Modelling underreported Spatio-temporal Crime Events

Álvaro J. Riascos Villegas[1,2], Jose Sebastian Ñungo[1,4], Lucas Gómez Tobón[1], Mateo Dulce Rubio[3], and Francisco Gómez[4]

[1] Quantil
[2] Universidad de Los Andes
[3] University of Carnegie Melon
[4] Universidad Nacional de Colombia **

March 16 2023

Abstract. Crime observations are one of the principal inputs used by governments for designing citizens' security strategies. However, crime measurements are obscured by underreporting biases, resulting in the so-called "dark figure of crime". Current approaches for estimating the "true" crime rate do not account for underreporting temporal crime dynamics. This work studies the possibility of recovering "true" crime incident rates over time using data from underreported crime observations and complementary crime-related measurements acquired online. For this, a novel underreporting model of spatiotemporal events based on the combinatorial multi-armed bandit framework was proposed. Through extensive simulations, the proposed methodology was validated for identifying the fundamental parameters of the proposed model: the "true" rates of incidence and underreporting of events. Once the proposed model was validated, crime data from a large city, Bogotá (Colombia), was used to estimate the "true" crime and underreporting rates. Our results suggest that this methodology could be used to rapidly estimate the underreporting rates of spatiotemporal events, which is a critical problem in public policy design.

1 Introduction

The observation of crime events constitutes a primary input used by government agencies for designing citizens' security strategies [1,2]. Different instruments aim to register these observations, including official crime record systems, citizen victimization surveys, and offender self-reports of crimes committed [2]. Nevertheless, the underreporting biases, introduced by unequal crime reporting across social groups and geographical areas [3,4,5,6], underrecording tendencies of official entities [7], which commonly prioritize the registration of high-impact offenses, methodological limitations in the selection of victims/offenders representative samples in the case of surveys [8,9], or simply, the lack of observers to report crime occurrences [8] highly impacts the number and type of offenses known through these mechanisms. Therefore, activities that, by some criteria, are considered crimes may occur without being registered by the systems devised to count them [10]. This phenomenon obscures our knowledge of crime dynamics and is known as the "dark figure of crime" [8].

The dark figure of crime has severe consequences: (1) it limits the deterrent capacity of the criminal justice system, (2) causes victims to become ineligible for public and private benefits, and (3) it affects insurance costs, among others [10]. In addition, in citizen security planning, which requires time-varying and trustable reports of crime incidences for resource assignation [2,11,12], dark crime figures may also result in misallocation of police resources, hampering short-time planning. Therefore, clarifying crime's dark figure over time constitutes a significant necessity in security planning.

Different strategies for estimating real crime incidences based on data describing crimes exist [13,14,15,16]. These strategies mainly rely on official crime reports and citizen victimization surveys and their covariates, including demographical and economic costs linked to the crime. Most approaches rely on victimization surveys, originally proposed to provide a ground truth of crime incidence. For instance, using these crime observations, Buil-Gil et al. provide long-term estimations of crime incidence for small areas [15]. Similarly, Akpinar et al. [17], and Buil-Git et al. [16] used surveys to simulate crime occurrences. Because of their design and intention, these victimization surveys provide a closer spatial picture of the criminal dynamic. However, despite the importance of these instruments for highlighting dark crime, they also result in noisy crime observations because of methodological limitations related to the sample design and its limited capability for capturing time-varying crime changes [8]. Alternatively, official crime registers, collected and available over time, provide indirect but

** Results of the project "Diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en Bogotá" funded by Colciencias with resources from the Sistema General de Regalías, BPIN 2016000100036. The opinions expressed are solely those of the authors.

time-updated views of the crime dynamics. Therefore, these observations have also been considered to unveil the dark figure of crime [13]. In particular, Gillespie adjusted the number of reported crimes of official statistics with the inverse probability of reporting a crime. This probability resulted from considering the costs of the crime and the benefit of informing it [13]. More recently, Chaudhuri et al. [14], and Moreira et al. [18] assumed crime as a linear function of demographic covariates and accounted for an additional term linked with inefficiency in the citizen's report, i.e., underreporting. Although these last approaches may provide a short-time estimation of actual crime incidences, they require exogenous covariates, which may also vary in time, limiting their capabilities for underreporting crime estimation over time. In summary, both data sources provide indirect information about the crime. Estimated crime rates provided by victimization surveys provide a closer spatial view of the crime dynamic, while estimations from official criminal records observe temporal crime dynamics. Nevertheless, current approaches still need to be expanded to offer time-varying estimates of crime incidence.

In recent years, several governmental agencies established alternative mechanisms to observe crime-related phenomena, including telephone citizen's reports [19], in-situ citizen's field reports [20], and mobile-based reports [21], among others. Currently, these observational mechanisms are deeply integrated into the citizen's security management information systems [2], registering large amounts of crime-related observations almost online. Although these observations still suffer from the dark figure of crime [8], together they may potentially provide valuable information for complementing official records. Recently, different works explored data integration/fusion approaches to provide more information about crime from multiple crime observation sources [22], particularly by spatially combining estimations of crime from different data sources, such as official crime reports, calls to the emergency line related to crime, and citizen's contraventions [22]. Nevertheless, these approaches are limited in uncovering crime underreporting over time (online) because there are no mechanisms for integrating partial observations arriving online into previous crime observation data.

The main objective of this work was to study the identification over time of the "true" unknown crime incidence rates based on official reports of crime incidents and the "true" underreporting rates based on complementary information, particularly crime-related data acquired gradually over time. In contrast with previous works aimed to describe underreporting in long-time scales by exploiting victimization surveys [15,17,16], official crime data [13,14,18], or combining multiple crime data sources [22], this work aims to integrate additional incremental evidence about crime once is available, allowing to gain knowledge about the crime phenomena gradually, instead of forcing to wait for final integration. The proposed online estimation relies on a new crime underreporting combinatorial multi-armed bandit model [23] aimed to elicit the "true" average incidence rates and estimate the underreporting rates for different spatial units over time. Importantly, the proposed approach maximizes the number of observed incidents and allows for limited budgets for acquiring complementary information [23]. We hypothesized that the online estimation of underreported crime data, considering partial complementary observations of the "true" crime, might help to estimate underreporting rates over time, further clarifying the dark figure of crime. Historical data of more than $35,000$ yearly crime incidents from two instruments for crime observation were used to study this hypothesis: 1) officially reported crimes and 2) telephone citizen reports on crimes in Bogotá (Colombia). The combination of these two data sources provided an approximation to the "true" crime incidents, which was aimed to be discovered by the proposed approach using officially reported crimes and partial observations of the citizens' reports acquired over time. In addition, the proposed strategy was also explored in the underreporting crime estimation evidenced in victimization surveys [17]. For this, Bogotá's victimization survey, which reports both victimization and underreporting rates, was used to simulate the "true" crime and underreported crime incidents, respectively. Then the proposed strategy identified the underreporting crime rates. The main contributions of this work are, first, the introduction of a new model aimed to provide estimates of the "true" crime incident rate over time, and second, the quantitative evaluation of the capacity of the model to unveil underreported crimes for different crime data sources. This work may have implications for designing cost-effective planning mechanisms for citizens' security planning.

