# Crime under-reporting in Bogotá: a spatial panel model with fixed effects

# Crime under-reporting in Bogotá: A Spatial Panel Model with Fixed Effects

Luis Chanci* Subal C. Kumbhakar† Luis Sandoval‡

This Version: October, 2023

**Abstract**

We examine spatial spillovers in violent crime and its under-reporting in Bogotá, Colombia, using a *Cuadrante* (quadrant) level data. To model spatial spillovers, we use a spatial panel model with fixed effects; and to address under-reporting, we use the stochastic frontier approach as a tool. The novel statistical approach is combined with a database of police-reported crimes in Bogotá to examine how influential surrounding areas with high criminal offenses are on crime (under)reporting. The results suggest that spatial correlations are highly significant and that under-reporting is mainly related to interactions with other localities, which have important public policy implications.

**JEL codes:** C23, C33, K42, O17.

**Keywords:** Bogotá (Colombia), Crime, under-reporting, Spillovers, Spatial model, Stochastic frontier.



*Escuela de Ingeniería Comercial, Facultad de Economía y Negocios, Universidad Santo Tomás, Chile. Email: *luischanci@santotomas.cl*.

†Binghamton University and Czech University of Life Sciences Prague, Czech Republic. Email: *kkar@binghamton.edu*.

‡Universidad Militar Nueva Granada. Email: *luis.sandoval@unimilitar.edu.co*.

# 1 Introduction

While crime is a serious social and economic problem,[1] working with crime data raises difficult challenges. Crime has a high variance over time and space,[2] seeming to be contagious in the sense that areas that are the most beset with violent crime tend to be clustered. Moreover, many victims do not report the crime, and, therefore, crime statistics tend to be underreported. This is true in every country/state/municipality/city.

This paper contributes to the literature on crime by examining the following two questions: (i) How significant are spatial spillovers for different types of crimes in Bogotá? (ii) How serious is under-reporting of crimes of different types? To address the first question, we use a comprehensive database containing police reports filed with the National Police of Colombia from 2010 to 2018. And to model crime spillover/interactions (neighborhood effects), we use a spatial model. To address under-reporting, we use the stochastic frontier (SF) model, in which the one-sided random variable is interpreted as under-reporting. Hence, the spatial SF panel model with fixed effects is deployed in dealing with both under-reporting and interconnections (spillover) of crime.

Under-reporting would imply that observed measures of the variables are below their true values (see, for instance, MacDonald 2000, MacDonald 2001, Allen 2007 or Chaudhuri et al. 2015 for discussions in cases like rape or property crimes; or see, for example, Millimet and Parmeter 2021b or Millimet and Parmeter 2021a for a broader discussion about skewed measurement). Thus, in empirical analyses exploring the role of potential factors that may explain differences in crimes across geographic locations,[3] under-reporting could be considered an additional (unobserved) component. The presence of under-reporting reduces the total number of crimes. We address this issue by using the stochastic frontier as a tool. In the SF literature, production inefficiency negatively affects a firm's output, meaning that observed output is less than the maximum possible output, which defines the production frontier (see, for instance, Kumbhakar and Lovell, 2000; Kumbhakar et al., 2022). We follow the SF literature to estimate under-reporting, which has a mathematical similarity with production inefficiency. Inefficiency in a production function will be under-reporting in our empirical crime model.

In addition, this proposed procedure is also aimed at contributing to the debate on the optimal allocation of scarce resources to fight crime in metropolitan areas. To illustrate, we use fine-gridded data for Bogotá as a case study, which is particularly interesting for two reasons. First, it is a representative case of a capital city in a developing country with serious crime problems. Second, many Latin American cities

[1] Crime is a critical issue in most developing countries. For instance, the United Nations Office on Drugs and Crime reported that in 2017 Latin America accounted for about 33 percent of the homicides in the world, while its population was only about 9 percent. Pessino et al. (2018) presents a review of the challenges that this issue puts on public spending in the region.

[2] Glaeser et al. (1996) state that this is perhaps one of the most interesting aspects of crime.

[3] To illustrate, Donohue and Levitt (2001) and Donohue and Levitt (2019) study whether the legalization of abortion and some other socioeconomic variables explain the reductions in crime in the U.S.; or, more recently, Higney et al. (2022) review the role of lead pollution.

now rely on the strategy called *Cuadrantes* (or quadrants) to offer patrolling services. In this strategy, large geographical areas (cities or metropolitan areas) are divided into smaller zones, each covered by: (i) a *Cuadrante*, an organizational scheme of about six police officers offering patrolling services; and (ii) a police post (for instance, a CAI - *Comando de Atención Inmediata*, in the Colombian case). Thus, by implementing a statistical technique parallel to the SF literature, one may think of a *Cuadrante* as a production unit whose outputs are the number of crimes and inputs are police officers, motor vehicles, and other resources used in reducing crime. Hence, estimates of under-reporting can also be viewed as a measure of a *Cuadrante*'s reporting performance (efficiency); that is, lower under-reporting means higher reporting efficiency. This configuration is a novel contribution to the literature because, although many countries have implemented a similar patrolling system, there is still no research on efficiency at the *Cuadrante* level. Furthermore, there are no studies that consider spillovers in criminal activities while examining the efficiency issue.

Our econometric framework and the case study relate to different strands of literature. Beginning with the seminal economic theory of crime and punishment proposed by Becker (1968) and Ehrlich (1975), in which the probability of being arrested and the severity of the sentence, if convicted, affect the cost of engaging in illegal actions. Thus, there is a theoretical expected deterrent effect of police law enforcement on criminal activities. After these works, a growing amount of empirical literature has focused its attention on testing the deterrence hypothesis and providing estimates for the associated elasticity (see, for instance, table 1 in Bun et al., 2020, for a summary). Conversely, other works depart from this theoretical deterrence effect and concentrate on providing additional empirical evidence about other potentially related factors that may affect crime and law enforcement.

Spatial correlations in crime have been documented previously and empirically assessed using related spatial models, as proposed in this paper (e.g, Anselin et al., 2000; Shi and Lee, 2018). Those spatial interactions when studying crime are essential and, according to the literature, may emerge through different mechanisms. Glaeser et al. (1996) point out a peer effect, where agents' decisions about criminal activities are affected by their neighbors' decisions. Billings et al. (2019) provide empirical evidence of this phenomenon by using random variations in school boundaries in Charlotte, the U.S.A. They find that when youths with similar socioeconomic characteristics share school, they are more likely to commit crimes. Using this intuition, we expect spatial spillovers across locations in Bogotá because of the residential segregation in this city and the fact that many students commute between neighborhoods to attend school.

Another potential mechanism is through law enforcement, which may lead to crime displacement. (e.g., Weisburd and Eck, 2004; Rincke and Traxler, 2011). Bronars and Lott (1998) study changes in the law to carry guns in the U.S., finding evidence suggesting that law enforcement deters crime but promotes criminals to move to different areas, highlighting spillover effects. Blattman et al. (2021) present the results of a random experiment in which locations differ in the intensity of police patrolling services in Bogotá in 2016. Their results show no statistically significant effects in this kind of "hot spot policing" strategy but suggest it produces a relocation of crime.

Apart from their central results, Blattman and co-authors present evidence suggesting that the crime reporting rate is far from perfect in Bogotá. Their post-treatment survey shows that about 27 percent of people reported a crime. Thus, one may start from this evidence suggesting that there are indeed crime under-reporting problems in this city and focus the attention on both the empirical modeling of crime under-reporting and the role of geographical spillovers.

To capture the role of spatial interactions, we follow the literature on Spatial Econometrics (e.g., Pace and Barry, 1997; LeSage and Pace, 2009; Elhorst, 2014; Salima et al., 2018) and allow for spatial interactions in both crime and under-reporting. This modeling approach allows us to account for a weighted combination of contemporaneous crime and under-reporting in the surrounding locations to the observation unit, facilitating further analyses of performance, the role of spillovers, or the strategies for clustering units if one wants to explore strategies to improve the system. It is, however, worth noting that, although we are using panel information, this approach to study spillover is not designed to address potential dynamic (spillover) effects in crime. But, Caetano and Maheshri (2018) present evidence suggesting that current crimes may not affect future crime levels.

The econometric specification, combining SF and spatial models (i.e., the spatial autoregressive SAR model) that we propose to examine spatial spillovers in violent crime and its under-reporting at the *cuadrante* level, is not entirely new. However, we extend existing methods in several important directions. For instance, some econometric papers research the unified specification of SF and SAR (Glass et al., 2016; Kutlu et al., 2020). They use the SAR-SF model for panel data, with and without considering endogeneity in some of the explanatory variables and the inefficiency term. Our paper makes an important contribution to this literature by showing how to deal with fixed effects in the SAR-SF model and presenting an alternative estimation technique to the one in Glass et al. (2016). In other words, our model is more general than the one in Glass et al. (2016) and Kutlu et al. (2020). In addition, while these papers use more conventional data sets of firms or countries, our paper contributes by applying the model to an important social issue, the analysis of urban crime, which, to the best of our knowledge, has not been explored before at the micro level the way we do in the present study.[4]

Hence, this paper has two main contributions. First, understanding the determinants of crime is an important societal goal. Accommodating spillover and under-reporting of different types of crimes enriches this goal even further. We contribute to understanding the importance of spillovers in different types of crimes to different regions by dealing with a case with significant diversity in crimes in local geographies. Second, data on crime, economic, and demographic characteristics for small areas are often difficult to obtain, and there are further difficulties in crime data because crimes are often underreported for various reasons. Such under-reportings are likely to be different for different crimes. Also, while statistical methods explore the importance of socioeconomic factors in determining crime, they tend to be agnostic

[4]It is often argued that SAR production models violate axioms of the production function. One can produce more without increasing its use of inputs and/or increasing efficiency - simply because the neighboring producers are producing more, and there is a positive feedback/spillover effect. There is no such problem in the spatial crime model because crimes of a locality can go up (down) if crimes in the neighboring localities go up (down), *ceteris paribus*.

about whether the reporting rate is sensitive to environmental and socioeconomic conditions. We propose an alternative statistical method to overcome these difficulties. Consequently, this paper assesses the level of under-reporting and gives some indications of the importance of spatial interactions using *cuadrantes* as the units of analysis. Our spatial crime panel model introduces dependency in crimes as well as under-reporting, all while controlling for fixed *cuadrante* effects.

We report estimates of under-reporting in terms of the shares that may result from spillovers. These shares provide insights into whether under-reporting is associated with the surrounding environment (neighbors' crime levels) rather than the *cuadrante*'s characteristics - (in)efficiency in patrolling services. We study six criminal offenses, and for a crime like residential burglary, we find that a significant portion of crime under-reporting is related to spatial interactions (spillovers). Conversely, for assault and robbery, results point out the existence of local spillover and suggest that under-reporting is closely associated with characteristics of the *cuadrante* itself. Therefore, our findings suggest that the role of spillovers varies across types of criminal offenses.

In terms of the design of public policies aimed at reducing crime rates, the proposed empirical approach and findings in this paper may aid in analyzing whether or not implementing, for instance, the so-called hot spot policing may be a good strategy. To illustrate, in the case of robbery, the level of externalities we document may cast doubts on arguments suggesting that this type of policy could positively affect surrounding neighborhoods and more remote *cuadrantes* in a metropolitan area.

The remainder of the paper is organized as follows. In Section 2, we present the SF model first and then introduce the spatial stochastic crime panel model with fixed effects, describing the estimation procedure. We use standard notations to facilitate a comparison with the existing literature in econometrics. This will help researchers to understand and develop models that combine the study of under-reporting and spatial interactions. In Section 3, we describe our sample, present some descriptive statistics, and include some maps that graphically illustrate the spatial distribution of different crimes in Bogotá. In Section 4, we present estimates of the Spatial Crime Panel with fixed *Cuadrante* effects and discuss our findings. Finally, in Section 5, we conclude the paper.

