Modelling Under-reported Spatio-temporal Events

Álvaro J. Riascos[1,2], Jose Sebastian Ñungo[1], Lucas Gomez[1]
Mateo Dulce[3]

[1]Quantil [2]University of the Andes [3]Carnegie Mellon University

Septiembre, 2022

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.

Contenido

1. Introduction
2. Related Work
3. Methodology
   - Algorithms
   - Validation
4. Application
   - Data Description
   - Results
5. Conclusions
6. References

Motivation

- The under-reporting of data is a common phenomenon in many data-related problems.
- For example: non-sampling errors in survey sampling, food inspection services, child services, pest controls, building's compliance safety regulations, animal poaching surveillance, crime incidents in a city, among many others.

Motivation

- The under-reporting of data is a common phenomenon in many data-related problems.
- For example: non-sampling errors in survey sampling, food inspection services, child services, pest controls, building's compliance safety regulations, animal poaching surveillance, crime incidents in a city, among many others.
- Under-reporting of socially sensitive events can undermine the credibility of official figures or used strategically.

Motivation

- The under-reporting of data is a common phenomenon in many data-related problems.
- For example: non-sampling errors in survey sampling, food inspection services, child services, pest controls, building's compliance safety regulations, animal poaching surveillance, crime incidents in a city, among many others.
- Under-reporting of socially sensitive events can undermine the credibility of official figures or used strategically.
- Models that simultaneously estimate incidence and under-reporting rates of events can be used to improve the allocation of public resources.

Motivation

- The under-reporting of data is a common phenomenon in many data-related problems.
- For example: non-sampling errors in survey sampling, food inspection services, child services, pest controls, building's compliance safety regulations, animal poaching surveillance, crime incidents in a city, among many others.
- Under-reporting of socially sensitive events can undermine the credibility of official figures or used strategically.
- Models that simultaneously estimate incidence and under-reporting rates of events can be used to improve the allocation of public resources.
- One can target and prioritize the allocation of resources to appropriately monitor and record incidents of interest.

Introduction

- This can be seen as a classical explore-exploit trade-off.

Introduction

- This can be seen as a classical explore-exploit trade-off.
- However, true incidence rate may be elusive: (1) Partial observation (2) Change in behavior.
- In this paper we focus narrowly on the first problem.

Introduction

- This can be seen as a classical explore-exploit trade-off.
- However, true incidence rate may be elusive: (1) Partial observation (2) Change in behavior.
- In this paper we focus narrowly on the first problem.
- To solve this problem, we introduce a combinatorial multi-armed bandit model with under-reporting..
- For the first problem, the literature provides performance guarantees.
- For the second, we capitalize on the asymptotic performance of maximum likelihood estimation.

Introduction

- We provide a potential application of our methodology to the problem of crime victimization and reporting rates at the scale of a large city.

Contenido

1. Introduction
2. Related Work
3. Methodology
   - Algorithms
   - Validation
4. Application
   - Data Description
   - Results
5. Conclusions
6. References

Related Work

- Unit non-response in survey sampling is a common phenomenon with two proposed ways to address the problem. The main techniques are: (i) weighted adjustment of estimators and (ii) data imputation [**?**].

Related Work

- Unit non-response in survey sampling is a common phenomenon with two proposed ways to address the problem. The main techniques are: (i) weighted adjustment of estimators and (ii) data imputation [**?**].
- A closely relate literature comes form the crime pr context, [**?**] addresses the problem of fairness in allocating problems where the monitoring of incidents is censored in a well define way.

Related Work

- Unit non-response in survey sampling is a common phenomenon with two proposed ways to address the problem. The main techniques are: (i) weighted adjustment of estimators and (ii) data imputation [**?**].
- A closely relate literature comes form the crime pr context, [**?**] addresses the problem of fairness in allocating problems where the monitoring of incidents is censored in a well define way.
- From an algorithmic point of view, our problem is similar the online resource allocation problem [**?**, **?**, **?**, **?**], among others.

Related Work

- Unit non-response in survey sampling is a common phenomenon with two proposed ways to address the problem. The main techniques are: (i) weighted adjustment of estimators and (ii) data imputation [**?**].
- A closely relate literature comes form the crime pr context, [**?**] addresses the problem of fairness in allocating problems where the monitoring of incidents is censored in a well define way.
- From an algorithmic point of view, our problem is similar the online resource allocation problem [**?**, **?**, **?**, **?**], among others.
- We draw heavily on [**?**, **?**, **?**], by adapting their online algorithms to our problem and estimating our parametrized model of under-reporting in a online setup.

