Mathematical Models of Crime: Prediction, Discrimination, Interpretability, Under-reporting and Equilibrium[1]

Alvaro J. Riascos*

*Universidad de los Andes and Quantil

LACIAM 2023

February 3, 2023

[1] Diseño y validación de modelos de analítica predictiva de fenómenos de seguridad y convivencia para la toma de decisiones en Bogotá. BPIN 2016000100036, recursos regalías. Cooperantes: Secretaría de Seguridad, Convivencia y Justicia de Bogotá,Universidad Nacional de Colombia y Quantil S.A.S.

Motivation

- Between january 2012 and september de 2015, **all homicides and 25 % of all crimes reported in Bogotá occured in 2 %** of street segments.
- During the same period these segments received only 10 % **of the attention of police resources** (Blattman et.al 2017).
- We would like to predict.[2]

[2] Á. J. R. Villegas, J. S. M. Pabón, M. Dulce Rubio, S. Quintero, J. G. Vargas and H. García, "Spatio Temporal Sparsity in Homicide Prediction Models," in IEEE Access, vol. 10, pp. 14359-14367, 2022.

Applications

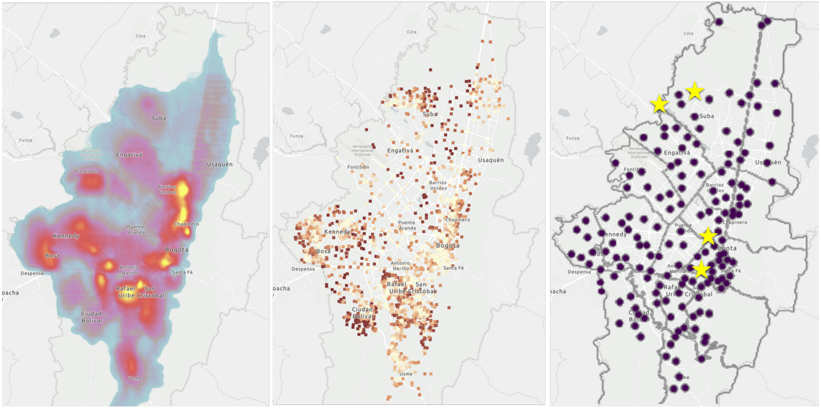

Figura 1: Hotspots, Cameras, Police Stations.

Challenges

No technology comes without **costs and challenges**:

- Interpretability. [3]
- Discrimination.[4]
- **Under-reporting and feedback loop**.
- **Equilibrium and causality**.

[3] Zero-Inflated Embeddings to Analyze Homicide Occurences Patterns. Benavides, H., Gomez O., Dulce M., Rodriguez, P., Riascos, A., and Moreno, J. 2nd International Conference on Computing and Data Science, 2021.

[4] Urcuqui, C., Moreno, J., Montenegro, C., Riascos, A., Dulce, M. (2020, November). Accuracy and Fairness in a Conditional Generative Adversarial Model of Crime Prediction. In 2020 7th International Conference on Behavioural and Social Computing (BESC) (pp. 1-6). IEEE.

Introduction: It is a common phenomenon

- 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.

Introduction: Implications

- **Bias prediction models due to feedback loop**.[5][6]
- **Can be used strategically**.
- Misallocation of resources: [7]

[5] Lum, K., and I., William. 2016. To predict and serve?

[6] Akpinar, N. and De-Arteaga, M., and Chouldechova, A. 2021. The Effect of Differential Victim Crime Reporting on Predictive Policing Systems.

[7] Aaron, Ch., and J. McCrary. 2018. Are U.S. Cities Underpoliced? Theory and Evidence.

Related Work

- Unit non-response in survey sampling: (i) weighted adjustment of estimators and (ii) data imputation.[8]

[8] Särndal, C., Swensson, B and Wretman, J. 2003. Model Assisted Survey Sampling.

[9] Elzayn, H., Shahin J., Jung, Ch., Kearns, M., Seth N., Roth, A. and S. Zachary. 2018. Fair Algorithms for Learning in Allocation Problems.

[10] Chen, W., Wang., Yajun.., Y, Yang., and W, Qinshi. 2014. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms.

Related Work

- Unit non-response in survey sampling: (i) weighted adjustment of estimators and (ii) data imputation.[8]
- Fairness in allocating problem where the monitoring of incidents is censored in a well defined way.[9]

[8] Särndal, C., Swensson, B and Wretman, J. 2003. Model Assisted Survey Sampling.