# 2 A Spatial Crime Panel Model with F.E.

## 2.1 Panel Stochastic Frontier Models: A bird's-eye view

Before introducing the spatial crime model the statistical/mathematical basis of which is a panel stochastic frontier (SF) model, we first introduce the panel SF model. This is followed by the panel SF model with a spatial structure imposed on it.

A typical panel production stochastic frontier (SF) model is written as

$$\begin{aligned} y_{it} &= \alpha_0 + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \\ \varepsilon_{it} &= \nu_{it} - u_{it} \ ; \qquad\qquad i = 1, \cdots, N \ ; \ t = 1, \cdots, T \end{aligned} \tag{1}$$

where $y_{it}$ is output (in logarithm); $\boldsymbol{x}_{it}$ is $k \times 1$ vector of input variables (in log); subscripts $i$ and $t$ denote production units and time, respectively; and $\alpha_0$ is the intercept. It can be unit-specific (i.e., we can use $\alpha_i$ instead of $\alpha_0$) to capture production heterogeneity specific to each unit. Finally, $\nu_{it} \lessgtr 0$ represents random productivity shock, and $u_{it} \geq 0$ is production inefficiency which is interpreted as a shortfall of output from the maximum possible output, given the input levels. Different variants of the panel SF and their estimation procedures are surveyed in the book by Kumbhakar et al. (2015) and more recently in the survey papers by Kumbhakar et al. (2022). The key in using the SF approach as a tool is to have the error term decomposed into a two-sided noise term and a one-sided inefficiency term. Estimating the model and predicting inefficiency for each observation is accomplished by using distributional assumptions on the error components. See Kumbhakar et al. (2022) for details on these and other advanced SF models.

Note that the SF models can be used as a tool in various atypical cases, such as in wage determination (especially when the offered wage is less than the maximum possible wage given the workers' characteristics), Covid 19, crime under-reporting (when reported cases are less than the actual case), investment shortfall in the presence of credit constraints, etc. In all these cases, a one-sided error term is used along with the two-sided noise term. The interpretation of the one-sided term is not inefficiency, as is the case in the production function. The SF tool is used to predict the one-sided term, the interpretation of which varies with the application. In our case, the one-sided term is under-reporting -not production inefficiency.

The production function model is extended to address spillover effects. The spillover effects of, say, research and development, merger, subsidy, etc., can work through outputs, inputs, and the noise and inefficiency terms. When specifying interaction between spatial units, the model may contain a spatially lagged dependent variable or a spatial autoregressive process in the error term, known as the spatial lag and the spatial error model (with or without inefficiency), respectively. To illustrate, the spatial lag model posits that the dependent variable $y$ depends on the dependent variable observed in neighboring units and on a set of observed local characteristics $x$. That is, the spatial autoregressive model is

$$y_{it} = \alpha_0 + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} \tag{2}$$

where $\sum_j w_{ij} y_{jt}$ is the endogenous spatial lag of the dependent variable and $\rho$ is its coefficient, $w_{ij}$ is the (spatial) weight that captures the effect of $j^{th}$ unit's $y$ on the $i^{th}$ unit's $y$. There is no spillover effect when $\rho = 0$. The above model may not necessarily be a production model.

## 2.2 A Spatial Stochastic Frontier Panel Model with Fixed Effects

We now introduce the spatial stochastic frontier panel model for analyzing crime statistics at the level of small areas, such as *cuadrantes*. The departure from the standard spatial SF panel production model is that the dependent variable is a crime, and the one-sided inefficiency term is crime under-reporting. Thus, the model allows one to control for two important characteristics when studying crime statistics: spatial spillover/correlations and under-reporting. To do so, we start by considering the following spatial lag SF crime model for a specific crime type denoted by $y$ in location $i = 1, ..., N$ and time period $t = 1, ..., T$:

$$\begin{aligned} y_{it} &= \alpha_i + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} \\ \varepsilon_{it} &= \nu_{it} - u_{it} \end{aligned} \tag{3}$$

where $\alpha_i$ is a location-specific fixed effect that captures heterogeneity in crimes (given $\boldsymbol{x}_{it}$); $\boldsymbol{x}_{it}$ is $k \times 1$ vector of exogenous variables explaining $y$; $\sum_j w_{ij} y_{jt}$ is the endogenous spatial lag of the dependent variable, where $w_{ij}$ is the (spatial) weight that captures the effect of $j^{th}$ location's crime on the $i^{th}$ location's crime. The weights are based on the geographic distance and are non-stochastic and pre-specified.

The last two terms in equation (3), $\nu_{it}$ and $u_{it}$, are unobserved random error terms. $\nu_{it}$ is an idiosyncratic noise that can take both positive and negative values. It can also be interpreted as a measurement error and is separated from under-reporting. The term representing under-reporting $u_{it}$ implies that the observed number of criminal offenses would be below its true values and, therefore, is defined as a one-sided error term $u_{it} \geq 0$. Furthermore, we allow for spatial dependence in under-reporting by introducing a spatial structure on $u_{it}$. Specifically, to better capture that under-reporting would imply that crime figures are below their true values, we follow Hou et al. (2023) and introduce a spatial moving average (SMA) process in the under-reporting term, $u$ and the noise term, $\nu$ (specification (6) in equation 2.8 in Hou et al. (2023)) as follows[5]

$$u_{it} = \xi \sum_{j=1}^{N} w_{ij} \dot{u}_{jt} + \dot{u}_{it} \tag{4}$$

$$\nu_{it} = \xi \sum_{j=1}^{N} w_{ij} \dot{\nu}_{jt} + \dot{\nu}_{it} \tag{5}$$

[5]Although we wrote two equations (viz., (4) and (5)), they are connected through $\varepsilon$, and are part of the same equation in (3) and for this is why the same $\xi$ in both. Also two separate spatial parameters in (4) and (5) could not be identified.

Note that $u_{it}$ in the SF literature is inefficiency (Kumbhakar and Lovell, 2000; Kumbhakar et al., 2015), and, since we use the SF as a tool for modeling under-reporting, we rely on the standard assumptions in the SF literature. These are: $\dot{\nu}_{it}$ and $\dot{u}_{it}$ are both i.i.d., and that $\dot{u}_{it} \sim \mathcal{N}^{+}(0, \sigma_{\dot{u}}^2)$ and $\dot{\nu}_{it} \sim \mathcal{N}(0, \sigma_{\dot{\nu}}^2)$. The notation $\mathcal{N}^{+}$ means positive values of the normal distribution (also known as a half-normal distribution).

There are two important features in the proposed econometric specification that are worth mentioning. First, as mentioned, there are related spatial models used in the literature(e.g., Glass et al., 2016; Sun and Malikov, 2018; Hou et al., 2023). In contrast with, for instance, Glass et al. (2016) or Hou et al. (2023), we added fixed effects $\alpha_i$ in our model. This term plays an essential role in dealing with heterogeneity in crime across spatial locations, especially when using detailed data sets. It, however, makes estimation much more difficult (as discussed below). Chen et al. (2014) introduced fixed effects. However, they did not consider spatial effects. Therefore, the models used by Glass et al. (2016) and Chen et al. (2014) are special cases of our empirical specification. Second, contrary to Glass et al. (2016), we add a spatial structure on $u$ as in (4). Thus, in our model, we have two types of spatial dependencies –one in the dependent variable $y$ and another one in the one-sided term $u$, and therefore our model is much more general than some of the existing models.[6]

## 2.3 Identification/Estimation

Since a model can be estimated consistently when it is identified, instead of dealing with these issues separately, we go straight to estimation. Given the distributional assumptions, both the spatial and the SF models are commonly estimated by exploiting the likelihood function. However, estimating the model in equation (3) combines several challenges that have been discussed separately in the Spatial Econometrics and SF literature. To illustrate, a direct (numerical) optimization of the resulting likelihood function is computationally challenging. It is necessary to compute in each iteration the scaled logarithm of a large matrix that is based on the spatial weights matrix (SWM) (see, for instance, LeSage and Pace, 2009). In our case, the number of locations $N$ is larger than 1,000, and the time periods $T$ are about 9, which would mean inverting a $9,000 \times 9,000$ matrix in each iteration. Additionally, the inclusion of fixed effects in the SF framework involves modifications to the likelihood function that raise additional computational issues (see, for instance, Greene, 2005; Chen et al., 2014, for a discussion about the inclusion of fixed effects in the SF). To the best of our knowledge, an estimator for the spatial panel SF with fixed effects, based on

[6] A general nesting spatial (GNS) econometric model may be a more flexible specification to model the spatial relations because it allows for spatial dependency in all the different variables (e.g., the dependent variable, the independent variables, and/or the error term). Thus, although our model is flexible, it may still miss spatial dependency in the explanatory variables. However, we do not find any compelling reason for including spatially lagged covariates in the present application. Also, in empirical applications using fine-gridded data, as we do in this paper by analyzing *cuadrantes*, finding the information for spatial dependence in the covariates is challenging. Nonetheless, methodologically speaking, the proposed spatial modeling approach we present for studying spillovers could naturally be extended to also include, for instance, spatially lagged explanatory variables, making the specification more general. For this, one needs to find economic reasons for including spatially lagged explanatory variables and better data sets.

the likelihood function, is still research in progress (for instance, Lai and Tran, 2021).

Following Hou et al. (2023), we employ an alternative two-step GMM estimation approach for the model in equation (3), which is much simpler and can be summarized as follows. First, we transform the model and use GMM, which avoids distributional assumptions. In contrast, distributional assumptions are essential in the use of the maximum likelihood method (ML). In addition to avoiding distributional assumptions, this step prevents the necessity of numerically integrating a computationally challenging term associated with the multivariate (cumulative) distribution (with a large dimension) in the log-likelihood function that emerges after the transformation to remove the fixed effects.[7] Second, we compute pseudo-residuals from the first stage. These pseudo-residuals are used to recover estimates of under-reporting using GMM in which the moment equations are based on distributional assumptions on the noise and inefficiency terms.

We now discuss these steps in details.

**Step 1: Transformation and estimation via GMM.** Although the model in equation (3) does not satisfy the assumption in standard regressions models that the expected value of the error equals to zero (i.e., $\mathbb{E}(\nu_{it} - u_{it}) = -\mathbb{E}(u_{it}) \neq 0$), it is possible to move the constant associated with the expected value $\mathbb{E}(u_{it})$ to the intercept and rewrite the model as

$$y_{it} = \alpha_i^* + \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \boldsymbol{x}_{it}^{'*} \boldsymbol{\beta_1} + \varepsilon_{it}^* \tag{6}$$

where for $\boldsymbol{x}_{it} = (1, \boldsymbol{x}_{it}^{'*})'$ we pick out the intercept; $\alpha_i^* \equiv \beta_0 + \alpha_i - \mathbb{E}(u_{it})$; and $\varepsilon_{it}^* \equiv \nu_{it} - u_{it} + \mathbb{E}(u_{it})$. Therefore, by construction, $\mathbb{E}(\varepsilon_{it}^*) = 0$, and the resulting model belongs to the family of spatial panel models with fixed effects.[8] Because in our data set $N$ is larger than 1,000, instead of using location dummy variables for the fixed effects, we first follow the literature on spatial panel data (e.g., Lee and Yu, 2010) and employ the transformation approach to eliminate the individual effects in equation (6). Let $Q = (I_T - (1/T)\iota_T \iota_T')$ be the $T \times T$ matrix used to compute the time-demeaned variables in the within transformation in the panel data literature, where $\iota_T$ is a $T \times 1$ vector of ones. Thus, $\left[F_{T,T-1}, \frac{1}{\sqrt{T}}\iota_T\right]$ is the associated orthonormal eigenvector matrix of $Q$, where $F_{T,T-1}$ is the $T \times (T-1)$ submatrix corresponding to the eigenvalues of one. Therefore, one can remove the fixed effects using $F$. Specifically, denote a variable with a tilde as the resulting after multiplying the variable by $F$. For instance, $\widetilde{\boldsymbol{y}}_i$ is a $(T-1) \times 1$ vector, for $i = 1, ..., N$, resulting from $F' \cdot \boldsymbol{y}_i$, where $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT})'$. Hence, this transformation gives the following model in equation (7)

[7] To illustrate, after the transformation, a correlation matrix emerges over various time spans, posing computational complexities for the application of the ML method.