The remainder of this paper is organized as follows. Section 2 introduces our formal underreporting model in a multi-armed bandit setting. To solve this problem we introduce and evaluate three well known algorithms using simulated data and we show how this strategy can be used to elicit the "true" crime incidence and underreporting rates in a large city. Section 4 explain our contribution and provides a general discussion and section 5 summarizes the results of this study.

2 Materials and methods

2.1 Model from Crime Underreporting over time

The problem that we want to solve is as follows. Suppose we repeatedly interact with an environment characterized by the realization of certain spatial events (i.e., crime events). Spatial events are modelled as count random variables $X_{i,t}$, where i indexes a spatial location and t indexes the round of the interaction. In each round we are given a chance to observe a finite non exhaustive number of locations (a subset S of all locations) and record the realization of these random variables (i.e., police can only visit a finite non exhaustive number of locations in the city). For those events that we did not observed in a particular round, we observe a filtered observation. That is, for each $i \notin S$, we observe a count random variable $\widetilde{X}_{i,t}$ (i.e., an underreported nuumber of crime events). Our main hypothesis is that the count random variable $\widetilde{X}_{i,t}$ is an underreported realization of the count variable $X_{i,t}$. To fix ideas the reader can think i denoting a location in a city, t a date, $X_{i,t}$ the number of crimes that occur at this location on a particular date, S as those locations that on date t are visited by police officers and $\widetilde{X}_{i,t}$ the reported crime incidents of those places not visited by the police on that particular date but still reported by, for example, some citizens. Our objective is, in a repeated interaction with this environment, to learn the true mean of the distributions of spatio-temporal events $X_{i,t}$ and filtered (or underreported) spatio-temporal events $\widetilde{X}_{i,t}$. To formally set up the problem to be solved, we use the same notation as in [24], and rewrite [25] and [26] algorithms in this notation.

A combinatorial multi-armed bandit (CMAB) problem with *underreporting* consists of M *base arms* associated with a set of random variables $X_{i,t}$ (i.e., crime events) and $\widetilde{X}_{i,t}$ (i.e., underreported crime events), with bounded support in $[0, 1]$, for $1 \leq i \leq M$ and $t \geq 1$. Variables $X_{i,t}$ indicate the random outcome of the i-th base arm in the t-th trial. Variables $\widetilde{X}_{i,t}$ indicate underreporting of events $X_{i,t}$. We assume that the set of variables $\{X_{i,t} \mid t \geq 1\}$ associated with base arm i, are independent and identically distributed over time t according to some distribution with unknown expectation μ_i. We also assume that the set of variables $\{\widetilde{X}_{i,t} \mid t \geq 1\}$ associated with underreporting of base arm i over time t, are independent and identically distributed according to some distribution with unknown parameters q_i. Note that $X_{i,t}$ and $\widetilde{X}_{i,t}$ may be correlated.

Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ be the vector of expectations of all base arms, and $\boldsymbol{q} = (q_1, q_2, \ldots, q_M)$ be the vector of the parameters of interest of the underreported base arms. Note that q, in our model, is not the mean of the vector $\widetilde{X}_t$. By allowing random variables of different base arms to be dependent we rationalize the common framework in which the random variables $\{X_{i,t} \mid i = 1, ...M\}$ represent the spatial events at M different locations. We also allow for the dependence of base arms and underreporting of base arms in the same period and across arms, as noted earlier.

In every period a decison maker or social planner (i.e., police planning department) must select a *super arm* (i.e., set of locations to be visited by police officers), which is a subset of the set of base arms. Let $\mathcal{S}$ denote the set of all possible super arms that can be played in a CMAB problem instance. For example, $\mathcal{S}$ can be the set of all subsets of base arms containing m base arms (this is our case). In each round, one of the super arms $S \in \mathcal{S}$ is selected and played, and every base arm $i \in S$ is triggered and played as a result (i.e., this means that the realization of crime, $X_{i,t}$ is observed). We assume that for arms outside super arm S, we observe underreported realizations of the base arms. More precisely, we assume that for $i \notin S$, we observe the random variables $\{\widetilde{X}_{i,t} \mid X_{i,t}\}$ (i.e., the realization of underreporting conditional to the true crime realization). That is, arms not in the super arm selected in some rounds, are fired but not observed and we only observe the random variable $\widetilde{X}_{i,t}$ conditional to $X_{i,t}$. In our simulation study and applications we assume that variables $X_{i,t}$ distribute as a Binomial random variable $B(n, \mu_i)$, and $\widetilde{X}_{i,t}$ conditional to $\mid X_{i,t}$, which we denote as $\{\widetilde{X}_{i,t} \mid X_{i,t}\}$, are distributed as a Binomial random variable with parameters $X_{i,t}$ and q_i, denoted by $B(X_{i,t}, q_i)$.

For each arm $i \in \{1, ..., M\}$, where M is the total number of arms, let $T_i(t)$ denote the number of times arm i has been triggered after the first t rounds in which t super arms have been played. If arm $i \in S$ is not triggered in round t when super arm S is played, then $T_{i,t} = T_{i,t-1}$. Analogously, let $\widetilde{T}_i(t)$ denote the number of times arm i has been underreported after the first t rounds in which t super arms have been played.

The final reward of a round depends on the outcomes of all triggered base arms in the super arm. Let $R_t(S)$ be a non-negative random variable denoting the reward of round t when super arm S is played. We assume that reward $R_t(S)$ has the form $R_t(S) = \sum_{i \in S} X_{i,T_{i,t}}$. In other words, our goal was to maximize the number of observed incidents. Underreported events do not contribute to the reward. The expected value of $R_t(S)$, $E[R_t(S)]$, is a function of S and the parameters μ_i of the arms in super arm S.

An algorithm for this problem is the selection of a super arm for each round t such that it maximizes the expected round t reward: $E[R_t(S)] = \sum_{i \in S} \mu_i$, for an unknown $\boldsymbol{\mu}$. To use the algorithms proposed by [24], [25] and [26], we must have access to a computational oracle that takes an expectation vector μ as input, and compute the optimal or near-optimal super arm S. In our case, the computational oracle is reduced to a sorting problem for which there are fast algorithms [27]

2.2 Algorithms

For completeness and to illustrate how we apply the algorithms [24], [25] and [26] to our underreporting problem, we provide the pseudo-algorithms that we implemented.

```
1: For each arm $i$, maintain: (1) variable $T_i$ as the total number of times arm $i$ is played so far; (2) variable $\widetilde{T}_i$ as the total
   number of times arm $i$ has been underreported (initially both 0); (3) variables $\hat{\mu}_i$, $\hat{q}_i$ as the mean of all outcomes $X_{i,t}$
   for $1 \leq i \leq M$ that have been observed up to round $t$ and the best estimate of the parameters characterizing $\widetilde{X}_{i,t}$,
   $1 \leq i \leq M$, which have been observed up to round $t$ (initially both 1).
2: $t \leftarrow 0$.
3: while true do
4:     $t \leftarrow t + 1$.
5:     For each arm $i$, set $\bar{\mu}_i = \min\left\{\hat{\mu}_i + \sqrt{\frac{3\ln t}{2T_i}}, 1\right\}$.
6:     $S = \text{Oracle}(\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_m)$.
7:     Play $S$. Observe the outcomes of base arms $i \in S$, and update all $T_i$'s and $\hat{\mu}_i$'s.
8:     For $i \notin S$, observe $\widetilde{X}_{i,t}$ conditional on the outcomes of base arm $i$ in step 7. Update $\hat{q}_i$:

           $\hat{q}_i \leftarrow \frac{\text{Empirical mean of underreporting so far observed}}{n\hat{\mu}_i}$        (1)

9: end while
```