[9] Elzayn, H., Shahin J., Jung, Ch., Kearns, M., Seth N., Roth, A. and S. Zachary. 2018. Fair Algorithms for Learning in Allocation Problems.

[10] Chen, W., Wang., Yajun.., Y, Yang., and W, Qinshi. 2014. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms.

Related Work

- Unit non-response in survey sampling: (i) weighted adjustment of estimators and (ii) data imputation.[8]
- Fairness in allocating problem where the monitoring of incidents is censored in a well defined way.[9]
- Our problem is similar the online resource allocation problem. [10]

[8] Särndal, C., Swensson, B and Wretman, J. 2003. Model Assisted Survey Sampling.

[9] Elzayn, H., Shahin J., Jung, Ch., Kearns, M., Seth N., Roth, A. and S. Zachary. 2018. Fair Algorithms for Learning in Allocation Problems.

[10] Chen, W., Wang., Yajun.., Y, Yang., and W, Qinshi. 2014. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms.

Related Work

- Unit non-response in survey sampling: (i) weighted adjustment of estimators and (ii) data imputation.[8]
- Fairness in allocating problem where the monitoring of incidents is censored in a well defined way.[9]
- Our problem is similar the online resource allocation problem. [10]
- We draw heavily on Chen. et.al (2014) and others, by adapting their online algorithms to our problem and estimating our parametrized model of under-reporting in a online setup.

[8] Särndal, C., Swensson, B and Wretman, J. 2003. Model Assisted Survey Sampling.

[9] Elzayn, H., Shahin J., Jung, Ch., Kearns, M., Seth N., Roth, A. and S. Zachary. 2018. Fair Algorithms for Learning in Allocation Problems.

[10] Chen, W., Wang., Yajun.., Y, Yang., and W, Qinshi. 2014. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms.

The Model

- **Spatial events**: $X_{i,t}$, where i indexes a spatial location and t indexes the round of the interaction.
- **Unobserved or filtered observations**: a random variable $\widetilde{X}_{i,t}$.
- Application: $X_{i,t}$ binomial with parameter μ, $\widetilde{X}_{i,t} \mid X_{i,t}$ binomial parameter q.
- **Objective**: in a repeated interaction with this environment learn the true mean of the distributions: $X_{i,t}$ and $\widetilde{X}_{i,t}$.

Algorithms:CUCB

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

1: For each arm i, set $\bar{\mu}_i = \text{mín}\left\{\hat{\mu}_i + \sqrt{\frac{3\ln t}{2T_i}}, 1\right\}$.

2: Play $S = \mathrm{Oracle}(\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_m)$.

3: Update all T_i's and $\hat{\mu}_i$'s.

4: For $i \notin S$, observe $\widetilde{X}_{i,t}$ conditional to outcomes played by base arms i.

5: Update $\hat{q}_i = \frac{\text{Empirical mean of under-reporting so far observed}}{n\hat{\mu}_i}$

Algorithms: LLR, UCB1

- **Learning with Linear Rewards (LLR) algorithm** in the following way. Replace in CUCB:

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

- **UCB1 algorithm** ignores the potential association between arms:

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

Validation: Basic parameters

- Four experiments: 12, 100, 1,000 and 50,000 arms.

M	k	T_{max}	n
12	2	1000	1000

Cuadro 1: 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: Convergence μ

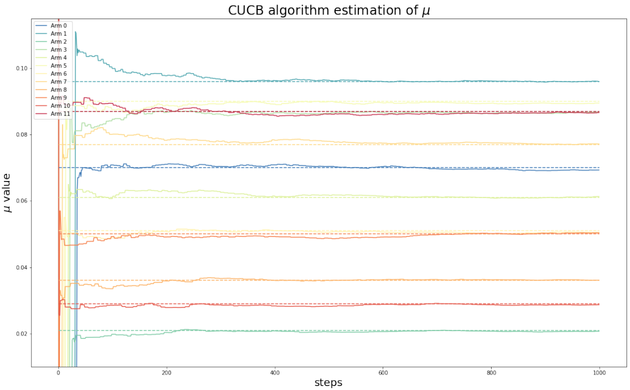

Figura 2: CUCB Convergence to true arms mean.