[8] Note that $\mathbb{E}(\varepsilon_{it}^*) = 0$, although $u_{it}$ follows the SMA structure in (4).

$$\widetilde{\boldsymbol{y}} \quad = \quad \rho\left(I_{(T-1)} \otimes W\right)\widetilde{\boldsymbol{y}} + \widetilde{\boldsymbol{X}}\boldsymbol{\beta_1} + \widetilde{\boldsymbol{\varepsilon}} \tag{7}$$

where $\otimes$ denotes Kronecker product and, for instance, $\widetilde{\boldsymbol{y}} = (\widetilde{y}_{(1,1)}, ..., \widetilde{y}_{(1,T-1)}, ..., \widetilde{y}_{(N,T-1)})'$ is a $N \times (T-1)$ vector where the data are sorted first by time and then by spatial units. Therefore, the model in equation 7 is a standard spatial lag model, which can be estimated by GMM (e.g., Lin and Lee, 2010; Liu and Saraiva, 2015).

In short, after implementing the transformation approach and estimation via GMM in the first stage, one can recover estimates for $\rho$ and $\boldsymbol{\beta_1}$ without using distributional assumptions on $\nu_{it}$ and $u_{it}$, avoiding the above-mentioned computational challenges associated with the ML method. An additional benefit of using GMM rather than ML, is related to potential issues with unknown heteroskedasticity (for instance, Debarsy and Ertur, 2019; Doğan and Taşpınar, 2014). However, the literature acknowledges a potential downside: practitioners may obtain $|\hat{\rho}| > 1$ when using GMM. This is an empirical issue that we did not encounter. It is possible to reparameterize $\rho$ so that $|\hat{\rho}| < 1$.[9]

**Step 2: Transformation and estimation of under-reporting via SF.** We use the estimates of $\rho$ and $\boldsymbol{\beta}$ from the first stage to compute the following pseudo-residuals, $r_{it}$,

$$r_{it} \quad = \quad y_{it} - \hat{\rho}\sum_j w_{ji}y_{jt} - \boldsymbol{x}_{it}^{'*}\hat{\boldsymbol{\beta}}_1$$

Based on equation (3), $r_{it} \approx \alpha_i + \beta_0 + \nu_{it} - u_{it}$. Let $\alpha_i + \beta_0 + \nu_{it} - u_{it} = \alpha_i^* + e_{it}$, where $\alpha_i^* = (\alpha_i + \beta_0 - \mathbb{E}(u_{it}))$ and $e_{it} = (\nu_{it} - (u_{it} - \mathbb{E}(u_{it})))$. Thus, we have that $\mathbb{E}(e_{it}) = 0$. And, therefore, one can estimate $\alpha_i^*$ from $\hat{\alpha}_i^* = \bar{r}_i$. In what follows, we use this term to get a simpler expression for estimating under-reporting. Define $\widetilde{r}_{it}$ as $r_{it} - \hat{\alpha}_i^*$. Thus, it is possible to obtain the following expression in equation (8), which would be the central equation for the second stage:

$$\widetilde{r}_{it} \quad \approx \quad \mathbb{E}(u_{it}) + \nu_{it} - u_{it} \tag{8}$$

which can be viewed as a SF model because

[9] Perhaps this discussion is more appropriate for a review in the spatial econometrics literature. Here we emphasize that an advantage of our first step is that one can recover estimates using standard spatial models without having to deal with under-reporting and distributional assumptions associated with it and the noise term.

$$\widetilde{r}_{it} = \text{constant} + \nu_{it} - u_{it} \tag{9}$$

especially when $u_{it}$ is i.i.d. and therefore $\mathbb{E}(u_{it})$ is a constant. Thus, one can use the distributional assumptions on $\nu_{it}$ and $u_{it}$ to estimate the SF model in (9) and get estimates of $\sigma_{\dot{u}}$, $\sigma_{\dot{\nu}}$ and the constant term. When $u_{it}$ follows the SMA structure as in (4), $\mathbb{E}(u_{it}) = \sigma_{\dot{u}}\sqrt{\frac{2}{\pi}}\tau_i'(I_N + \xi W)\iota_N \equiv \mu$, which is a constant, where $\tau_i$ is an $N \times 1$ vector whose $i$th element is 1 and other elements are 0.

Because we are assuming the SMA process in equation (4) to allow for potential spatial dependency in under-reporting, the use of ML is not straightforward. We closely follow the estimation approach in Hou et al. (2023) – henceforth, HZK. They propose GMM estimation of a semiparametric spatial stochastic frontier model, specifying various spatial structures on the composite error. Although their model has functional coefficients and our specification has constant parameters that include fixed effects, the models in the second stage are related, and, therefore, one can implement their approach. In brief, the GMM estimation of the remaining parameters ($\xi$, $\sigma_{\dot{u}}^2$, $\sigma_{\dot{\nu}}^2$), and the prediction of $u_{it}$, relies on exploiting moments of the composite error using the distributional assumptions. To illustrate, the second-moment condition, based on the variance-covariance structure of the error term in equation (9), is

$$\begin{aligned}\mathbb{V}(\boldsymbol{\nu}_t - \boldsymbol{u}_t) &= \sigma_{\dot{\nu}}^2(I_N + \xi W)(I_N + \xi W)' \\ &\quad + \left(1 - \frac{2}{\pi}\right)\sigma_{\dot{u}}^2(I_N + \xi W)(I_N + \xi W)'\end{aligned}$$

Thus, one can compute the sample counterpart of the left side as $T^{-1}\sum_t(\widetilde{r}_t - T^{-1}\sum_t \widetilde{r}_t)(\widetilde{r}_t - T^{-1}\sum_t \widetilde{r}_t)'$, which is a $N \times N$ matrix. Similar to HZK moments can be used to estimate $\sigma_{\dot{\nu}}^2, \sigma_{\dot{u}}^2$ and $\xi$. Therefore, we use the estimated value of $\sigma_{\dot{\nu}}^2$ to obtain $\mathbb{E}(u_{it})$ first, and then $(\alpha_i + \beta_0) = \alpha_i^* + \mathbb{E}(u_{it})$. Note that $\beta_0$ cannot be separated from $\alpha_i$.

Finally, based on the spatial structure of the composite error and insights from Jondrow et al. (1982), HZK propose an approach to predict $u$. Specifically, after estimating the parameters, one can predict under-reporting using $u_{it} = \tau_i(I_N + \xi W)vec\left\{\mathbb{E}(\dot{u}_{it}|\tau_i(I_N + \xi W)^{-1}(\nu_t - u_t))\right\}$, where

$$\mathbb{E}(\dot{u}_{it}|\tau_i(I_N + \xi W)^{-1}(\nu_t - u_t)) = \mu_{it}^* + \sigma_* \frac{\phi\left(-\mu_{it}^*/\sigma_*\right)}{1 - \Phi\left(-\mu_{it}^*/\sigma_*\right)}, \tag{10}$$

for $\mu_{it}^* = -\left(\frac{\sigma_{\dot{u}}^2}{\sigma_{\dot{u}}^2 + \sigma_{\dot{\nu}}^2}\right)(\dot{\nu}_{it} - \dot{u}_{it})$ and $\sigma_* = \sqrt{\left(\frac{\sigma_{\dot{u}}^2\sigma_{\dot{\nu}}^2}{\sigma_{\dot{u}}^2 + \sigma_{\dot{\nu}}^2}\right)}$. The unobserved component $(\nu_t - u_t)$ in (10) is replaced by its estimate $(\tilde{r}_{it} - \mu)$.

**Bootstrap inference.** The estimation approach we propose involves several steps, and therefore, establishing the asymptotic properties of the estimator is non-trivial. Thus, we rely on wild bootstrap to

compute the standard errors, representing a good alternative with well-documented properties (see, for example, HZK for simulation results). Specifically, for the model in (3), the wild bootstrap data-generating process is

$$\boldsymbol{y}^* = (I_{NT} - \hat{\rho}(I_T \otimes W))^{-1} \left( \iota_T \otimes \widehat{(\alpha + \beta_0)} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 + (I_T \otimes (I_N + \hat{\xi}W)\hat{\boldsymbol{\varepsilon}}^*) \right)$$

where $\widehat{\varepsilon}^*_{it} = \widehat{\varepsilon}_{it}\,\eta_{it}$ is the resampled composite residual, where $\eta_{it}$ is as a random variable with mean 0 and variance 1.[10] We, therefore, proceed as follows. First, we estimate the parameters in the model $(\hat{\rho}, \widehat{\alpha_i + \beta_0}, \hat{\boldsymbol{\beta}}_1, \hat{\xi}, \hat{\sigma}^2_u, \hat{\sigma}^2_\nu)$ following the proposed two-steps estimation approach. While $\hat{\rho}$ and $\hat{\boldsymbol{\beta}}_1$ are recovered during the first stage, the estimates of the fixed effect (up to the constant) are computed using $\hat{\alpha}^*_i$ and $\widehat{\mathbb{E}(u)} = f(\hat{\sigma}_u, \hat{\xi})$. Moreover, the estimates $\hat{\boldsymbol{\varepsilon}}$ are obtained from $\hat{\boldsymbol{\varepsilon}}_t = (I_N + \hat{\xi}W)^{-1}\hat{\boldsymbol{\varepsilon}}_t$, where the vector $\hat{\boldsymbol{\varepsilon}}_t = \widehat{\boldsymbol{\nu}_t - \boldsymbol{u}_t}$ is computed using $\tilde{r}_{it}$ and $\widehat{\mathbb{E}(u)}$. Second, for a draw of $\eta$, we generate $\hat{\boldsymbol{\varepsilon}}^*$ and compute $\boldsymbol{y}^*$. Third, using $\boldsymbol{y}^*$ and $\boldsymbol{X}$, we conduct the estimation of the parameters. Finally, we repeat these steps several times to obtain the bootstrap standard errors of the estimators (equation 3). Further details can be found in HZK.

**Direct and Spillover Effects.** The two-step approach allows us to estimate under-reporting, $\hat{u}$. However, according to the literature, the parameters in the spatial model are affected by the spatial matrices, and direct interpretation of estimates must be conducted with caution (LeSage and Pace, 2009; Glass et al., 2016; Kutlu et al., 2020). Specifically, after estimating equation (3), the total estimated role of $u$ on $y$ would be captured by the vector $\ddot{u} = \left[I_T \otimes (I_N - \rho W)^{-1}\right] u$ rather than $u$ alone. Glass et al. (2016) and Kutlu (2018) propose a way of analyzing this term by bringing ideas from the SF literature.[11] This idea, in our case, would mean that we could break down the estimates of under-reporting into two parts: a share that is related to spillovers or the relationship with other geographic locations (like the role that the surrounding neighborhoods play in crime under-reporting) and another share that is not related to spillovers (like more idiosyncratic features). Hence, the use of this decomposition will tell us how significant spatial spillovers are for crime under-reporting in Bogotá.

Define $SU^{spillover}_{it}$ as the share of under-reporting of the $i^{th}$ location that is resulting from spillovers from the other surrounding geographic locations. Also, define $SU^{direct}_{it}$ as the share of under-reporting that is resulting from reasons other than spillovers. Each share is computed following equation (11) (see, Kutlu, 2018, for more details):

$$SU^{direct}_{it} = \frac{\left[(I_N - \rho W)^{-1} u_t\right]_{ii}}{\ddot{u}_{it}}, \qquad SU^{spillover}_{it} = \frac{\sum_{i \neq j} \left[(I_N - \rho W)^{-1} u_t\right]_{ij}}{\ddot{u}_{it}} \tag{11}$$

[10] The classical weighting scheme in wild bootstrap. To illustrate, following Mammen (1993), $\eta_{it}$ is equal to $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$.