With this notation we write the **Learning with Linear Rewards (LLR) algorithm** of [25] as follows. Replace Step 5 in 2.2 with:

$$\bar{\mu} = \hat{\mu}_i + \sqrt{\frac{(M+1)\ln t}{T_i}} \tag{2}$$

Finally, we consider the **Upper Confidence Bound, version 1 (UCB1) algorithm**, of [26] which ignores the potential association between arms at any moment in time. This is a major handicap in its performance as has been pointed out in [25]. Replace Step 5 in 2.2 as follows. Rather than choosing a super arm every time period, the algorithm updates only the arm that maximizes:

$$\hat{\mu}_i + \sqrt{2\frac{\ln t}{T_i}} \tag{3}$$

We now turn to the validation of our proposed methodology to estimate underreporting.

2.3 Model Validation

To validate our strategy to elicit the "true" incidence rate, underreporting parameters, and maximize the discovered events simultaneously, we extensively study the model with binomial arms distributions and binomial conditional underreporting. We report the results of the four experiments. In all of our validation simulations and in our two applications, we assume that the size of the super arms is at most 10% of the number of arms. This is because in our applications to crime underreporting, the number of arms that can be efficiently monitored and checked by police officers are at most 10% of the area of the city.

In the first experiment we used 12 arms and considered the superarms of at most two arms as shown in Table 1. The true mean $\boldsymbol{\mu}$ and parameters $\boldsymbol{q}$ for the first set of simulations are listed in Table 2. Figures 1 and 2 show how the

CUCB algorithm converges to the true values of μ and q over time for the different arms, which are represented by dashed horizontal lines in both figures. The graphs for UCB1 and LLR are similar and are not shown for brevity.

M	m	T_{max}	n
12	2	1000	1000

Table 1: Global parameters. M is the number of arms, m the size of the super arm, T_{max} the maximum number rounds played and n is the parameter of the Binomial distribution.

Arm	$\boldsymbol{\mu}$	$\boldsymbol{b}$
0	0.070	0.244759
1	0.096	0.694755
2	0.021	0.593902
3	0.087	0.631792
4	0.061	0.440257
5	0.090	0.083726
6	0.051	0.712330
7	0.077	0.427863
8	0.036	0.297780
9	0.050	0.492085
10	0.029	0.740296
11	0.087	0.357729

Table 2: True values of $\boldsymbol{\mu}$ and $\boldsymbol{q}$ for each arm in simulations.

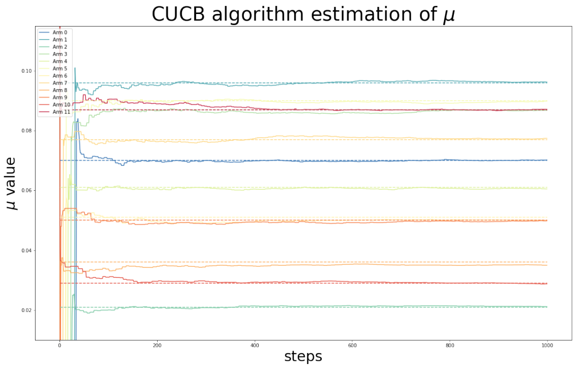

Fig. 1: CUCB Convergence to true arms mean.

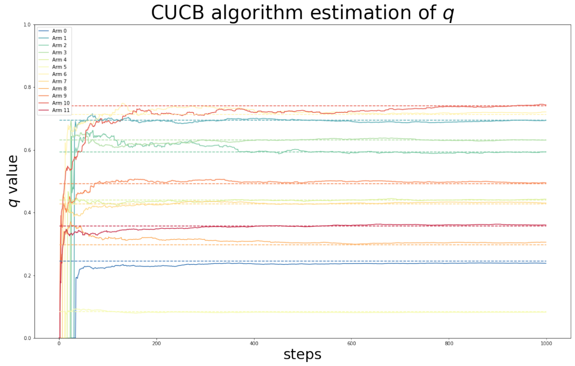

Fig. 2: CUCB Convergence to true arms underreporting parameters.

Figure 3 shows the Euclidean distance between the estimated $\widehat{\mu}_t$ and true value of μ in each round t of the algorithms. Additionally, figure 4 shows the number of times each algorithm triggered each arm. Note that, by construction, UCB1 visits only one arm per round while the other two algorithms visit all arms in the superarm in each round. Hence, after $1,000$ rounds, the other algorithms visited mores arms. Note also that there were minor differences in the number of times each arm is visited by CUCB and LLR algorithms. Finally, given that this was a small simulation experiment, there is no major computational burden. The results clearly show that all algorithms in this small experiment can recover the true means of all arms and the true parameters of the underreporting distributions. Figure 3 shows how the CUCB algorithm (green line) outperforms the other two algorithms in terms of convergence speed.

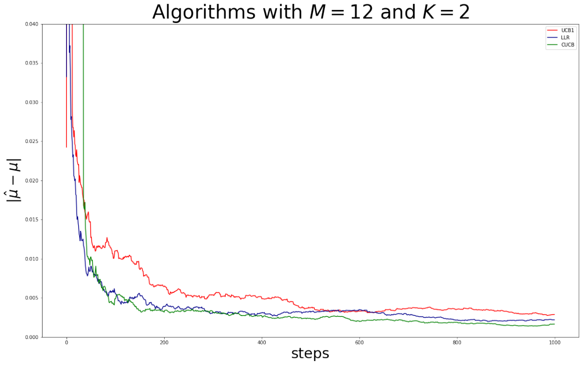

Fig. 3: Convergence error of true arms mean for each algorithm. The error is measured as the Euclidean distance between the true mean vector and estimated mean vector per round.

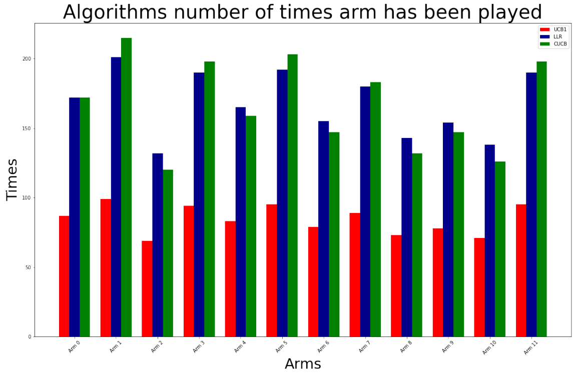

Fig. 4: Number of visits (i.e., fired arms) of algorithms to each arm.