Validation: Convergence q

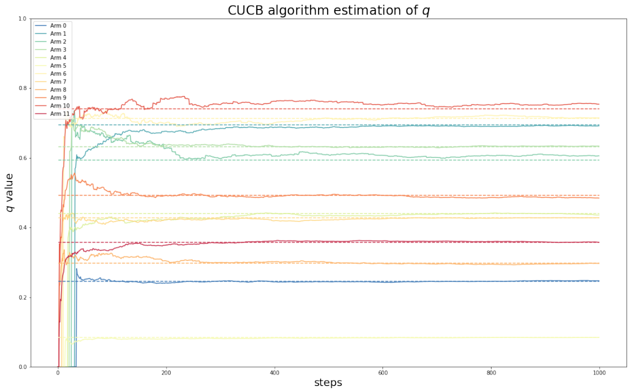

Figura 3: CUCB Convergence to true arms under-reporting parameters.

Validation: Visits

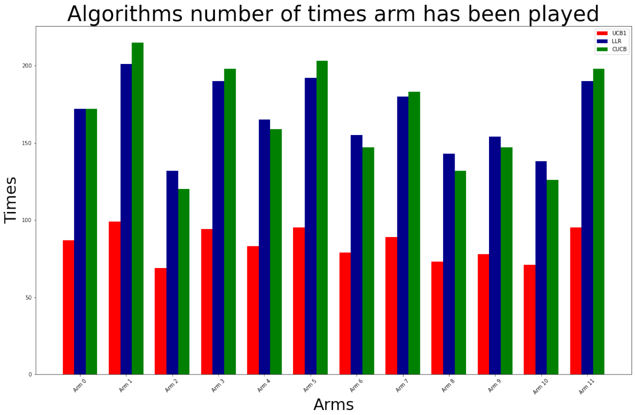

Figura 4: Number of visits (i.e., fired arms) of algorithms to each arm.

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

Cuadro 2: Time to completion. 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.

Data: Evidence of under-reporting Bogotá

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 %

Cuadro 3: Results of Bogotá's City Chamber of Commerce, 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.

Data: Real Crime Estimation

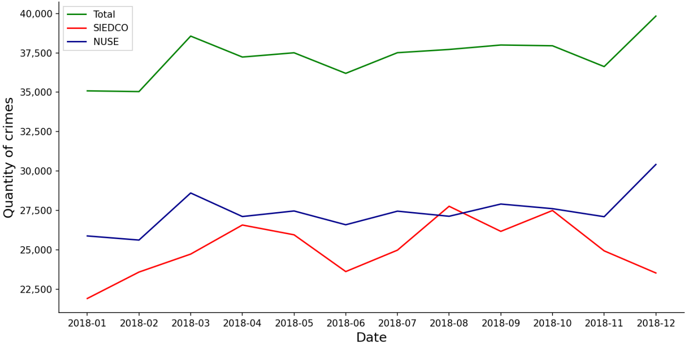

Figura 5: Crimes by source of information: SIEDCO is the official source of information of crimes in Bogotá. NUSE is the security emergency call center of the city. Total is the sum of both sources eliminating double counting.

Arms

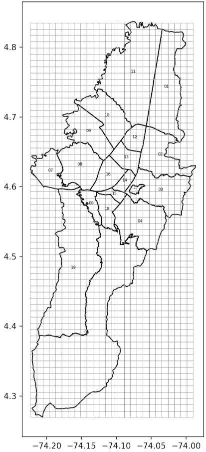

Figura 6: Bogotá, capital city of Colombia. Figure shows the 19 jurisdictions in which the city is divided and our grid of 1 km^2 cells.

Results

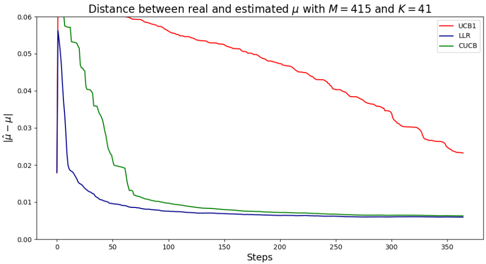

Figura 7: Convergence of the vector of incidence rates μ to the mean of all crimes per cell across time.

Results

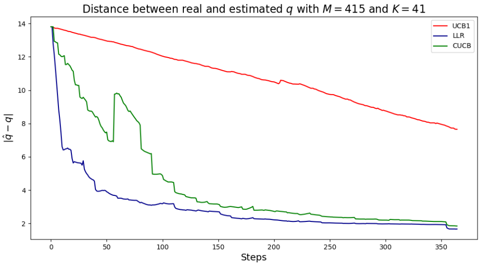

Figura 8: Convergence of estimated vector q per round to the empirical mean of the under-reporting rate for the whole sample. Euclidean distance reported.