[11] Kutlu (2018) and Kutlu et al. (2020) state that the approach in Glass et al. (2016) may be highly sensitive to outliers, proposing the use of the share approach as a more intuitive and robust to an outlier.

Once again, these measures represent shares and, therefore, are aimed at presenting an intuitive overview of what percentage of the estimated under-reporting is related to the characteristics of the *cuadrante* itself and what percentage is associated with the overall criminal environment in the surrounding neighborhoods (spillover effect or indirect effect).

# 3 Data

We combine different types of information, such as police-recorded crimes, maps, and other geographically-referenced information. In this section, we describe the data and sources. We also present descriptive statistics and illustrative maps of the spatial distribution of the crimes in our rich dataset.[12]

**Police recorded crimes** In 2010 the National Police of Colombia implemented a new strategy to offer patrolling services called *Plan Cuadrante*. In general terms, large geographical areas were divided into small zones, each covered by: (i) a *Cuadrante*, an organizational scheme of six police officers offering patrolling services, and (ii) a police post called CAI (Comandos de Atención Inmediata, in Spanish). This police patrolling model was implemented in the major cities of Colombia.

The data we use in this paper are a collection of several crimes linked to each *Cuadrante* in the capital city of Colombia, Bogotá. This is a rich dataset that contains detailed information about the reported crime, such as the specific day of the week and hour, the location (neighborhood), and in some cases, some information about the victim.[13] The primary source is the Criminal Statistics System of the National Police of Colombia - SIEDCO (2019) (Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional, in Spanish).[14]

When possible, we combine several similar crimes into one general category. For instance, violent offenses with similar fatal outcomes are merged with homicides. Thus, we study the following six criminal offenses: (i) residential burglary, (ii) personal injuries, (iii) homicides, (iv) theft and robbery, (v) extortion, and (vi) sexual assault. Moreover, we aggregate the data to a yearly frequency. Considering the reduced geographical area covered by a *cuadrante*, these computations facilitate a significant reduction in the number of zeros while allowing for cross-sectional and time-series variations in the crime variables. The final dataset contains information on $N = 1,049$ *Cuadrantes* over the period 2010-2018.[15]

[12] Note that there are zero values for each crime type. So one cannot use a log transformation of $y$. To deal with zero values, we use a transformed dependent variable, the inverse hyperbolic sine (IHS) of the number of reported crimes, $y$. The IHS transformation of $y$ is $\ln(y + \sqrt{1 + y^2})$.

[13] There is no personal information in the data that may reveal a victim's identity.

[14] Data from La Secretaría Distrital de Seguridad, Convivencia y Justicia de Bogotá, https://scj.gov.co/es/oficina-oaiee/estadisticas-mapas and Datos Abiertos, https://datos.gob.cl/; accessed in 2019.

[15] To check the quality of our dataset, we compute some aggregated statistics using larger geographical areas, such as *Localidades*. When figures in our dataset do not match other governmental records (e.g., reports by the

Detailed spatially organized crime data is difficult to obtain, and although it offers some advantages, there are important caveats to mention. On the one hand, a positive point in using a *cuadrante* as the smallest unit of observation is that it facilities the empirical analysis because the inputs used in the fight against crime are somehow similar across units. On the other hand, such a narrow spatial definition raises several data challenges. Specifically, there is a lack of socioeconomic information that, according to the literature (e.g., Donohue and Levitt, 2001; Levitt, 2004; Sen, 2007), may play a role in explaining geographical differences in crime rates. To illustrate, the index of 'Effective Legalized Abortion Rate' in Donohue and Levitt (2001), presented as an important factor in explaining the sharp decrease in crime in the U.S. during the 1990s, is constructed using historical abortion rates and crimes by cohorts. To the best of our knowledge, there is no such information, neither at the *cuadrante* level nor for the whole city, Bogotá. Likewise, there is no localized information about the population or other socioeconomic characteristics. This means that the dependent variable in the empirical model relies on figures about the number of crimes rather than crime rates. One viable assumption here could be that population density remains somehow stable across units over time. Notwithstanding, in the context of the small geographical zones (*cuadrantes*) of Bogotá there is hardly any variation in the socioeconomic characteristics across nearby units, and, as a result, the impact of these variables could not be econometrically estimated. Still, we enhance our econometric model by including individual fixed effects, allowing for heterogeneity across units. Also, to improve our estimates, we use time dummy variables in the vector $x_{it}$, controlling for aggregated effects, as these variables can capture any temporal variations in the variables.[16] Nevertheless, we conduct a robustness check exercise, and in an extension to the central application, we explore the role of some socioeconomic variables by aggregating crime figures to larger areas, such as *localidades*.

**Spatial information** We use geographical information systems to organize the maps, compute the spatial weights matrices, and illustrate some results. Maps and geographically-referenced data are from Datos Abiertos Bogotá (2021), and we process this referenced information using ArcGIS.[17]

The spatial weight matrix follows a distance approach. The weights $w_{ij}$ are based on the normalized inverse geographic distance between location $i$ and location $j$, which is equivalent to the distance between

Secretaría de Seguridad, Convivencia y Justicia de Bogotá, retrieved from: scj.gov.co/en/oficina-oaiee/boletines; accessed: January 2021), we report a missing value in our data. Therefore, in the case of personal injuries, we dropped the observations for 2014. For robbery and residential burglary the period was reduced to 2013-2018.

[16] There may be permanent under-reporting which will be captured by the fixed effects. It is not possible to identify time-invariant (permanent) under-reporting from time-invariant fixed effects. See Greene (2005) and Chen et al. (2014).

[17] We used three software: ArcGIS, Stata, and R. The first two software, ArcGIS and Stata, were used to create the spatial matrices and manage the (raw) crime reports. We then, taking the spatial matrices and the dataset as inputs, code and estimate the econometric model using R. Some R commands involved, used for instance to compare results with other standard models in the literature, are the `spgm` that is part of the `splm` package (version 1.6-2) (Millo and Piras, 2012), the `stsls` that is part of the `spdep` package (version 1.2-8), and the `spatialreg` package (version 1.2-9) (Pebesma and Bivand, 2023). To facilitate replication and empirical implementation of our proposed model, we have made available an online Jupyter notebook with the central parts of the R code we developed. It also includes the main results.

the patrolling area assigned to *cuadrantes* $i$ and $j$. Thus, $w_{ij} = (1/d_{ij})/(\sum_j^N (1/d_{ij}))$ for $i \neq j$, where $d$ represents spatial distance. We also combine these functions with a distance cut-off criterion, such that $w_{ij} = 0$ for any $d_{ij}$ larger than the cut-off. We select a cut-off of three kilometers (about 1.9 miles) after considering the typical size of an area covered by a *cuadrante*, the geographic area of the *localidades*, the whole area of the city, the average displacement distance between neighborhoods in Bogotá, among other criteria. As a robustness check exercise, we also study how variations in the spatial weights matrix affect our main findings. In the last part of the paper, we present results using the queen contiguity-based weights (Fotheringham and Rogerson, 2009; LeSage and Pace, 2009).

In some cases, producing understandable results for more than 1,000 *cuadrantes* may be challenging. We thus take advantage of the fact that Bogotá is geographically divided into 20 localities (also known as *localidades*, in Spanish) and sometimes present results referring to these (larger) areas.

**Descriptive statistics** Table 1 summarizes the descriptive statistics of the six crime variables in the data set. The third and fifth columns in this table show, for instance, that the average number of homicides is about one, and the maximum is 16. As an illustration, to provide an order of magnitude, these numbers can be translated into an average annual rate of homicides of about 17 per 100,000 population for the whole city. In other words, this average number, computed for only one metropolitan area, is about nine times the average value reported for an entire province in a developed country like Canada. Thus, the figures portray crime as a serious social issue in Colombia.

We explore how these crime figures are distributed over the geographical space. Figure 1 illustrates heat maps for Bogotá using the data for some selected criminal offenses in 2015.[18] The color scales are all based on the same percentiles, and the figure is organized in four panels. In each panel, the minor areas, delimited with gray lines, represent a *cuadrante*. Larger areas, delimited with dark blue and labeled with words like Suba or Bosa, are *localidades*. Panel (a) illustrates the spatial distribution of homicides. The highest numbers of homicides (areas in red) are reported in the southern part of the city, in *localidades* like Ciudad Bolivar. The lowest figures for homicides are in the North, which includes *localidades* like Suba or Usaquen. Panel (b) shows a different pattern for residential burglary. This criminal offense is mainly concentrated in the North. Panel (c) suggests that most robberies occur in the central part, which includes the downtown area of Bogotá, and panel (d) indicates that most personal injuries take place in the western region.

[18] The figure provides an illustration rather than a precise scaled map.

Table 1: Summary statistics.

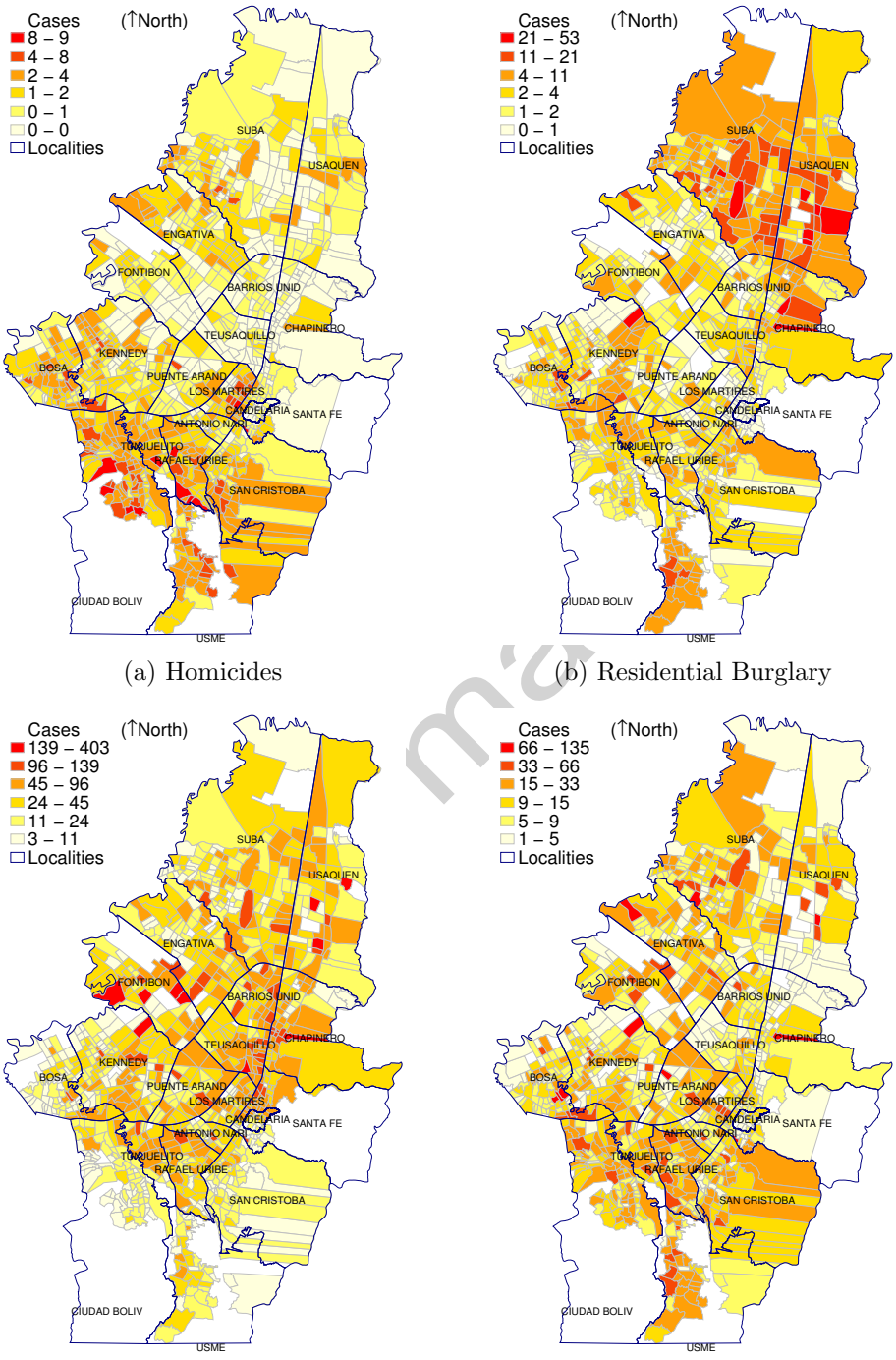
| Variable | Min | Mean | Median | Max | Standard Deviation | Number of observations | Moran's I |
|---|---|---|---|---|---|---|---|
| Residential Burglary | 0 | 4 | 3 | 231 | 6 | 7,343 | 0.225 *** |
| | | | | | | | ( 33.174 ) |
| Personal Injuries | 0 | 13 | 9 | 877 | 18 | 8,392 | 0.061 *** |
| | | | | | | | ( 9.315 ) |
| Homicides | 0 | 1 | 1 | 16 | 1 | 9,441 | 0.315 *** |
| | | | | | | | ( 45.647 ) |
| Theft and Robbery | 0 | 50 | 32 | 2,255 | 66 | 6,294 | 0.175 *** |
| | | | | | | | ( 25.605 ) |
| Extortion | 0 | 0 | 0 | 11 | 1 | 6,294 | 0.121 *** |
| | | | | | | | ( 17.711 ) |
| Sexual Assault | 0 | 1 | 1 | 183 | 3 | 6,294 | 0.029 *** |
| | | | | | | | ( 6.052 ) |

