


Figure 8: Fraction of prediction time steps with no true hot spots in district separated into instances with predicted and no predicted hot spots. Ratios are computed over all evaluation time steps (189 days) and all simulation runs (50 runs) with 50 hot spots selected at each step.

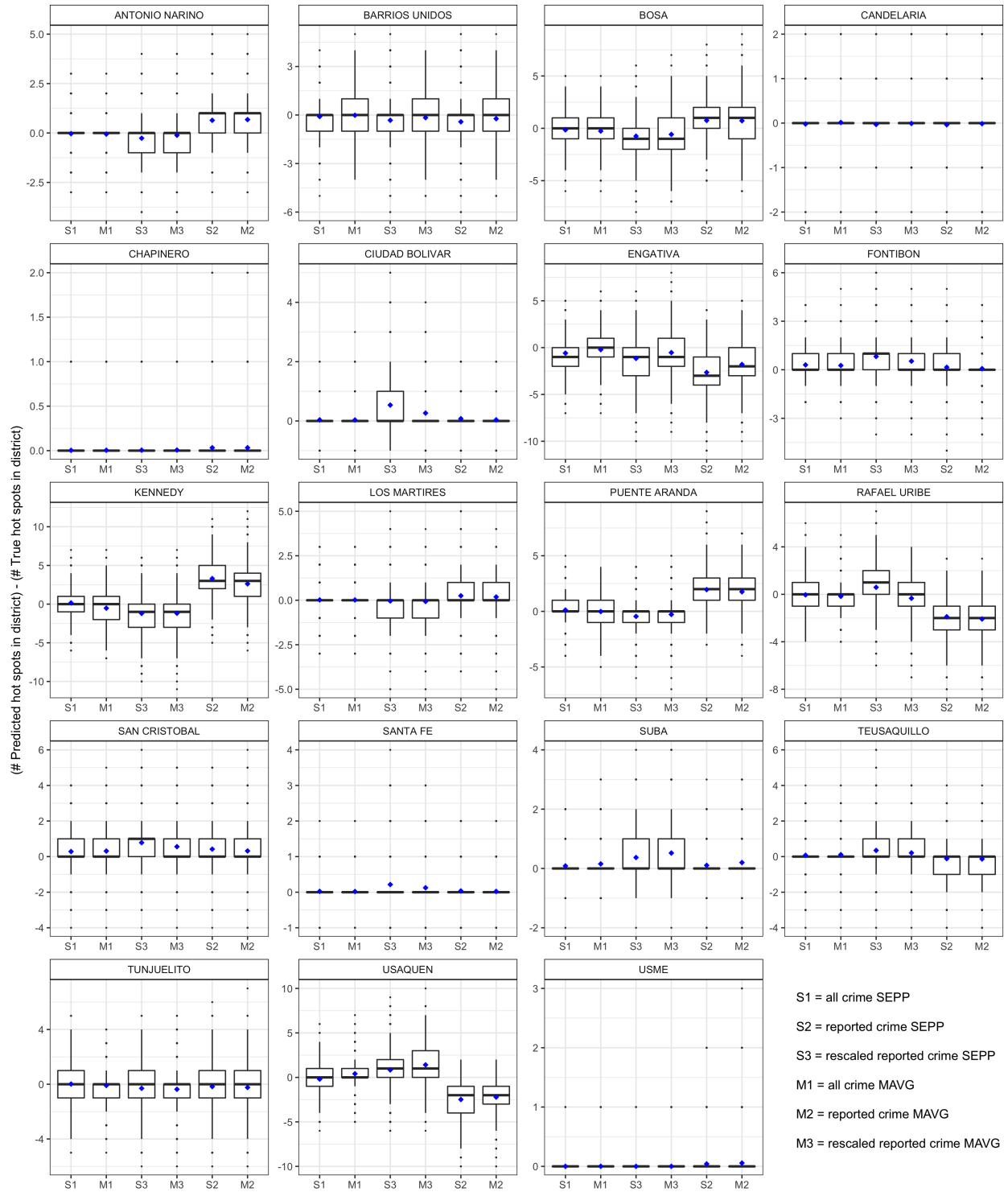


Figure 9: Absolute number of overpredicted hot spots over all evaluation time steps (189 days) and all simulation runs (50 runs) with 50 hot spots selected at each step.