Contenido

1 Introduction

2 Related Work

3 Methodology
- Algorithms
- Validation

4 Application
- Data Description
- Results

5 Conclusions

6 References

Methodology

CUCB

Combinatorial Upper Confidence Bound Algorithm (CUCB) with under-reporting

```
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 numer of times arm $i$ has been under-reported (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 $\widetilde{X}_{i,t}$, $1 \leq i \leq M$, that have been observed up to round $t$ (initially both $1$), respectively.
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:     $S = \mathrm{Oracle}(\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_m)$.
7:     Play $S$. Observe outcomes of played base arms $i$, and update all $T_i$'s and $\hat{\mu}_i$'s.
8:     For $i \notin S$, observe $\widetilde{X}_{i,t}$ conditional to outcomes played by base arms $i$ in step $7$. Update $\hat{q}_i = \frac{\text{Empirical mean of under-reporting so far observed}}{n\hat{\mu}_i}$
9: end while
```

LLR, UCB1

- With this notation we write the **Learning with Linear Rewards (LLR) algorithm** of [?] in the following way. Replace step 5 in CUCB algorithm by:

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

- We also consider the **UCB1 algorithm** of [?] that ignores the potential association between arms at any moment of time:

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

Validation: Basic parameters

- We did 4 experiments.

M	k	T_{max}	n
12	2	1000	1000

Table: Global parameters. M is the number of arms, K the size of the super arm, T_{max} the of maximum number of simulations and n is the number of trials of each binomial distribution.

Validation: Convergencia μ

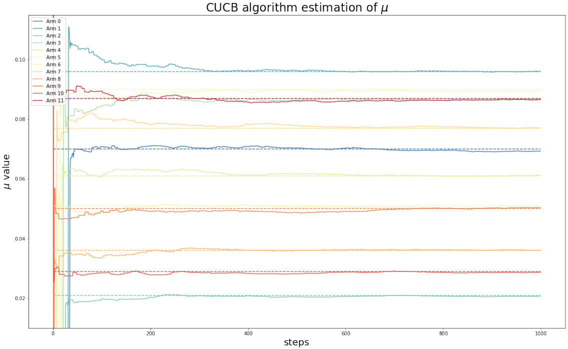

Figure: CUCB Convergence to true arms mean.

Validation: Convergencia q

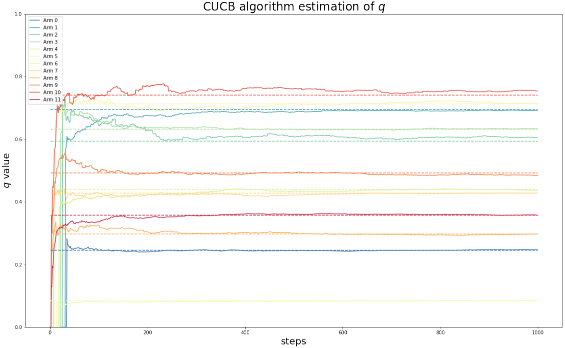

Figure: CUCB Convergence to true arms under-reporting parameters.

Validation: Error

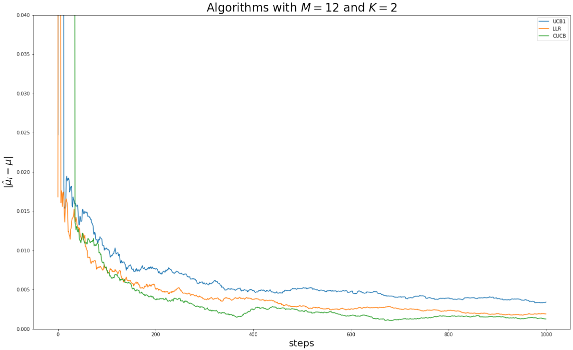

Figure: Convergence error of true arms mean for each algorithm. The error is measures as the euclidean distance between the true mean vector and estimated mean vector per round.