*Notes*: 1. This table presents summary statistics of the number of criminal offenses in Bogotá, Colombia. 2. The unit of analysis is the *cuadrante* using data on criminal offenses on 1,049 *Cuadrantes* in Bogotá over 2010-2018. 3. Differences in the number of observations are explained by missing information in some years. 4. The last column shows the Moran's I measure of spatial autocorrelation, with $E(I) = -0.001$. Z-test in parentheses. 5. *** significant at the 1%; ** 5%; * 10%.

Overall, panels (a) and (b) in figure 1 suggest important spatial correlations in homicides and residential burglary in Bogotá. To explore more about the statistical correlation properties in those patterns, we compute and present the Moran's I measure of spatial autocorrelation for each crime in the last column of Table 1. As the results suggest, there is statistical evidence to reject the null hypothesis that there is no spatial autocorrelation in the crime variables that we will use in the empirical analysis in the next section.[19] In addition, these results support evidence for positive spatial correlation coefficients, $\rho > 0$.

[19]The statistics on the table are based on analytical solutions. In this case, the use of a Monte Carlo simulation approach generates similar results (very small p-values of about 0.001) for all crimes.

Figure 1: Criminal Offenses in Bogotá. Spatial Distribution in 2015.




(a) Homicides (b) Residential Burglary

(c) Theft and Robbery (d) Personal Injuries

*Notes:* 1. The figure has four panels. 2. Each panel illustrates the spatial distribution of a crime in the Metropolitan area of Bogotá. 3. The unit of analysis is the *cuadrante* (quadrant) and larger areas delimited with dark-blue lines and with their names are *localidades* (localities). 4. The color scale are different across panels, but all are based on the same percentiles. 5. Areas in red suggest *cuadrantes* with the highest number of cases.

# 4 Empirical Results

This section first presents the estimates of the parameters in the Spatial Stochastic Crime Panel Model with Fixed Effects in equation (3) for different criminal offenses. Next, based on these estimates, it describes the other statistics of interest, such as under-reporting (by *cuadrante*) or their shares as explained by spillover effects. After presenting the estimates, we conduct a general discussion about the potential implications of our results. We then conclude this section with a robustness check exercise and an extension to the application where we explore the role of some socioeconomic variables that may be related to crime.

## 4.1 Results for the Spatial Crime Panel model

Panel A of Table 2 reports results from residential burglary, personal injuries, homicides, robbery, extortion, and sexual assault from the main model (equation 3). The first two rows in the table summarize the results from the GMM estimation from the first stage, especially the information for the spatial autoregressive coefficient, $\rho$. The other rows show the estimates of $\sigma_{\dot{u}}$, $\sigma_{\dot{\nu}}$, and $\xi$ estimated from the second stage, and, in the bottom part, the associated p-values for testing spatial correlation of the estimated residuals. The results are as follows. First, the spatial autocorrelation parameter estimates are positive and statistically significant for most criminal offenses. As panel A shows, the estimates range from about 0.2, for robbery, to about 0.9, for residential burglary, which is very high. A high value of $\rho$ for burglary, for example, means that the effect of neighborhood burglary is quite important in explaining residential burglary. Second, the estimates of the coefficient associated with the SMA process in under-reporting, $\xi$, are lower. Excluding robbery, the average value is below 0.5. Also, the coefficient is only statistically significant for half of the criminal offenses. In the case of robbery, the estimate is significant and greater than one, suggesting that spatial interactions in under-reporting play an important role in explaining crime.[20] One potential interpretation is that external variations in under-reporting in a *cuadrante* directly affect the neighbors in the sense of local spillovers, which would be relevant later when we discuss the potential effects in terms of public policy. Lastly, the results of the test of spatial correlation for the estimated residuals fail to reject the null hypothesis that there is no spatial autocorrelation.

Furthermore, as a comparison/robustness exercise, in panels B and C of Table 2, we also present results when crimes are modeled using standard spatial panel models. One model is the spatial autoregressive SAR panel (in Panel C), which only includes the spatially lagged dependent variable, and the other model is the SAC panel model (in Panel B), which combines endogenous interaction effects and interaction effects among the error terms. The SAC model would be more general because it includes all possible types of interaction effects and would be comparable to the primary model in Panel A. Neither of these two models considers under-reporting, as we do in our main specification. These models in panels B and C were estimated using standard software packages that are widely used by practitioners (e.g., Pebesma and Bivand, 2023). Overall, the estimates from these models support our first finding: a large and statistically

[20] Note that we are modeling underreporting as a MA process. In such a case, there are no requirements for the coefficient to be below one in absolute value.

significant coefficient on the spatially lagged dependent variable for most criminal offenses.

Table 2: Results Spatial Stochastic Crime Model with FE by crime type

| | ***Crime*** | | | | | |
|---|---|---|---|---|---|---|
| ***Parameter*** | Residential Burglary | Personal Injuries | Homicides | Theft and Robbery | Extortion | Sexual Assault |
| *Panel A. Spatial Stochastic Crime Model with FE:* | | | | | | |
| $\rho$ | 0.915*** | 0.562** | 0.792*** | 0.241 | 0.933*** | 0.819*** |
| | ( 0.158 ) | ( 0.249 ) | ( 0.138 ) | ( 0.435 ) | ( 0.100 ) | ( 0.146 ) |
| $\xi$ | 0.185 | 0.582*** | 0.152*** | 1.220** | 0.00002 | 0.061 |
| | ( 0.635 ) | ( 0.125 ) | ( 0.028 ) | ( 0.498 ) | ( 0.206 ) | ( 0.070 ) |
| $\sigma_\nu^2$ | 0.291 | 0.187 | 0.223 | 0.223 | 0.083 | 0.194 |
| $\sigma_u^2$ | 0.091 | 0.193 | 0.086 | 0.009 | 0.115 | 0.203 |
| Resid. Moran's I (pvalue) | 0.205 | 1.000 | 0.217 | 1.000 | 1.000 | 0.535 |
| *Panel B. SAC Model with FE:* | | | | | | |
| $\rho$ | 0.855*** | 0.186 | 0.754*** | 0.524*** | 1.027 *** | 0.811*** |
| | ( 0.143 ) | ( 0.133 ) | ( 0.202 ) | ( 0.169 ) | ( 0.124 ) | ( 0.148 ) |
| $\lambda$ | 0.181 | 0.386 | 0.135 | 0.586 | -0.417 | 0.029 |
| $\sigma_v^2$ | 0.161 | 0.157 | 0.126 | 0.100 | 0.086 | 0.173 |
| *Panel C. SAR Model with FE:* | | | | | | |
| $\rho$ | 0.915*** | 0.562*** | 0.792*** | 0.241 * | 0.933*** | 0.819*** |
| | ( 0.112 ) | ( 0.102 ) | ( 0.164 ) | ( 0.131 ) | ( 0.150 ) | ( 0.132 ) |
| $\sigma_v^2$ | 0.162 | 0.160 | 0.127 | 0.106 | 0.086 | 0.173 |
| Observations | 7,343 | 8,392 | 9,441 | 6,294 | 6,294 | 6,294 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes*: 1. The table reports results for six separate regressions, one in each column. Panel A presents results for the main specification in this paper with spatial dependence in both $y$ and $u$. 2. For a comparison, Panel B presents the estimates using a Spatial autoregressive combined model (SAC) panel model, $Y = \iota_T \otimes \alpha + \rho(I_N \otimes W)Y + XB + e$ with $e = \lambda(I_N \otimes W)e + v$, and Panel C a Spatial Autoregressive (SAR) panel model $Y = \iota_T \otimes \alpha + \rho(I_N \otimes W)Y + XB + v$ with $v \sim \mathcal{N}(0, \sigma_v^2)$. None of these two models consider under-reporting. 3. Controls means the use of time dummies as regressors but their coefficients are not reported. We used data on criminal offenses from 1,049 *Cuadrantes* in Bogotá during 2010-2018. Standard errors using wild bootstrap are in parentheses. 4. *** significant at the 1%; ** 5%; * 10%.

In short, the overall results are consistent with the patterns we found in the illustrative maps presented

in the descriptive statistics, especially for residential burglary. Hence, the estimates in Table 2 confirm our initial conjecture about the importance of considering spatial dependence in modeling crime data.

## 4.2 Under-reporting and Spillovers

So far, we have presented results following a direct reading of estimates in the specification in equation (3). However, the primary goal of the spatial stochastic crime panel model with fixed effects is to assess under-reporting and the importance of spillovers (or the shares). Thus, we now first present raw under-reporting estimates $\hat{u}$ and then, to facilitate the interpretation, report how much of the under-reporting is from spillover effects.