Results

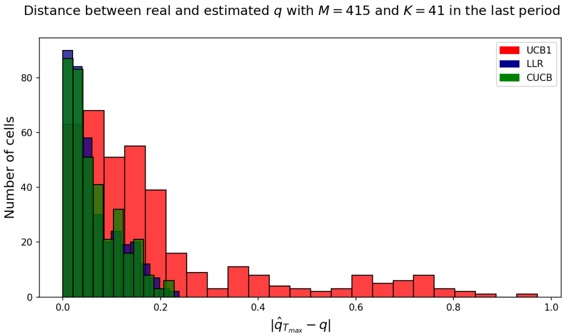

Figura 9: Histogram of convergence of estimated error of q in the last round to the empirical mean of the under-reporting rate for the whole sample. Absolute value reported.

Results

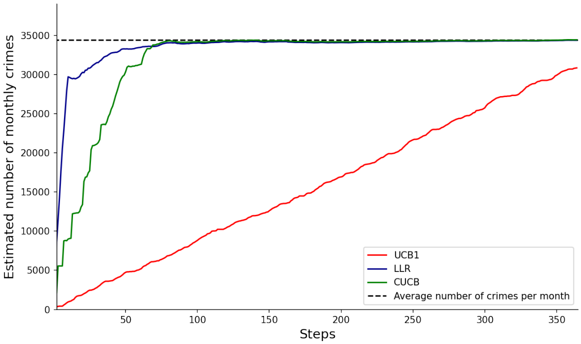

Figura 10: Convergence of the estimated total number of crimes to the observed number of crimes in the city.

Results

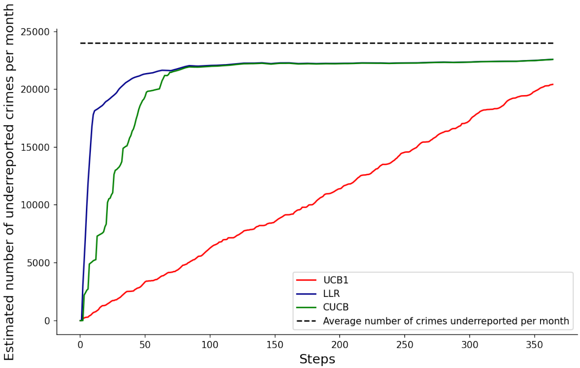

Figura 11: Convergence of the estimated total (aggregate across cells of) number of under-reported crimes implied by the model.

Results

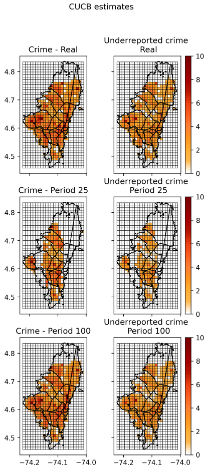

Figura 12: Heat map illustrating the convergence, using CUCB algorithm, of the estimated crime and under-reporting of events in the city, to the real values. The first column, second and third rows shows the heat map of the estimated crime incidence rates after 25 iterations and 100 iterations, respectively. 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.

Introduction

- Prediction models ignore **strategic reaction**.
- We use a unique experimental data to identify the causal impact of police patrolling on crime.
- Use of a **structural model of crime location choice**.
- Estimate **own-and cross-elasticities of crime to patrolling time**.
- **Evaluate alternative patrolling strategies**.

Related Work

- Comprehensive study for the US: David Weisburd and Malay K. Majmundar. 2018. Proactive Policing: Effects on Crime and Communities.
- Aaron, Ch., and J. McCrary. 2018. Are U.S. Cities Underpoliced? Theory and Evidence:
    1. Police elasticity (number of policemen) of violent crime between -0.289 to -0.361.
    2. Property crimes of -0.152 to -0.195.

Related Work

- Blattman, Ch., Green, D.,Ortega, D. and S. Tobón. 2021. Place-based interventions at scale: The direct and spillover effects of policing and city services on crime.
  1. Randomly assigned 756 (206) streets to an 8-month treatment of doubled police patrols (greater municipal services) and measure the direct effect.
  2. Measures spillovers (indirect effects) in streets in a radius of 250 meters: $52,095$ $(21,286)$.
  3. Confidence intervals suggest they can rule out total reductions in crime of more than 2 %.