Validation: Visits

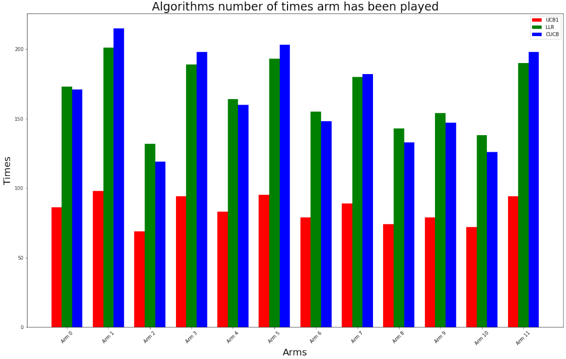

Figure: Number of visits (i.e., fired arms) of algorithms to each arm.

Validation: Error μ

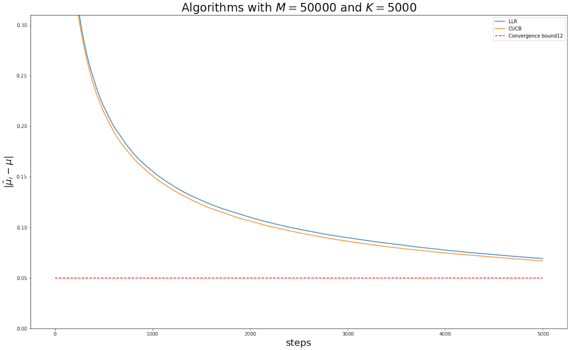

Figure: Convergence error of true arms mean for each algorithm. The error is measured as the euclidean distance between the true mean vector and estimated mean vector per round.

Validation: Time to Completion

	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: Time to completion of one round (note this was previously reported wrong in Quantil seminar). 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.

Contenido

1. Introduction
2. Related Work
3. Methodology
    - Algorithms
    - Validation
4. Application
    - Data Description
    - Results
5. Conclusions
6. References

Data Description

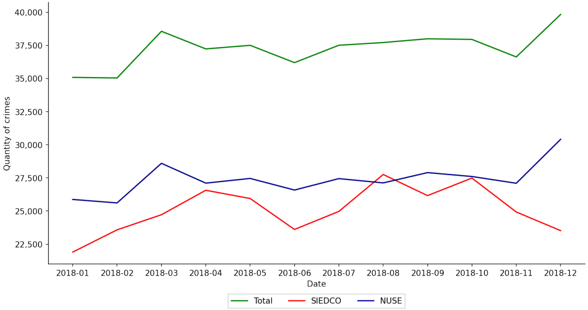

Figure: Crimes by source of information: SIEDCO is the official source of information of crimes in Bogotá. NUSE is the security emergency call center of the city. Total is the sum of both sources eliminating double counting as explained in the main body of the text.

Data Description

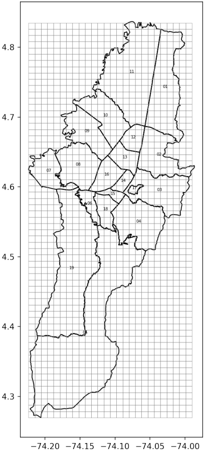

Figure: Bogotá, capital city of Colombia. Figure shows the 19 jurisdictions in which the city is divided and our grid of 1 km^2 cells.

Data Description

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: 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 under-reporting simulated form our Poisson model. The table also reports the population of each jurisdiction and victimization rate.

Results

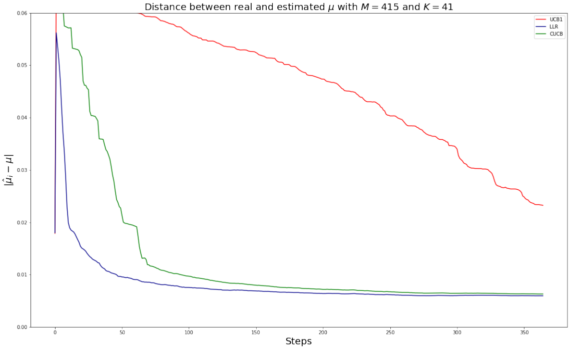

Figure: Convergence of the vector of incidence rates μ to the mean of all crimes per cell across time.