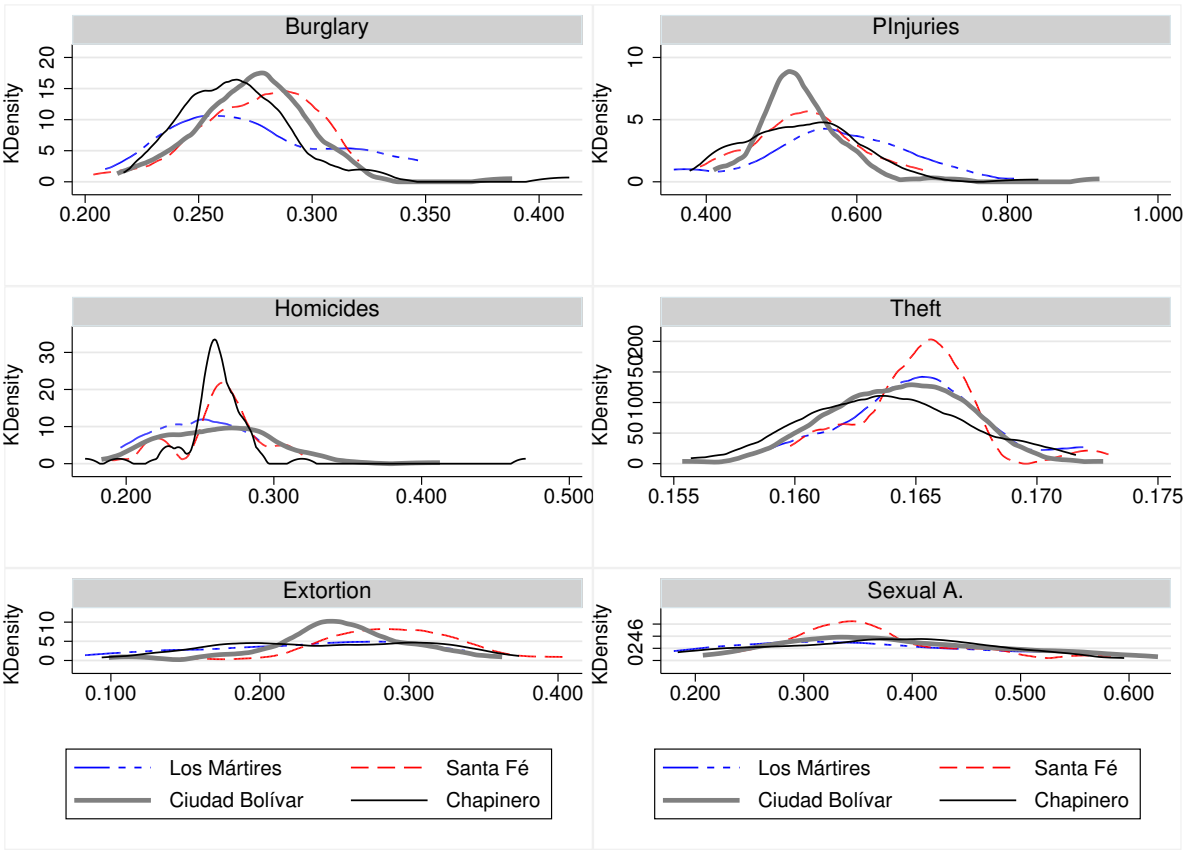
**Under-reporting.** We follow equation (10) and compute estimates of underporting from the estimates of $E(u_{it}|\varepsilon_{it}) \equiv u_{it}$ by *cuadrante* and over time. Considering that there are large numbers of *cuadrantes* (more than 1,000), we present some of the results at the *localidad* level.

Overall, for most crimes, average estimates of under-reporting are in the range of 0.2 - 0.3. In particular, the average estimate for a crime like burglary is about 0.23. Thus, the true figures for residential burglary would be about 23 percent higher than those observed in the reports. A similar interpretation applies to other crimes.[21] To get a better idea of these under-reporting, in Figure 2 we provide kernel density plots of the estimates of under-reporting for each criminal offense. We construct these figures using results for *cuadrantes* located in four *localidades* in the year 2015. As pointed out in Glass et al. (2016), in the context of SF production models, it is not easy to conduct a direct reading of the estimated values of $E(u|\varepsilon)$, but plots are helpful in getting an idea of variations over time and differences across selected *localidades*. Thus, it can be seen that the mean and shape of the density plots of under-reporting do show large differences across spatial units, especially the ones at the tail ends. This result holds for most criminal offenses. Also, *cuadrantes* located in the *localidad* of Ciudad Bolívar presented on average more under-reporting than *cuadrantes* in Los Mártires. Also, we do not observe in the data that the ranking of *localities*, based on under-reporting, varies over time. This result, however, might be an artifact of the assumption that $u$ is iid.

As mentioned before, part of the novelty in using *cuadrantes* as the unit of analysis is that one can rank locations according to each crime's under-reporting benchmark and identify those specific geographical areas that may be susceptible to improvement. Thus, Figure 3 presents illustrative heat maps in which geographical zones can be raked based on under-reporting. The color palette is based on quintiles of under-reporting for a particular crime. In the case of homicides in panel (a), under-reporting seems

[21] Under-reporting for each crime is measured relative to a benchmark (frontier), which is estimated separately for each crime and therefore varies among crimes. As a result, comparing under-reporting across different crimes is not recommended. In this context, comparisons should be made among *localidades*, considering each crime separately. The same principle applies to making cross-country comparisons of, for example, banking efficiency when frontier estimation is conducted separately for each country.

Figure 2: Kernel Density under-reporting




*Note:* This figures has six panels, each for a criminal offense in four *localidades* (a group of *cuadrantes*). Estimated under-reporting $\hat{u}$ in x-axis.

to be evenly distributed in space. The results for robbery and personal injuries in panels (c) and (d), respectively, suggest that there is more predicted under-reporting in the eastern and northern zones of Bogotá. Moreover, when exploring the dataset of estimates in more detail, we find that the *localidades* of Suba and Chapinero have the highest average estimates of crime under-reporting. Also, the *localidad* with the largest number of quadrants in the top quintiles of estimated under-reporting across all the criminal offenses is Suba.

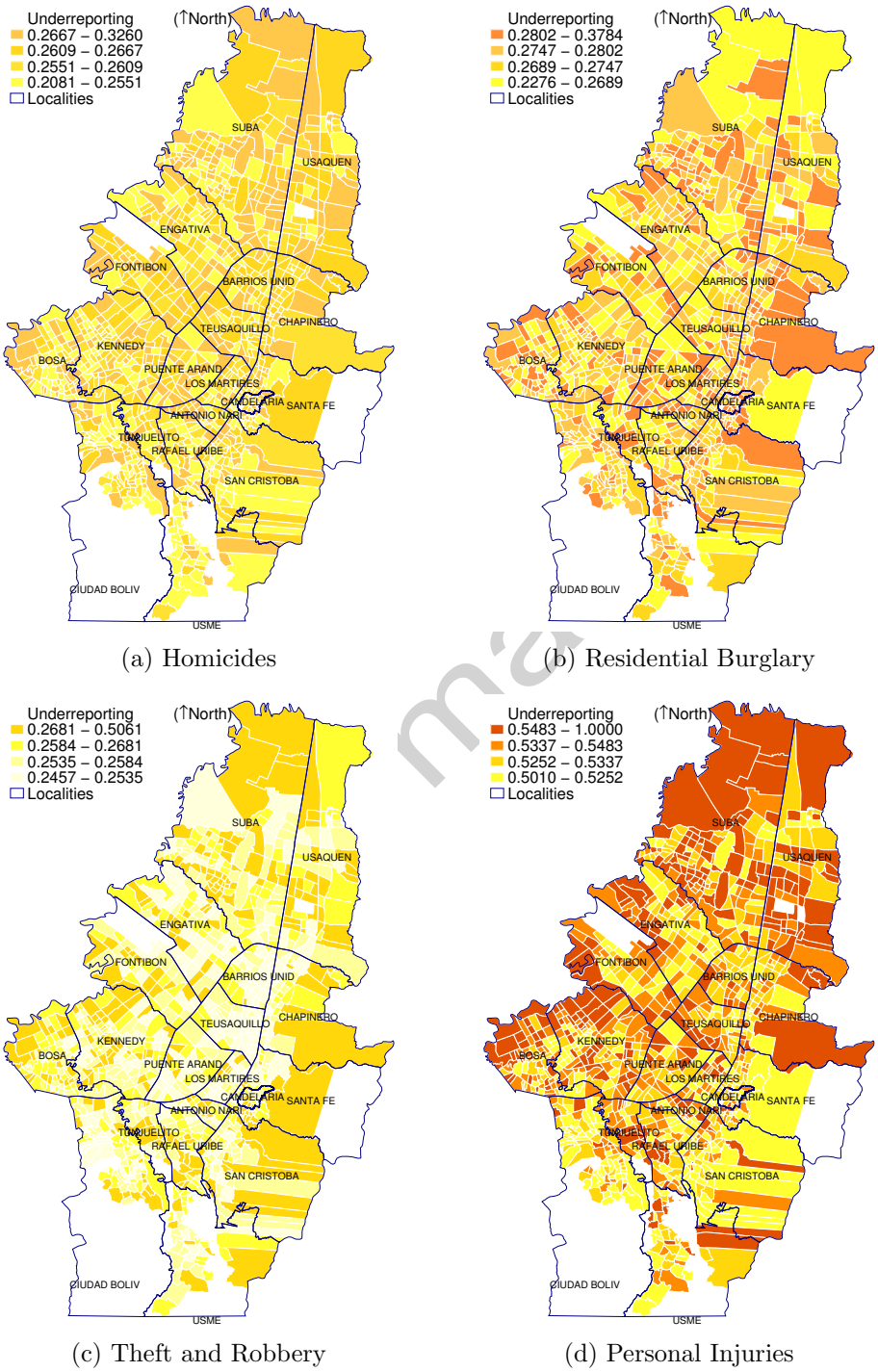
On the other hand, while searching in the data for the location of the *cuadrante* with the best performance, in terms of lowest under-reporting, we find that it is one police post located in the *localidad* of Usaquén. Its estimates of under-reporting are in the lowest quintile for all criminal offenses. Also, for a particular criminal offense like robbery, we find that a *cuadrante* in the *localidad* Usme registers the lowest estimated under-reporting.

**Spillovers.** We now examine under-reporting in terms of idiosyncratic and environmental factors / characteristics. Following the SF literature, the first component may be related to, for instance, unobservables associated with the unit under study. In other words, because not all police officers are the same, one may expect some variations in the quality of patrolling services. As described in section 2, we follow the literature and call this first component the direct effect. The second component concerns environmental externalities, such as the effect of being surrounded by too many hot neighbors exhibiting high crime rates. Therefore, we follow equation (11) and compute the contribution of each component in terms of their shares of under-reporting.

Table 3 presents the results of the estimated shares of direct and indirect (spillover) effects for the six criminal offenses. The figures in the table suggest that for residential burglary, homicides, extortion, and sexual assault, a substantial fraction of under-reporting is associated with spillover effects (indirect effects) rather than characteristics of the *cuadrante* itself (direct effect). Thus, excluding robbery and personal injuries, on average, about 80 percent of the under-reporting is mainly related to the spillover effects in the *cuadrantes* of Bogotá. This outcome could imply that, for various reasons, a victim chooses not to report a crime in specific *cuadrantes* when they perceive the entire area negatively and consider reporting to be not worthwhile.

Finally, when reviewing the results of the shares of spillovers by *localidades* in the dataset, the figures suggest that there are no significant differences in the shares across *localidades*. This result seems consistent with the results we presented using the kernel density plots, where we did not find significant visual differences in the empirical distributions of under-reporting across *localidades*. Furthermore, the figures in the table suggest that heterogeneity in results mainly emerges when the direct and indirect effects are compared.

Figure 3: Predicted under-reporting. Spatial Distribution of $\hat{u}$ in Bogotá.




(a) Homicides (b) Residential Burglary

(c) Theft and Robbery (d) Personal Injuries

*Notes:* 1. The figure has four panels, each illustrates the spatial distribution of under-reporting for a criminal offense in the Metropolitan area of Bogotá: Panel (a) Homicides, (b) Residential Burglary, (c) Theft and Robbery, and (b) Personal Injuries. 2. The unit of analysis is the *cuadrante* (quadrant) and larger areas delimited with dark-blue lines, and with their names, are *localidades* (localities). 3. The color scale are different across panels, but all are based on the same quintiles. 4. Areas in red suggest *cuadrantes* with the highest values of estimated under-reporting, $\hat{u}$.