The next experiment solved an increasingly challenging task. In each arm, we drew random true mean incidence rates, $\boldsymbol{\mu}$ and parameters, $\boldsymbol{q}$ for each arm. Figure 5 shows the case of $1,000$ arms and at most 100 super arms. Since UCB1's performance is highly surpassed by the CUCB and LLR algorithms, we do not report the outcome of this algorithm in the next two exercises. Figures 6 and figure 7 report the cases of $10,000$ and $50,000$ arms with at most $1,000$ and $5,000$ super arms, respectively. These figures show on the Euclidean distance between the true mean and the estimated values in each round the vertical axis. In addition, Table 3 reports the time required for each algorithm to complete $1,000$ rounds (we used a portable PC, with Intel i7-16 GB of RAM).

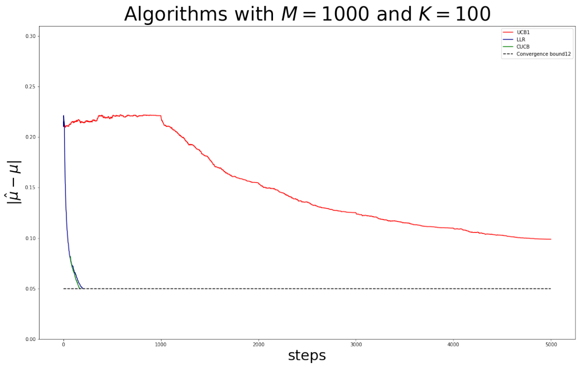

Fig. 5: Convergence error of true arms mean for each algorithm. The error is measured as the Euclidean distance between the true mean vector and estimated mean vector per round.

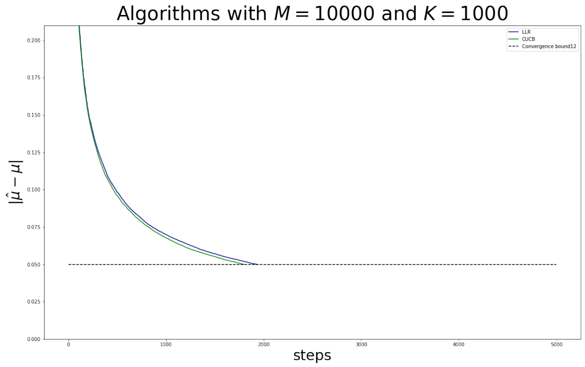

Fig. 6: Convergence error of true arms mean for each algorithm. The error is measured as the Euclidean distance between the true mean vector and estimated mean vector per round.

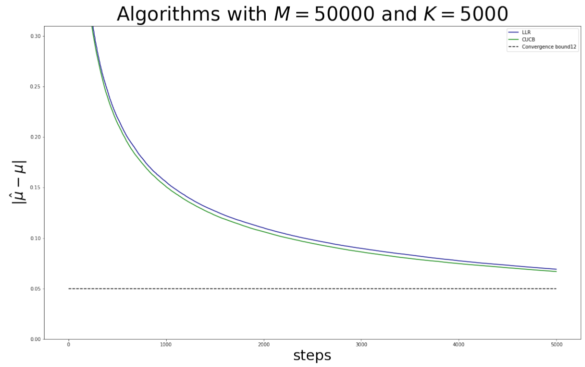

Fig. 7: Convergence error of true arms mean for each algorithm. The error is measured as the Euclidean distance between the true mean vector and estimated mean vector per round.

The next table quantifies the time to completion of $1,000$ rounds for each algorithm. With many arms, CUCB and LLR have a similar performance, but after $1,000$ rounds UCB1 fails to converge.

	Case 1	Case 2	Case 3
UCB1	3 sec	38 sec	3 min 31 sec
LLR	4 sec	51 sec	4 min 15 sec
CUCB	4 sec	53 sec	4 min 12 sec

Table 3: Time to completion of $1,000$ rounds of each of the three algorithms. Case 1: $M = 1,000$ and $K = 100$. Case 2: $M = 10,000$ and $K = 1,000$. Case 3: $M = 50,000$ and $K = 5,000$. Sec is seconds, min is minutes.

2.4 Estimating crime underreporting

We provide two applications of our underreporting algorithm, showing that it is an effective way of estimating the true mean of crime incidents and underreporting in Bogotá, the capital city of Colombia. First we discuss how we built the two data sets for our applications. The first was the real crime and underreporting dataset and the second was, the simulated dataset. We divided the city into 1 km^2 cells. This resulted in 500 cells with at least one crime during year 2018. These cells were the focus of our study. In both applications we assumed that the size of the superarms was at most 10% of the number of arms. This is because the number of arms that can be efficiently monitored and spot checked by police officers is at most 10% of the area of the city's area.[1] Figure 8 shows the 19 jurisdictions in which the city is divided and our grid of 1 km^2 cells that we used as arms.

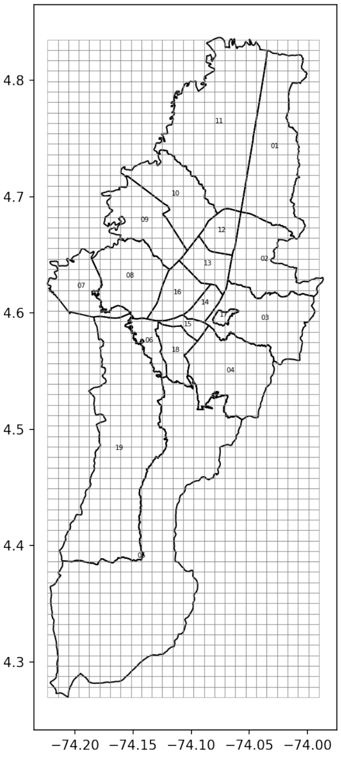

Fig. 8: Bogotá, capital city of Colombia. Figure shows the 19 jurisdictions in which the city is divided and our grid of 1 km^2 cells.

2.5 Crime data

Our dataset contained daily time-stamped information on the spatial location of each criminal event reported in Bogotá from January 2018 to December 2019. The source was the Criminal, Contraventional and Operating Information System (SIEDCO). The dataset was assembled by the Colombian National Police and was provided by the Bogotá Security Office. Although SIEDCO is the official crime source in the city, there is evidence of substantial underreporting as can be deduced from two different sources. The first source is citizens crime reports to the security and emergency call center NUSE (*Número Único de Seguridad y Emergencias* in Spanish). By comparing the different reports in SIEDCO and NUSE, it can be observed that many reports in NUSE do not appear in SIEDCO and viceversa.

Our main approach to capturing the totality of violent crimes consists of combining both data sets. To avoid double counting of crimes, we eliminated all crimes a for which there was another crime b that belonged to the same crime category, occurred at a distance of less than 500 meters and both where reported within a period of less than 8 hours. Figure 9 shows the total number of crimes reported by each source, SIEDCO and NUSE, and the Total number of crimes which is the sum of SIEDCO plus NUSE eliminating double counting as explained previously. This Total series (the green line in Figure 9) is called the *real* dataset.

[1] According to official statistics [28], between years 2012 − 2015, all homicides and 25% of crime in the city took place in 2% of street segments.