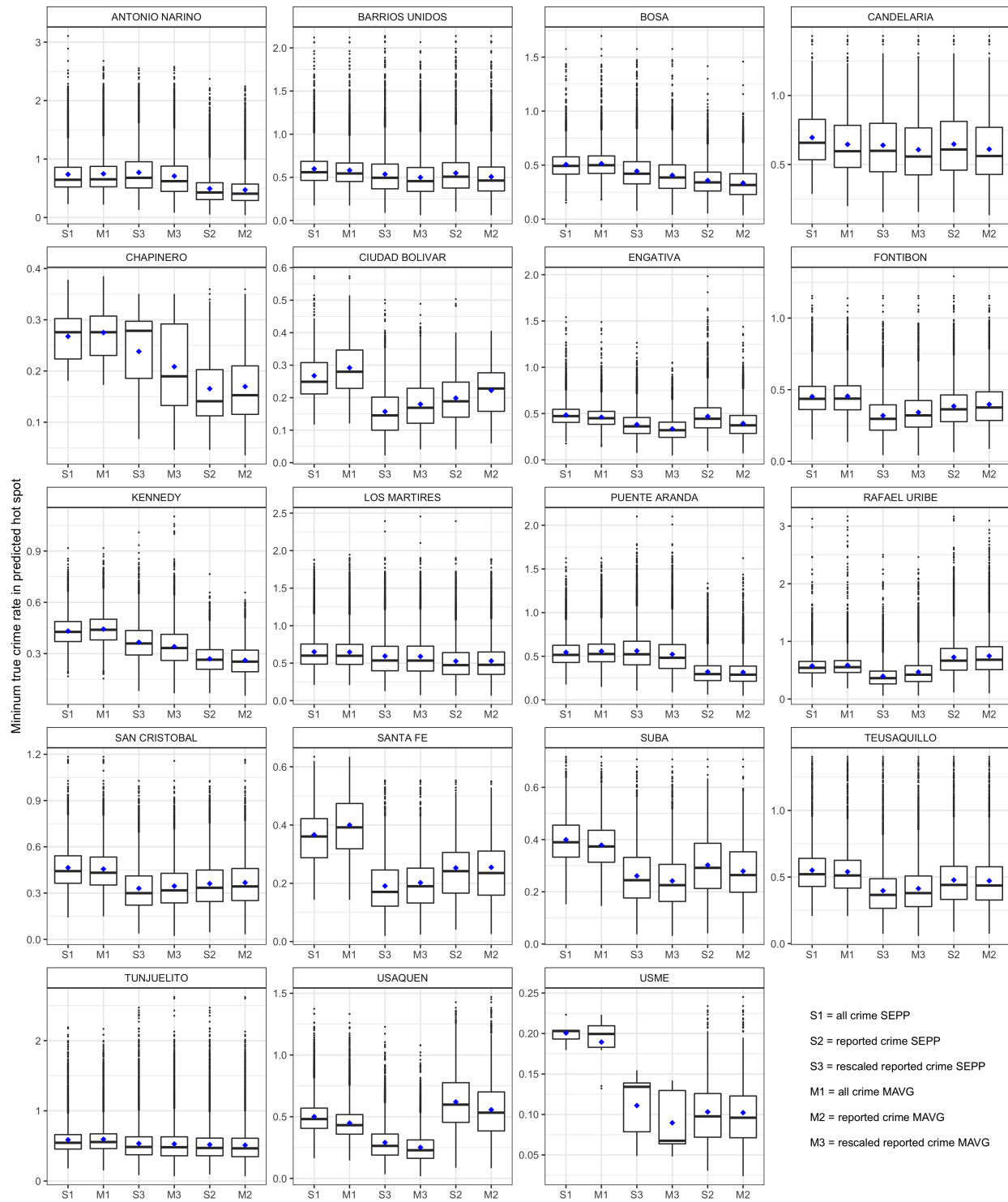


Figure 10: True crime thresholds for hot spot selection in Bogotá districts. Each point corresponds to an evaluation time step (189 days) and a simulation run (50 runs). A total of 50 hot spots is selected at each step, and cases in which no hot spot is predicted within the district are omitted for visualization.