Table 3: Shares of under-reporting. Average Direct and Indirect Effects.

| **Criminal Offenses** | Direct Effect ($SU^{direct}$) | Indirect Effect ($SU^{spillover}$) |
|---|---|---|
| | mean/(s.d.) | mean/(s.d.) |
| Residential Burglary | 0.092 | 0.909 |
| | ( 0.012 ) | ( 0.012 ) |
| Personal Injuries | 0.440 | 0.560 |
| | ( 0.043 ) | ( 0.079 ) |
| Homicides | 0.215 | 0.785 |
| | ( 0.026 ) | ( 0.026 ) |
| Theft and Robbery | 0.760 | 0.240 |
| | ( 0.005 ) | ( 0.005 ) |
| Extortion | 0.073 | 0.927 |
| | ( 0.022 ) | ( 0.022 ) |
| Sexual Assault | 0.187 | 0.813 |
| | ( 0.042 ) | ( 0.042 ) |

*Notes*: 1. The table contains summary statistics for the estimated shares of under-reporting according to equation (11). 2. The mean is the average value of the shares over time periods and *cuadrantes* for each criminal offense. 3. Standard deviations (s.d.) are in parentheses.

### 4.3 General discussion.

We note that our overall results support the evidence that spatial relationships and spillovers play an important role in modeling crime in Bogotá, but the degree of it varies by crime. In terms of under-reporting, a *cuadrante* surrounded by high crime and under-reporting seems to face more crime and under-reporting problems in comparison with, for instance, a *cuadrante* located in a safer *localidad*.

Collectively, the estimates presented in Table 2, along with the proportions of spillovers detailed in Table 3, indicate significant instances of crime spillovers. First, the coefficient value of the lagged dependent variable in criminal offenses like burglary, extortion, or homicides points to global spillovers are high in relation to other applications in the spatial econometrics literature. This type of spillover means that actions in one *cuadrante* will affect near and far *cuadrantes* and that there could be feedback effects. This situation seems plausible because we are studying crime in our application and that all the observations in the sample are connected because they belong to one metropolitan region. As an illustration, consider the case of homicides, in which a typical incident of gang war in Colombia (for instance, to control extortion, drug smuggling, and other illegal activities) may quickly escalate, producing responses between different *cuadrantes*. This situation ultimately leads to many homicides, mainly in near but also in far locations

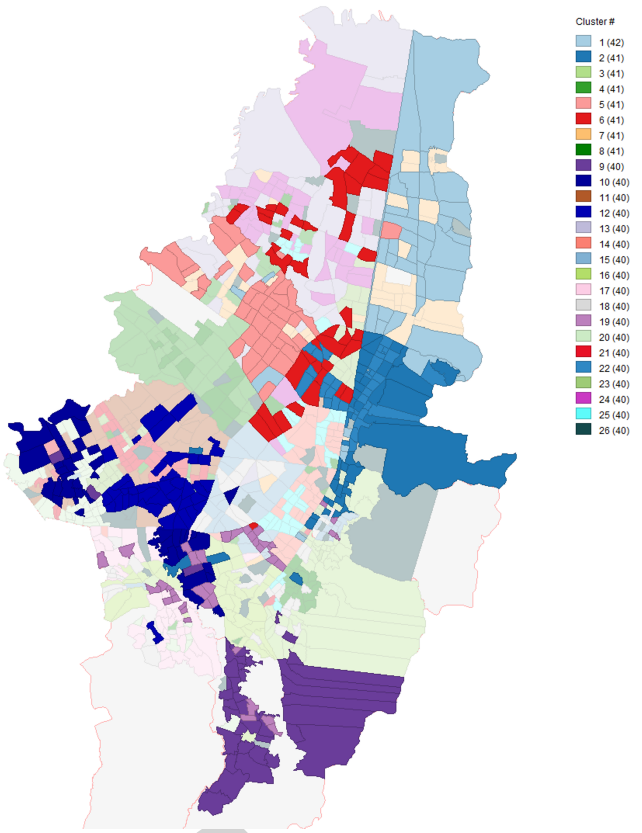
inside the city. Another example could be burglary, for which offenses are usually perpetrated by offenders residing in far distance *cuadrantes* in Bogotá. Second, the results for robbery and personal injuries point more to local spillovers, in the sense that there are spillovers between nearby locations without necessarily a feedback effect. The coefficient associated with the spatially lagged under-reporting, and the shares of direct effects are important for these two criminal offenses.

In terms of public policies aimed at reducing crime, the results of spillovers may suggest that, for instance, a type of 'hot spot' policy could be beneficial as it will deliver additional externalities in surrounding areas. Furthermore, in cases like theft and robbery, the share of direct effects suggests that idiosyncratic characteristics play an important role. This may motivate a closer look at the specific functional attributes of the *cuadrantes* to improve reporting rates. In such a case, the large number of *cuadrantes* may make it challenging for the national police or policymakers to work 'case by case' on performance. Therefore, creating clusters based on the estimates of under-reporting may make it feasible to operate with groups of *cuadrantes* that do not necessarily belong to the same administrative division (e.g., *localidad*). To illustrate, Figure 4 shows some clusters of under-reporting constructed using the max-p regions model (Duque et al., 2012).[22]

There is, however, a note of caution when drawing conclusions regarding the design of public policies aimed at tackling crime. Although we find that spatial spillovers are relevant, it may not be apparent the direction in which such externalities may work after an external policy intervention. On the one hand, one could expect that nearby areas benefit from positive spillovers after implementing place-based policies. This is one of the arguments of those in favor of using 'hot spot' policing, where disproportionate police efforts are directed to high-crime areas. Thus, according to the results in this paper, one may expect positive externalities from this type of policy for criminal offenses in Bogotá, such as residential burglary, extortion, or sexual assault. On the other hand, it is a fact that the size and composition of the police force in a city do not quickly change. Therefore, place-based policies are commonly based on the reallocation of existing resources. Hence, surrounding areas may be negatively impacted by displaced criminals that look for less protected areas. For instance, Blattman et al. (2021) document displacements in property crime in Bogotá after variations in patrolling services on high-crime streets in this city in 2016. Nevertheless, we view the methodology proposed in this paper as a novel approach for capturing spatial spillovers and estimating (unobserved) under-reporting in crime statistics. Moreover, using SF as a tool for computing under-reporting facilities, the construction of rankings of locations based on under-reporting, which helps in the identification of areas that need improvements in policing services in Bogotá.

[22]This model combines the similarity of under-reporting with locational similarity and endogenously determines the number of regions. Because of space constraint, we decided not to delve into max-p further.

Figure 4: Illustration of spatial clustering of estimated under-reporting in Bogotá.




*Notes:* This figure illustrates the spatial distribution of the clusters of under-reporting $\hat{u}$ in Bogotá. Clusters using the max-p regions model (see, Duque et al., 2012, for more details). The unit of analysis is the *cuadrante* and some selected larger areas delimited with light-red lines are *localidades*.

## 4.4 Robustness checks

### 4.4.1 Selection of the Spatial Weights Matrix

The connectivity matrix plays an essential role in spatial models, and there are different researches on selection techniques (see, for instance, Debarsy and Ertur, 2019). Our results are based on a spatial weights matrix constructed using an inverse distance criterion with a pre-defined cut-off of 3 kilometers (about 2 miles). Thus, one may question how robust the results are after modifying this matrix, for instance, by simply changing the distance cut-off point or using rook/queen contiguity-based weights. We now examine the sensitivity of our main findings to variations in the spatial weights matrix.

We first highlight that, although a commonly used data-driven approach might be a good strategy in some cases, such as when the linkages are not obvious, in our case, two distinct features motivate a direct

choice of the spatial matrix. First, spatial connections are defined in terms of geographical boundaries. *cuadrantes* that are located close to the crime-infested *cuadrantes* are likely to have higher weights than those that are farther away. As we described earlier, we see 3 kilometers as a reasonable cut-off after analyzing the typical size of a *cuadrante*, the whole area of the city, and the average displacement distance between *cuadrantes*. Second, the proposed econometric technique is relatively new and does not rely on maximum likelihood estimator (MLE), which makes distributional assumptions on the error terms, to recover the spatial correlation parameter. Furthermore, the error term in our case is composed of statistical noise and under-reporting. We wanted to avoid using distributional assumptions on the error components in the first step. We, therefore, decided not to use the AIC or BIC for selecting the spatial matrix, as these criteria depend on the values of the likelihood function, which will change with a change in the distributional assumptions.

Notwithstanding, as an alternative procedure to select another spatial matrix for reviewing the robustness of our central results, we explore how the residual sum of squares (RSS) is minimized depending on the specification of the spatial matrix. Thus, Table 4 presents the RSS from the first stage for three spatial weights matrices, based on (1) queen continuity, (2) inverse distance with a cut-off of 3 kilometers, and (3) inverse distance with a cut-off of 4 kilometers. According to the results in the table, an alternative candidate is the queen matrix. Hence, compared to the inverse distance approach we originally used, these queen contiguity weights involve a distance cut-off somehow below two kilometers. In other words, this means that we would be examining the role of a reduction in the distance cut-off, which, in our context, is equivalent to assigning more importance to crimes in closer *cuadrantes*.

Table 4: Residual Sum of Squares and the Spatial Matrix.

| SWM based on | RSS |
|---|---|
| Queen Contiguity | 888.1 |
| A cut-off of 3 kilometers | 1,018.4 |
| A cut-off of 4 kilometers | 1,035.1 |

*Notes*: This table presents the Residual Sum of Squares (RSS) from the first stage GMM for different spatial weights matrices (SWM).

Table 5 presents the results of using a spatial matrix based on queen contiguity. We observe some variations in the estimates, although to a lesser extent for most criminal offenses. First, recognizing that the estimates of the spatial correlation coefficients, $\rho$, are obtained from variations in $\sum_j w_{ij} y_{jt}$, we expect more fluctuations in the estimates of $\rho$ (the outcome of this sum changes with the cut-off because the number of neighbors varies). Therefore, when comparing the estimates of $\rho$ in Table 5 with results in Table 2, one may conclude that results for residential burglary, homicides, and sexual assaults remain in the same range, especially after considering the magnitude of the standard errors in Table 5.[23] On

[23]In the case of sexual assault, we notice that the coefficient is somehow larger than one. GMM estimates

the other hand, the two criminal offenses for which we initially found lower spillovers, personal injuries, and robbery, are the ones presenting a significant increase in the estimates of $\rho$ after the reduction in the cut-off distance. For these criminal offenses, the results may confirm that it is precisely the combination of local characteristics rather than spillovers that could explain under-reporting.

We, therefore, stress the importance of the arguments employed when computing the weights in the case of crime. Researchers interested in the methodology we propose in this paper may want to avoid using a general rule of thumb when calculating the weights and better adjust according to their context. The computation of weights is an open research area, and further discussion escapes our goal in this paper. We, however, suggest an analysis based on the characteristics of the city or country in which researchers are implementing the approach we propose.

Table 5: Results Spatial Stochastic Crime Model with FE using a Queen SWM.

| | ***Crime*** | | | | | |
|---|---|---|---|---|---|---|
| ***Parameter*** | Residential Burglary | Personal Injuries | Homicides | Theft and Robbery | Extortion | Sexual Assault |
| $\rho$ | 0.831*** | 0.993 *** | 0.869 *** | 0.696 *** | 0.254 | 1.063 *** |
| | ( 0.078 ) | ( 0.013 ) | ( 0.084 ) | ( 0.211 ) | ( 0.242 ) | ( 0.003 ) |
| $\xi$ | $5.6\text{x}10^{-6}$ | $8.3\text{x}10^{-6}$ | $7.8\text{x}10^{-6}$ | 0.00003 | 0.281 *** | 0.00015 |
| | ( 0.000 ) | ( 0.067 ) | ( 0.055 ) | ( 0.026 ) | ( 0.0001 ) | ( 0.026 ) |
| $\sigma^2_\nu$ | 0.332 | 0.041 | 0.001 | 0.010 | 0.153 | 0.091 |
| $\sigma^2_u$ | 0.0004 | 0.301 | 0.267 | 0.192 | 0.039 | 0.265 |
| Observations | 7,343 | 8,392 | 9,441 | 6,294 | 6,294 | 6,294 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes*: 1. The table reports the results for six independent regressions, each in one column. 2. Results based on the main specification with spatial dependence in both $y$ and $u$ (equation 3). 3. Using data on criminal offenses in 1,049 *cuadrantes* in Bogotá over 2010-2018. 4. Standard errors using wild bootstrap in parentheses. 5. *** significant at the 1%; ** 5%; * 10%.