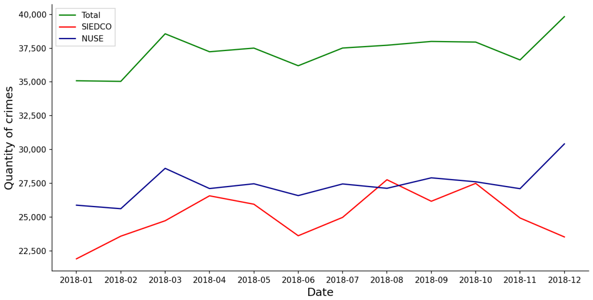

Fig. 9: Crimes by source of information: SIEDCO is the official source of information of crimes in Bogotá. NUSE is the security and emergency call center of the city. Total is the sum of both sources eliminating double counting as explained in the main body of the text.

The second source is Bogotá's City Chamber of Commerce (*Cámara de Comercio de Bogotá*, in Spanish) victimization and reporting survey [29]. This is a biannual crime perception and victimization survey that asks individuals if they had been victims of some crime in the last six months and in case they did, if they had reported this crime. The survey is representative of the whole city, stratified at the level of 19 jurisdictions of Bogotá, and interviews $6,500$ people. In 2021, the survey reported an average victimization rate of 17% and, among those, only 49% said they had reported the event to the police. Using this survey, we simulated a second dataset using standard crime models. In detail, using the historical data constructed in the first dataset, which we call the real crime dataset, we fit a Poisson model on each arm (cell) and we used this Poisson model to simulate crime in each round of our algorithms. This is our simulation model of crime. To estimate underreporting, we used the most recent results of the victimization and reporting survey [29] for which we disaggregated results at the level of jurisdiction (taken from [17]). We mapped each of our cells to each jurisdiction and estimate underreporting using the data shown in Table 4

ID	District	Pop.	Vict. Rate	Rep. Rate
15	Antonio Nariño	109,176	15%	33%
12	Barrios Unidos	243,465	12%	22%
07	Bosa	673,077	13%	26%
17	Candelaria	24,088	12%	22%
02	Chapinero	139,701	9%	28%
19	Ciudad Bolívar	707,569	8%	17%
10	Engativá	88,708	11%	20%
09	Fontibón	394,648	10%	19%
08	Kennedy	1,088,443	13%	28%
14	Los Mártires	99,119	17%	25%
16	Puente Aranda	258,287	14%	32%
18	Rafael Uribe Uribe	374,246	12%	15%
04	San Cristóbal	404,697	13%	21%
03	Santa Fe	110,048	17%	17%
11	Suba	1,218,513	5%	19%
13	Teusaquillo	1,53,025	14%	19%
06	Tunjuelito	19,943	17%	23%
01	Usaquén	501,999	18%	13%
05	Usme	457,302	9%	33%

Table 4: Results of Bogotá's City Chamber of Commerce, Cámara de Comercio de Bogotá, victimization and reporting survey 2014. We use reported rates form each jurisdiction to estimate underreporting simulated from our Poisson model. The table also reports the population of each jurisdiction and victimization rate.

3 Results

Consider our first application in which we have done our best to estimate the real crime rate and underreporting in each cell of Bogotá in 2018 (what we call the real dataset). Below we present the results of running the three algorithms on these datasets. Figures 10 and 11 show the convergence of the vector of incidence rates $\boldsymbol{\mu}$ and the vector of parameters $\boldsymbol{q}$ respectively, for each algorithm. In each case the reference vectors are the mean of all crimes in each cell over the period and the mean of the vector of estimated underreporting parameters in each cell over the entire period. In either case, the error is measured as the Euclidean distance between two high dimensional vectors with 415 components. Therefore, a reported error of for example, 0.4 in figure 10, or 2 in figure 11 represents means errors per component of 0.96^{-3} and 4.8^{-3} respectively. In addition, these parameters are unknown in this real-world application.

As expected, the estimated parameters is not perfect because the real dataset may not satisfy some of our working hypothesis. In particular, the number of crimes reported per cell i as a proportion of the total number of crimes in the cell, $\frac{\widetilde{X}_i}{X_i}$, may not be a stationary distribution. In addition, the distribution of $\widetilde{X}_i \mid X_i$ may not be a binomial random variable, $B(X_i, q_i)$.[2]

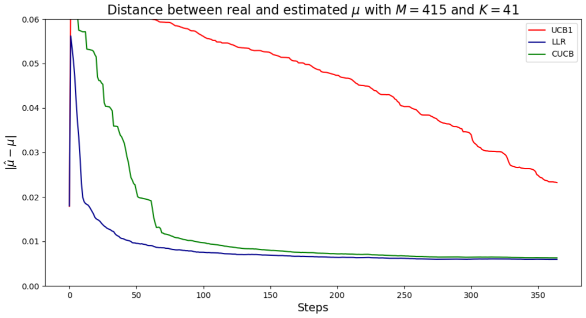

Fig. 10: Convergence of the vector of incidence rates $\boldsymbol{\mu}$ to the mean of all crimes per cell across time. The error is measured as the Euclidean distance between vectors with 415 components.

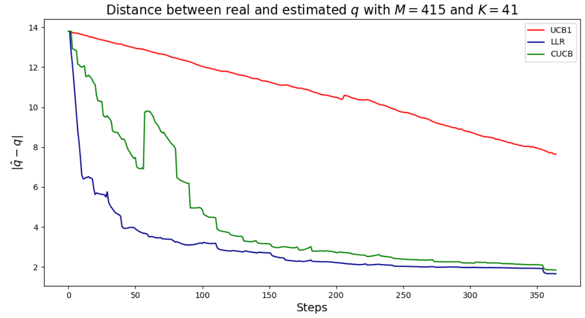

Fig. 11: Convergence of estimated vector $\boldsymbol{q}$ per round to the empirical mean of the underreporting rate for the whole sample. The error is measured as the Euclidean distance between vectors with 415 components.

We further explore the nature of this convergence. Figure 12 shows a histogram of the error between the empirical mean of the ratio $\frac{\widetilde{X}_i}{X_i}$ and that implied by our model in the last round per cell (error in absolute value). As can be seen from Figure 12, the CUCB and LLR algorithms converge in almost all cells with an error smaller than 0.2, after $1,000$ rounds.

[2] Since many cells report zero crime, care must be taken to empirically estimate these ratios. To do these, we estimate the mean $\widetilde{X}_i \mid X_i$ whenever $X_i \neq 0$, otherwise we set the ratio to zero. We compare these statistics to those implied by the model: $q_i(1-(1-\mu_i)^n)$

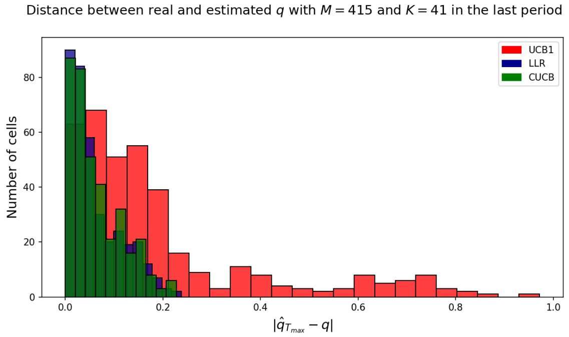

Fig. 12: Histogram of convergence of estimated error of $\boldsymbol{q}$ in the last round to the empirical mean of the underreporting rate for the whole sample. Absolute values reported.