Spatial Discrete Choice Model

- N potential criminal offenders with symmetric preferences, each of them deciding between $J+1$ locations in the city to commit a crime.
- The associated utility u_{ij}, of agent i, of selecting location j, is given by

$$u_{ij} = \alpha P_j + X_j \beta + \xi_j + \varepsilon_{ij} \tag{3}$$

where:

  - P_j police presence in location j.
  - X_j: K observed characteristics of the location.
  - ξ_j:unobserved characteristics of location j.
  - ε_{ij}: idiosyncratic error term.

Spatial Discrete Choice Model

- Assume $\varepsilon_{ij}, \varepsilon_{ij'}$ are i.i.d. extreme value type I distributed, location choice probabilities are:

$$s_{ij}(P_j, X_j, \xi_j; \alpha, \beta) = \frac{\exp(\alpha P_j + X_j\beta + \xi_j)}{1 + \sum_{k=1}^{J} \exp(\alpha P_k + X_k\beta + \xi_k)} \quad (4)$$

  where option $j = 0$ is assumed to be the **outside option**.

- By symmetry of preferences:

$$S_j(P_j, X_j, \xi_j; \alpha, \beta) = s_{ij}(P_j, X_j, \xi_j; \alpha, \beta)$$

Spatial Discrete Choice Model

- Own- and cross-elasticities of crime:

$$\frac{\partial S_j}{\partial P_\ell} = \begin{cases} \alpha S_j(1 - S_j) & \text{if } j = \ell \\ -\alpha S_j S_\ell & \text{if } j \neq \ell \end{cases} \tag{5}$$

and

$$E_{S_j, P_\ell} \equiv \frac{\partial S_j}{\partial P_\ell} \frac{P_\ell}{S_j} = \begin{cases} \alpha(1 - S_j)P_j & \text{if } j = \ell \\ -\alpha S_\ell P_\ell & \text{if } j \neq \ell \end{cases}. \tag{6}$$

Estimation

- To estimate the structural parameters $\theta = (\alpha, \beta)$ from equation (**??**) we note that:

$$\delta_j = \texttt{log}(S_j) - \texttt{log}(S_0) = \alpha P_j + X_j\beta + \xi_j, \qquad (7)$$

Estimation: Endogeneity

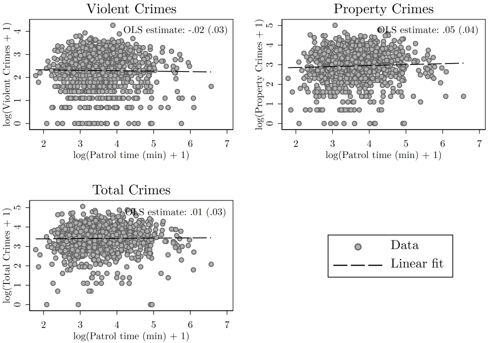

Figura 13: OLS estimation: Biased estimates

Estimation: TSLS

- Starting in January 2016 and during 8 months, 756 out of 1,919 street segments labeled as crime hot spots (out of the 136,984 street segments) received a doubled patrolling time
- We used this randomized treatment to **instrument** the police presence P_j and identify α.

Results: Estimation (Double Selection)

Table 2: TSLS α estimates for the discrete spatial location choice model

	Violent crimes				Property crimes				Total crimes			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
α	-0.004*	-0.003*	-0.005*	-0.006**	-0.004**	-0.005***	-0.009**	-0.008*	-0.005**	-0.005***	-0.008**	-0.008*
	(0.002)	(0.001)	(0.003)	(0.003)	(0.002)	(0.002)	(0.004)	(0.005)	(0.002)	(0.002)	(0.004)	(0.004)
Observations	1,050	1,050	1,050	1,050	1,050	1,050	1,050	1,050	1,050	1,050	1,050	1,050
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Past police presence	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Locality FE	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
R-squared	-0.023	0.309	0.310	0.380	0.000	0.339	0.340	0.414	-0.035	0.187	0.190	0.281

Notes: *** p<0.01, ** p<0.05, * p<0.1. Cluster robust standard errors at the level of locality in parentheses.

Figura 14: α TSLS estimation after double selection.