Results

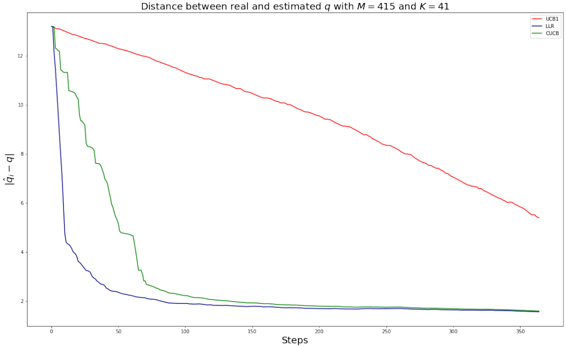

Figure: Convergence of estimated vector q per round to the empirical mean of the under-reporting rate for the whole sample. Euclidean distance reported.

Results

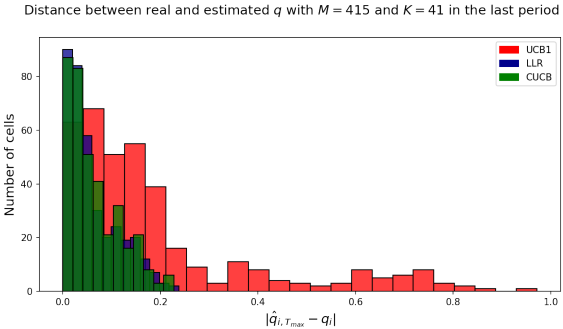

Figure: Histogram of convergence of estimated error of q in the last round to the empirical mean of the under-reporting rate for the whole sample. Absolute value reported.

Results

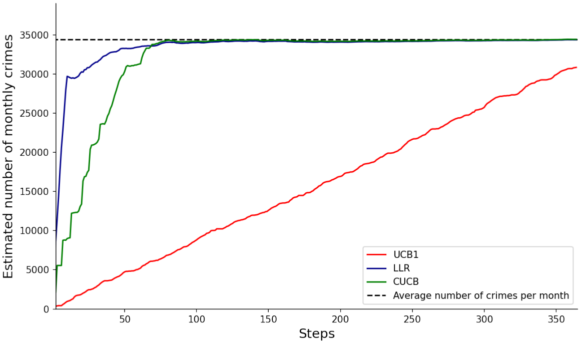

Figure: Convergence of the estimated total number of crimes to the observed number of crimes in the city.

Results

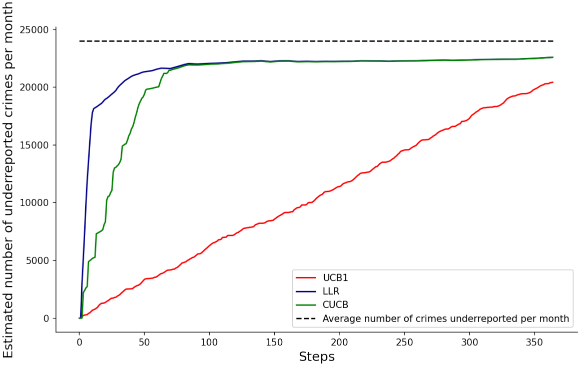

Figure: Convergence of the estimated total (aggregate across cells of) number of under-reported crimes implied by the model.

Results

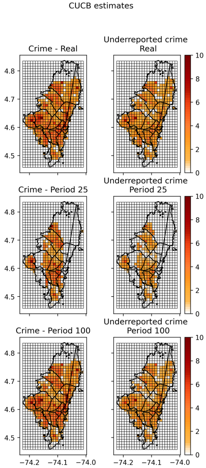

Figure: Heat map illustrating the convergence, using CUCB algorithm, of the estimated crime and under-reporting of events in the city, to the real values. The first column, second and third rows shows the heat map of the estimated crime incidence rates after 25 iterations and 100 iterations, respectively.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

Contenido

1. Introduction
2. Related Work
3. Methodology
   - Algorithms
   - Validation
4. Application
   - Data Description
   - Results
5. Conclusions
6. References

In a nutshell

- We have introduced an under-reporting model of spatio-temporal events that fits well certain applications.
- Used three well known multi-armed bandit algorithms [**?**], [**?**], [**?**] and validated our methodology using simulations that showed the effectiveness of the CUCB algorithm [**?**].
- We then applied our methodology to crime victimization and reporting in Bogotá.
- In both cases, our method performs well and suggests that this methodology could be used to estimate, in an online setup, the under-reporting of events, an important problem in public policy.

Contenido

1. Introduction
2. Related Work
3. Methodology
   - Algorithms
   - Validation
4. Application
   - Data Description
   - Results
5. Conclusions
6. References

References