The next two Figures 13 and 14 show the model implications for aggregate crime and underreporting (compare to Figure 9). Specifically, Figure 13 shows the aggregate expected crime rate over all cells in each round, $n(\mu_1 + ... + \mu_{415})$. Note that the CUCB and LLR algorithms converge approximately to the most recent observation of Total in Figure 9. In addition, Figure 14 shows the expected value of total underreporting in each round: $n(\mu_1 q_1 + ... + \mu_{415} q_{415})$. The CUCB and LLR algorithms converged approximately to the most recent NUSE observations. However, as noted before, the convergence of the vector of parameters $\boldsymbol{q}$ is not equally good across all cells, and hence there is an aggregate discrepancy.

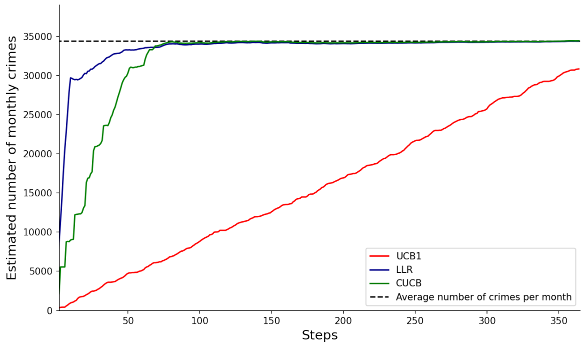

Fig. 13: Convergence of the estimated total number of crimes to the observed number of crimes in the city.

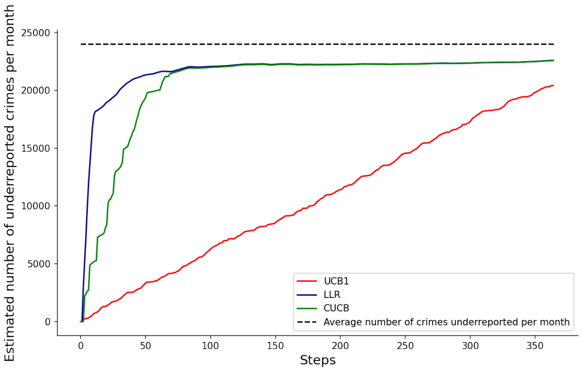

Fig. 14: Convergence of the estimated total (aggregate across cells) of the number of underreported crimes implied by the model.

A more illustrative presentation of the results is shown in Figure 15. We only report the results for the CUCB algorithm. The first column and row of the panel in Figure 15 show a heat map of the estimated real crime incident rates in the city and how the CUCB algorithm discovered these crime incidents. The first, second and third rows (left column), show the heat maps of the estimated crime incidence rates after 25 iterations and 100 iterations of CUCB, respectively. The first row, second column, show real underreporting as measured by NUSE dataset. The second and third rows (second column), show the heat map of the estimated underreporting crime after 25 and 100 iterations of CUCB, respectively.

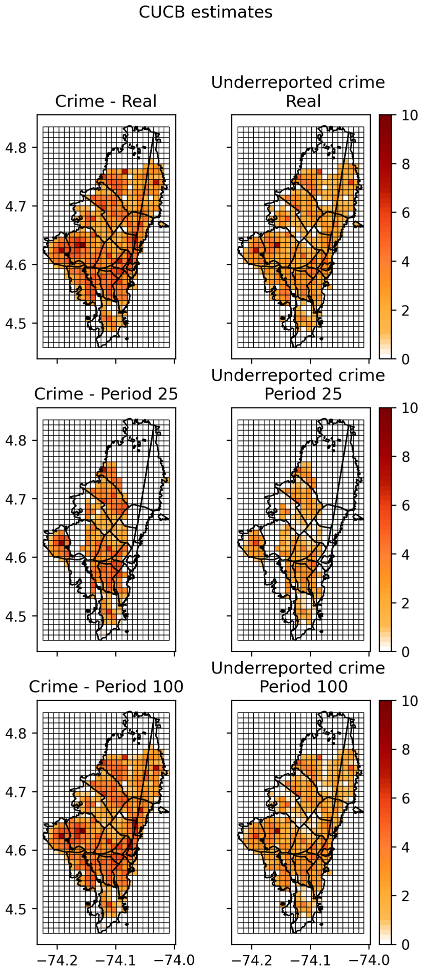

Fig. 15: Heat map illustrating the convergence, using the CUCB algorithm, of the estimated crime and underreporting of events in the city, to the real values. The first column, second and third rows show the heat maps of the estimated crime incidence rates after 25 and 100 iterations, respectively. The second column, first row shows real underreporting as measured by NUSE dataset. The second column, second and third rows show the heat maps of the estimated underreporting crime after 25 iterations and 100 iterations, respectively.

In our second application, we estimated a standard crime model. Using historical data, we fitted a Poisson distribution to each cell and used the Bogotá's City Chamber of Commerce 2014 victimization and reporting survey to estimate underreporting in each cell (note that the underreporting rate is the same for all cells that are mapped to the same jurisdiction). Figure 16 shows the convergence of the vector of the true incidence rates $\boldsymbol{\mu}$ to the true values. The error was measured as the Euclidean distance between the vectors. Note that the UCB1 algorithm failed to converge after $1,000$ rounds.

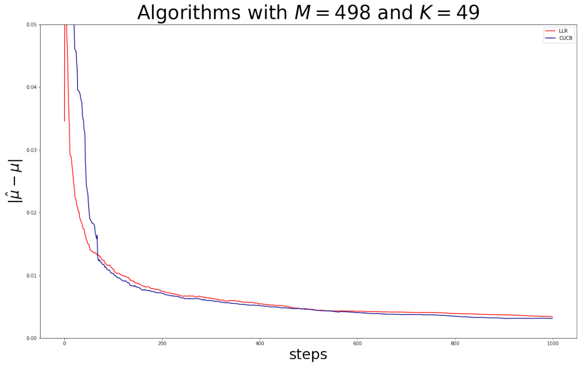

Fig. 16: Results for second application simulating data with standard crime Poisson model. Figure shows the convergence of the vector true incidence rates μ to the true values. Error measured as Euclidean distance between vectors.

Finally, in Figure 17 we report the convergence of the vector of parameters $\boldsymbol{q}$ in the underreporting distribution. The error was measured as the Euclidean distance to the true parameters. Algorithm UCB1 is not shown because it was considerably outperformed by the other two algorithms.

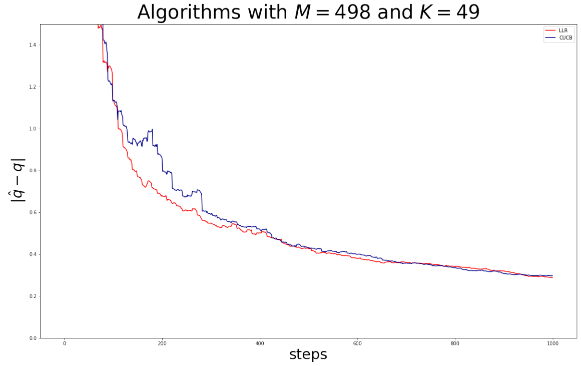

Fig. 17: Results for second application simulating data with a standard crime Poisson model. Figure shows the convergence of the vector parameters $\boldsymbol{q}$ to the true values. Error measured as the Euclidean distance between vectors. UCB1 not reported because it is outperformed by the other two algorithms.