Results:Estimation Direct Elasticities

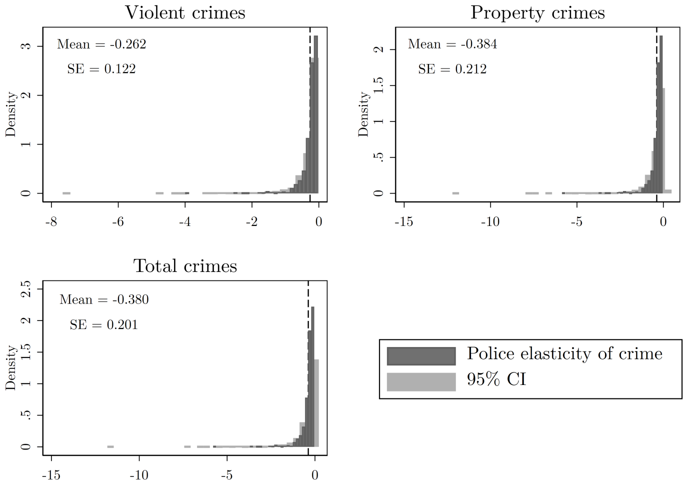

Figura 15: Impact of police presence on crime in the same location.

Results:Estimation Cross Elasticities

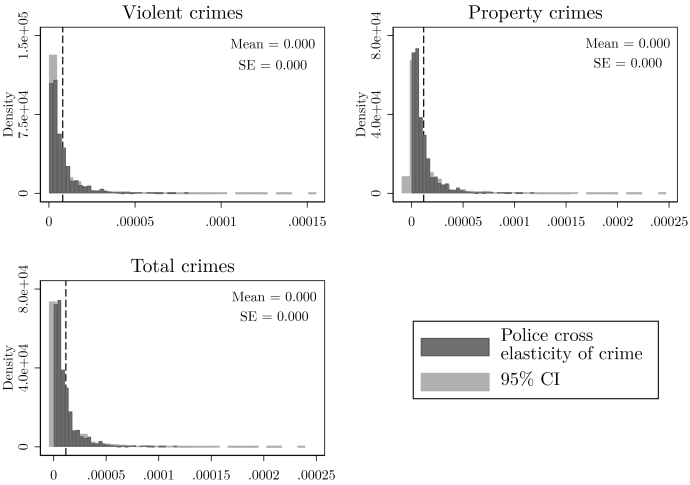

Figura 16: Impact of police presence on crime in different locations.

Results: Optimal Policy

	Violent crimes		Property crimes		Total crimes	
	Mean	Sum	Mean	Sum	Mean	Sum
Observed	11.84	12,435	23.86	25,053	35.70	37,488
Model Predictions						
(1) Original assignment	11.14	11,698	21.58	22,658	32.59	34,222
(2) Optimal assignment	10.35	10,863	19.75	20,739	30.91	32,459
Difference (2)-(1)	-0.79	-862	-1.83	-1,919	-1.68	-1,763

Results: Alternative Policies

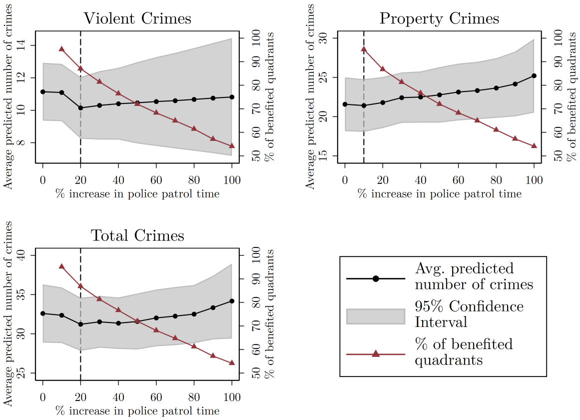

Figura 17

Results: Alternative Policies

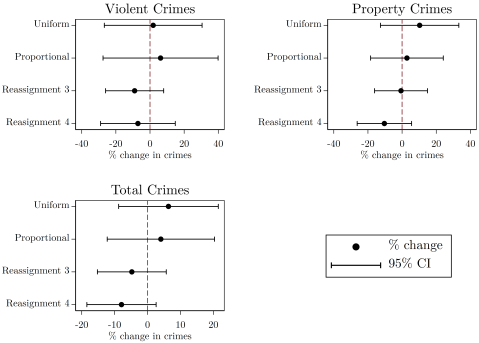

Figura 18

Conclusions

- Switching to the optimal time allocation policy, average reduction of
  - $-7\,\%$ violent crimes per quadrant.
  - $-8.5\,\%$ in property crimes.
  - $-5.2\,\%$ in total crimes.
- That is a reduction of 862 violent crimes, $1,919$ property crimes and $1,763$ in one year.

Thanks

www.alvaroriascos.com