4 Discussion

This paper studies the "true" crime incident rates estimated over time using data from underreported crime observations and complementary crime-related measurements acquired incrementally. Two crime-related observational mechanisms exhibiting underreporting, namely, official crime registers and victimization surveys, were studied to estimate their underreporting and "true" incidence rates, unveiling their dark crime figures on time. In contrast to previous approaches for estimating crime underreporting, which mainly focused on the long-term adjustment of underreporting rates, this study describes for the first time the online estimation of the spatial crime rates by sequentially integrating time-varying complementary crime-related observations.

The underreporting of spatio-temporal events is ubiquitous in many social problems [30], and particularly for crime characterization [31,8]. All systems that describe crime dynamics, including official crime registers and citizen surveys, provide informative but limited observations of crime occurrences [31]. Concerns about the dark crime figure have been present since the first initiatives to study crime quantitatively [32] until the modern artificial intelligence strategies for crime prediction [17]. Underreporting is present not only in the spatial dimension but also in the temporal dimension [33]. Previous works on crime underreporting focused on constructing average crime rate estimations for long-term windows [13,15,16,14,22]. Most of these works aim to quantify the spatial underreporting of official crime registers, assuming the citizen's surveys on victimization as the crime ground truth. This work also provides similar spatial estimations of underreporting but accounts for the temporal dimension, providing spatiotemporal estimates of the "true" crime incidences. Previous work on crime prediction also considers a time-based crime characterization but does not account for the underreporting phenomena [1,34,35]. Our results show that combining complementary crime-related data sources over time may help gradually illuminate the dark figure of crime, as illustrate Figure 15.

The proposed approach estimates the "true" crime incidence over time using synthetically constructed ground truths of crime. It is worth noting that constructing "real" crime reference databases is a challenging problem, mainly because it is almost impossible to directly measure this phenomena [8]. Our experimental configuration relies on two settings that explore the proposed approach capabilities to discover underreporting on two simulated ground truth crime databases. The first setting aimed to investigate the capacity of the proposed approach to complement official crime reports, with information supplied frequently by citizens'. Previous works suggest that official crime reports are biased by underreporting [3,4,5,6]. This limitation may result from unequal reporting rates across the population and space. The explored setting aimed to cover, at least partially, this reporting gap by considering, in addition, complementary reports, particularly citizens' telephone calls crime-related reports [19,36]. Therefore, a first ground truth crime database was constructed by combining these two datasets. Our results suggest that the proposed method provides good-quality ground truth crime estimations early in time (see Figure 10 and Figure 17), even for estimations of the total number of crimes (see Figure 13 and Figure 14). Nevertheless, the underreporting described for this setting should be cautiously interpreted because of the potential contamination of false crime reports, naturally observed in telephone reports of crime incidents [19,36], which may result in over/sub estimation of the underreporting and "real" crime rates. The second setting aimed to overcome this limitation by considering a citizen victimization survey, which also accounts explicitly for underreporting [29]. Crime occurrences with underreporting were simulated in time and compared with ground truth reports of crime obtained from the same survey. Our results show that again in this alternative setting, the proposed method resulted in fast, good-quality estimations of crime underreporting, see Figure 16 and Figure 17. However, these results should also be carefully interpreted because surveys provide long-term average descriptions of crime events, and the simulation process considered does not account for particular crime dynamics in time.

This work introduces a novel underreporting model of spatiotemporal events. The model relies on the multi-armed bandit framework, which provides efficient algorithms and convergence guarantees for online learning of the mean of the true arms distributions. Three well-known multi-armed bandit algorithms [24], [25], [26] were explored for the online estimation task. Importantly, the capacity of the model was extensively studied in several controlled simulated scenarios, given highly competitive results, as shown in figures 1, 2 and table 3. Results in the crime underreporting estimation suggest the CUCB algorithm's effectiveness in identifying the proposed model's fundamental parameters. Furthermore, these results indicate that the combinatorial nature of CUCB may help to accelerate the crime underreporting discovery process, likely improving its exploratory capacity [24], as shown in the comparison between the algorithms in Figures 12.

Several studies have pointed out the potential pitfalls of using discovered crime incidents, biased or underreported, to train machine learning models that will be used for crime prediction and police allocation [37,**?**,17]. Previous approaches used urns models to show how a naive online learning algorithm cannot succeed in estimating the true distribution of events when discovered events and reported events have different incidence rates, and there is a feedback loop between the discovered events and the instrument used to monitor locations [38]. However, implementing this model in a large multi-armed setting is computationally expensive. Therefore, the approach proposed leans toward more computationally efficient techniques used in multi-armed bandit problems.

This work has some limitations. First, the evidence we report relies on simulated data. However, an actual implementation of the strategy in operational settings requires a mechanism to acquire "true" observations of crime sequentially. This work considered the citizens' telephone calls related to crime. However, alternative observational mechanisms can be implemented by considering, for instance, the information provided by citizens using other channels beyond the telephone or the information provided to police in situ during street surveillance, among others. Nevertheless, crime observation through these mechanisms may elicit strategic crime changes, which are not considered in this work [39,11]. Future work may consider a closed-loop estimation of crime underreporting considering criminal adaptation [40,11]. Second, the estimated underreported crime rates correspond to the particular case of Bogotá (Colombia), a large Latin American city with a specific crime dynamic and citizen's reporting habits. Further work may explore the possibility of computing these estimations for other cities where reporting and crime dynamics may change. Finally, the estimation of the "true" crime incident rates from official crime records could be informed by citizen's victimization's surveys, as recently explored for quantification of underreporting [6].

5 Conclusions

This paper studied the estimation of "true" crime over time from underreported crime observations by sequentially considering complementary crime observations. For this, a novel multi-armed bandit model for underreporting estimation was proposed. Efficient algorithms for online learning of the mean of the true-crime distributions in different areas were studied and validated for identifying the fundamental model parameters. This strategy was applied for estimating crime underreporting on two data sources: official crime reports and citizens' victimization surveys. In the first setting, an estimate of the "true" crime incidence rate per geographical unit (1 km^2 cells) and crime underreporting (our true crime scenario) were computed, and underreporting rates were estimated. The second experiment used an estimated Poisson model of crime incidence to simulate real crimes and estimate underreporting using a victimization and reporting survey conducted by the Bogotá's City Chamber of Commerce. In both cases, our method performs well and suggests that this approach can be used to estimate, in an online setup, the underreporting of events. These findings may have implications in public policy because underreporting socially sensitive events can undermine the credibility of official figures and can be strategically used by government agents or influential citizens.

References

1. W. L. Perry, *Predictive policing: The role of crime forecasting in law enforcement operations.* Rand Corporation, 2013.
2. G. Grana and J. Windell, *Crime and intelligence analysis: an integrated real-time approach.* Routledge, 2021.
3. T. C. Hart and C. M. Rennison, *Reporting crime to the police, 1992-2000.* US Department of Justice, Office of Justice Programs Washington, DC, 2003.
4. M. Xie and J. L. Lauritsen, "Racial context and crime reporting: A test of black's stratification hypothesis," *Journal of quantitative criminology*, vol. 28, pp. 265–293, 2012.
5. M. Xie and E. P. Baumer, "Neighborhood immigrant concentration and violent crime reporting to the police: A multilevel analysis of data from the national crime victimization survey," *Criminology*, vol. 57, no. 2, pp. 237–267, 2019.
6. D. Buil-Gil, A. Moretti, and S. H. Langton, "The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis," *Journal of Experimental Criminology*, pp. 1–27, 2021.
7. L. Jaitman and V. Anauati, "The dark figure of crime in latin america and the caribbean," *Journal of Economics, Race, and Policy*, vol. 3, no. 1, pp. 76–95, 2020.
8. C. R. Block and R. L. Block, "Crime definition, crime measurement, and victim surveys," *Journal of social issues*, vol. 40, no. 1, pp. 137–159, 1984.
9. D. Buil-Gil, J. Medina, and N. Shlomo, "Measuring the dark figure of crime in geographic areas: Small area estimation from the crime survey for england and wales," *The British Journal of Criminology*, vol. 61, no. 2, pp. 364–388, 2021.
10. W. G. Skogan, "Dimensions of the dark figure of unreported crime," *Crime & Delinquency*, vol. 23, no. 1, pp. 41–50, 1977.
11. A. Mukhopadhyay, K. Wang, A. Perrault, M. Kochenderfer, M. Tambe, and Y. Vorobeychik, "Robust spatial-temporal incident prediction," in *Conference on Uncertainty in Artificial Intelligence*, pp. 360–369, PMLR, 2020.
12. A. Mukhopadhyay, G. Pettet, S. M. Vazirizade, D. Lu, A. Jaimes, S. E. Said, H. Baroud, Y. Vorobeychik, M. Kochenderfer, and A. Dubey, "A review of incident prediction, resource allocation, and dispatch models for emergency management," *Accident Analysis Prevention*, vol. 165, p. 106501, 2022.
13. R. W. Gillespie *et al.*, "Crime underreporting: theory and implications for the statistical analysis of crime/bebr no. 602," *Faculty working papers; no. 602*, 1979.
14. K. Chaudhuri, P. Chowdhury, and S. C. Kumbhakar, "Crime in india: specification and estimation of violent crime index," *Journal of Productivity Analysis*, vol. 43, pp. 13–28, 2015.
15. D. Buil-Gil, J. Medina, and N. Shlomo, "Measuring the dark figure of crime in geographic areas: Small area estimation from the crime survey for england and wales," *The British Journal of Criminology*, vol. 61, no. 2, pp. 364–388, 2021.
16. D. Buil-Gil, A. Moretti, and S. H. Langton, "The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis," *Journal of Experimental Criminology*, pp. 1–27, 2021.

17. N.-J. Akpinar, M. De-Arteaga, and A. Chouldechova, “The effect of differential victim crime reporting on predictive policing systems,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, (New York, NY, USA), p. 838–849, Association for Computing Machinery, 2021.
18. G. C. Moreira, A. L. Kassouf, and M. Justus, “An estimate of the underreporting of violent crimes against property applying stochastic frontier analysis to the state of minas gerais, brazil,” *Nova Economia*, vol. 28, pp. 779–806, 2018.
19. G. Antunes and E. J. Scott, “Calling the cops: Police telephone operators and citizen calls for service,” *Journal of criminal justice*, vol. 9, no. 2, pp. 165–180, 1981.
20. E. W. Welch and S. Fulla, “Virtual interactivity between government and citizens: The chicago police department’s citizen icam application demonstration case,” *Political communication*, vol. 22, no. 2, pp. 215–236, 2005.
21. C. Oduor, F. Acosta, and E. Makhanu, “The adoption of mobile technology as a tool for situational crime prevention in kenya,” in *2014 IST-Africa Conference Proceedings*, pp. 1–7, IEEE, 2014.
22. A. M. Reyes, J. Rudas, C. Pulido, L. F. Chaparro, J. Victorino, L. A. Narváez, D. Martínez, and F. Gómez, “Multimodal prediction of aggressive behavior occurrence using a decision-level approach,” in *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, vol. 2021, pp. 163–169, 2021.
23. J. Zuo and C. Joe-Wong, “Combinatorial multi-armed bandits for resource allocation,” *arXiv*, 2021.
24. W. Chen, Y. Wang, Y. Yuan, and Q. Wang, “Combinatorial multi-armed bandit and its extension to probabilistically triggered arms,” *arXiv*, 2014.
25. Y. Gai, B. Krishnamachari, and R. Jain, “Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards,” *arXiv*, 2010.
26. P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, pp. 235–256, 05 2002.
27. D. E. Knuth, *The art of computer programming: Volume 3: Sorting and Searching*. Addison-Wesley Professional, 1998.
28. C. Blattman, D. Green, D. Ortega, and S. Tobón, “Place-based interventions at scale: The direct and spillover effects of policing and city services on crime,” Working Paper 23941, National Bureau of Economic Research, October 2017.
29. C. de Comercio de Bogotá, “Encuesta de percepción y victimización de bogotá 2021,” 2022.
30. D. V. Shah, J. N. Cappella, and W. R. Neuman, “Big data, digital media, and computational social science: Possibilities and perils,” *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 6–13, 2015.
31. S. L. Myers, “Why are crimes underreported? what is the crime rate? does it" really" matter?,” *Social Science Quarterly*, vol. 61, no. 1, pp. 23–43, 1980.
32. T. L. Penney, “Dark figure of crime (problems of estimation),” *The encyclopedia of criminology and criminal justice*, pp. 1–6, 2014.
33. G. . P. V. U. S. Collaborators *et al.*, “Fatal police violence by race and state in the usa, 1980–2019: a network meta-regression,” *The Lancet*, vol. 398, no. 10307, pp. 1239–1255, 2021.
34. S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, “Crime prediction using spatio-temporal data,” in *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, pp. 277–289, Springer, 2020.
35. J. Victorino, M. Barrero, J. Rudas, C. Pulido, L. Chaparro, C. Estrada, L. Á. Narváez, and F. Gümez, “Prediction based on time-series of aggressive behaviors. a case study bogotá, colombia,” in *2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE)*, pp. 114–119, IEEE, 2022.
36. W. Spelman and D. K. Brown, *Calling the police: Citizen reporting of serious crime.* US Department of Justice, National Institute of Justice Washington, DC, 1984.
37. R. Richardson, J. Schultz, and K. Crawford, “Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice,” *New York University Law Review*, 2019.
38. D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, “Runaway feedback loops in predictive policing,” *arXiv*, 2017.
39. R. Di Tella and E. Schargrodsky, “Do police reduce crime? estimates using the allocation of police forces after a terrorist attack,” *American Economic Review*, vol. 94, no. 1, pp. 115–133, 2004.
40. H. Elzayn, S. Jabbari, C. Jung, M. Kearns, S. Neel, A. Roth, and Z. Schutzman, “Fair algorithms for learning in allocation problems,” *arXiv*, 2018.