### 4.4.2 The role of socioeconomic characteristics

As mentioned, we are mainly interested in researching spillover in urban crime while examining efficiency at the *cuadrantes* level. The advantage of following the literature on efficiency analysis is that one can rank units depending on their estimated under-reporting, which may help police departments, researchers, or policymakers to evaluate the patrolling system. However, this empirical application using fine-gridded data (*cuadrantes*) poses a challenge to the covariates: the geographic area covered by a *cuadrantes* is very

may produce this kind of result. However, in this case, the standard errors are large enough to consider that the estimate is still in a reasonable range.

small, and there is no panel localized socioeconomic information to employ as exogenous variables. From theoretical models of crime (e.g., Becker's model), some candidates for socioeconomic characteristics in empirical exercises are the number of police officers, the rate of legalized abortions, poverty measures, unemployment rates, and educational level, among others. In our case, these variables do not change across *cuadrantes* because of their size and close proximity (we are working with only one metropolitan area), and the main variation is temporal. Thus, the effects of socioeconomic variables are captured by the time dummies.

Nevertheless, one may still want to explore the role of some of those socioeconomic factors in crime. To illustrate this in our empirical setting, we searched for official information and surveys representing the smallest possible geographical area in Bogotá to construct a panel dataset. There is some information at an aggregated level and for a few years. Specifically, to compute unemployment rates, one of the most reviewed socioeconomic variables in the literature, we employ the Bogotá Multipurpose Survey. This is a household survey that provides information on homes and inhabitants' social, economic, and urban characteristics. It is a repeated cross-sectional that is available for the years 2011, 2014, 2017, and 2021. Because the survey expansion factors are only available at the level of *localidades*, we define this geographical area as the unit of observation to construct an alternative panel.[24] Thus, as a proxy for education at the locality level, we use the share of public schools that perform well in a (national) high school exit test (Saber Pro 11, from the Ministry of Education). Also, we use infant mortality rates to proxy for other socioeconomic characteristics, such as health (from Secretaría de Salud de Bogotá).

We, therefore, employ an alternative smaller longitudinal dataset consisting of 17 out of the 20 *localidades* in four nonconsecutive years.[25] Furthermore, the spatial weighting matrix is based on queen continuity, considering that *localidades* cover extensive areas compared to a *cuadrante*. Finally, we use a within transformation in the first stage to avoid losing additional observations.[26]

Table 6 presents the results for *localidades*, including the socioeconomic variables. The first robust result is that the estimates associated with the spatially lagged dependent variable are still high and statistically significant across criminal offenses. However, because of the differences in scale, a direct comparison with those results using *cuadrantes* should be made with caution. Also, the shares of direct and indirect effects remain in the same range for almost all crimes. Regarding the role of the socioeconomic variables, we find that the coefficient associated with the unemployment rate shows a positive coefficient that is statistically significant for criminal offenses like personal injuries, extortion, and sexual assault. To illustrate, in the case of extortion, a one percentage point increase in unemployment rates is associated with about an eight percentage point increase in this crime. This value is comparable to the figures reviewed in Bennett and Ouazad (2020).

[24] The most recent surveys are also representative at the UPZ level, a smaller geographical area, but creating panel data with this unit is not feasible yet.

[25] We use those *localidades* that are well-defined across the different surveys.

[26] The transformation approach of using the eigenvector matrix implies losing one year's observations. Reducing the sample size in this way and including time effects in the model raises multicollinearity issues due to the small sample size. Nonetheless, we checked for the effect of using the within transformation in the full dataset of *cuadrantes*, and results in Table 2 were unaffected.

Table 6: Results Spatial Stochastic Crime Model with FE using socioeconomic information.

| | ***Crime*** | | | | | |
|---|---|---|---|---|---|---|
| ***Parameter*** | Residential Burglary | Personal Injuries | Homicides | Theft and Robbery | Extortion | Sexual Assault |
| $\rho$ | 0.599* | 0.517* | 0.696** | 0.778*** | 0.872*** | 0.993*** |
| | ( 0.306 ) | ( 0.312 ) | ( 0.183 ) | ( 0.259 ) | ( 0.168 ) | ( 0.026 ) |
| Unemployment | 0.082 | 0.113* | 0.024 | 0.111 | 0.100** | 0.036* |
| | ( 0.077 ) | ( 0.065 ) | ( 0.052 ) | ( 0.103 ) | ( 0.049 ) | ( 0.055 ) |
| School | 0.001 | 0.001 | 0.008 | -0.0002 | -0.009 | -0.005 |
| | ( 0.006 ) | ( 0.008 ) | ( 0.010 ) | ( 0.007 ) | ( 0.016 ) | ( 0.014 ) |
| Infant mortality | 0.047 | 0.018 | 0.046 | 0.024 | 0.022 | 0.015 |
| | ( 0.047 ) | ( 0.034 ) | ( 0.036 ) | ( 0.030 ) | ( 0.029 ) | ( 0.051 ) |
| $\xi$ | $2.8\text{x}10^{-8}$ | $2.0\text{x}10^{-7}$ | 0.00002 | $4.2\text{x}10^{-8}$ | $6.8\text{x}10^{-7}$ | $1.1\text{x}10^{-8}$ |
| | ( 0.088 ) | ( 0.060 ) | ( 0.009 ) | ( 0.055 ) | ( 0.057 ) | ( 0.030 ) |
| $\sigma_{\nu}^2$ | 1.001 | 0.999 | 0.858 | 1.411 | 0.668 | 0.0002 |
| $\sigma_{\dot{u}}^2$ | 0.012 | 0.152 | 0.019 | 0.001 | 0.024 | 0.824 |
| Mean Direct Eff. | 0.461 | 0.530 | 0.377 | 0.303 | 0.213 | 0.254 |
| Mean Indirect Eff. | 0.539 | 0.470 | 0.623 | 0.697 | 0.787 | 0.746 |
| Resid. Moran's I (pvalue) | 1.000 | 0.997 | 1.000 | 1.000 | 0.996 | 1.000 |
| Observations | 68 | 68 | 68 | 68 | 68 | 68 |
| Time Effects | Yes | Yes | Yes | Yes | Yes | Yes |

*Notes*: 1. The table reports the results for six independent regressions, each in one column. 2. Results based on the main specification with spatial dependence in both $y$ and $u$ (equation 3). 3. Using data on criminal offenses in 17 *localidades* in Bogotá in 2011, 2014, 2017, and 2021. 4. The notation 'Resid. Moran's I (pval)' means the p-value associated to the spatial dependence test applied to the residuals. 5. Standard errors using wild bootstrap in parentheses. 6. *** significant at the 1%; ** 5%; * 10%.

Finally, we note that although working with aggregated areas like *localidades* allows exploring the role of socioeconomic variables, one loses insights from the disaggregated data. Estimates of under-reporting for a whole *localidad* are less informative than those from *cuadrantes*. The novelty in using *cuadrantes* and following the SF literature is that one can debate regarding the patrolling system's performance, the spillovers, and potential strategies for grouping units to work on improving them.

# 5 Conclusions

Researchers working with observational crime data face two crucial challenges: under-reporting and spatial spillovers. In this paper, we use a new tool borrowed from stochastic Frontier to estimate under-reporting.

In doing so, we also take crimes in the neighborhood into account by considering a spatial modeling approach in both the crime-dependent variable and under-reporting. As a result, our econometric model uses a novel Spatial Stochastic Crime Panel Model with Fixed *Cuadrante* Effects.

We use a rich data set of crimes in a city with substantial crime problems: Bogotá, Colombia. We study six criminal offenses in this city: residential burglary, personal injuries, homicides, theft and robbery, extortion, and sexual assault. As expected, the data show distinct spatial patterns, such as a concentration of residential burglary in the northern area of Bogotá.

We find large and statistically significant spatial correlation coefficients in our main empirical specification, with a median value of about 0.8. The estimates are then used to compute under-reporting. We find that they are about 23 percent. Overall, results do not suggest considerable differences in under-reporting across spatial units, *cuadrantes*, but variations in the role of spillovers. For most criminal offenses, we find that about 80 percent of under-reporting is associated with interactions with other localities rather than the characteristics of the *cuadrantes* itself. This figure switches for theft and robbery.

The results in this paper may help policymakers to identify geographical areas affected by under-reporting and possibly understand more about the nature of the city's crime. However, further conclusions about whether the evidence favors, for instance, particular policing approaches need to be drawn with caution.

Finally, two potential extensions to the empirical model in this paper may help researchers interested in understanding more about under-reporting and spatial patterns. First, one could extend the model to include functional coefficients instead of a constant coefficient for the spatial component. This approach may provide, for instance, further insights into whether differences in the intensity of patrolling services (i.e., hot spot policing) affect the spatial patterns of crime. Second, researchers may want to explore additional ideas from the (in)efficiency literature and model the first moments of the under-reporting term. Adding additional socioeconomic information through a functional form for the mean of $u$ may help researchers understand additional perspectives of under-reporting. We leave those extensions for future research.



## Statements and Declarations


- Competing Interests. Luis Chanci acknowledges the financial support from Universidad Santo Tomás, Chile, proyecto interno de investigación "Spillovers and Efficiency: A Spatial Autoregressive Stochastic Frontier Panel Data Model with Fixed Effects" in 2022. Luis Sandoval acknowledges the financial support from Universidad Militar Nueva Granada, project INV ECO 3170 "Temporal and spatial changes in the measurement of crime in Bogotá during 2010-2020."


- CRediT authorship contribution statement. **Luis Chanci:** Conceptualization, Methodology, Writ-

ing Original draft preparation, Writing Review & Editing, Software. **Subal C. Kumbhakar:** Conceptualization, Methodology, Writing Original draft preparation, Writing Review & Editing. **Luis Sandoval:** Data Curation.

- Data availability statements. The crime data that support the findings of this study were obtained from La Secretaría Distrital de Seguridad, Convivencia y Justicia de Bogotá, https://scj.gov.co/es/oficina-oaiee/estadisticas-mapas and Datos Abiertos, https://datos.gob.cl/. Luis Sandoval accessed to the data in 2019.
- Acknowledgments. We thank Tomás Berríos and Jorge Lobos for excellent research assistance. We thank seminar participants at the Society for Economic Measurement (SEM) 2022 conference and the Latin American and Caribbean Economic Association (LACEA) 2022 Annual Meeting for helpful comments. We thank the Associate Editor and two anonymous referee for their constructive comments.

## References

Allen, D. (2007). The reporting and underreporting of rape. *Southern Economic Journal 73*(3), 623–641.

Anselin, L., J. Cohen, D. Cook, W. Gorr, and G. Tita (2000). Spatial analyses of crime. *Criminal justice 4*(2), 213–262.

Becker, G. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy 76*(2), 169–217.

Bennett, P. and A. Ouazad (2020). Job displacement, unemployment, and crime: Evidence from danish microdata and reforms. *Journal of the European Economic Association 18*(5), 2182–2220.

Billings, S., D. Deming, and S. Ross (2019). Partners in crime. *American Economic Journal: Applied Economics 11*(1), 126–50.

Blattman, C., D. P. Green, D. Ortega, and S. Tobón (2021). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. *Journal of the European Economic Association 19*(4), 2022–2051.

Bronars, S. and J. Lott (1998). Criminal deterrence, geographic spillovers, and the right to carry concealed handguns. *The American Economic Review 88*(2), 475–479.

Bun, M. J., R. Kelaher, V. Sarafidis, and D. Weatherburn (2020). Crime, deterrence and punishment revisited. *Empirical economics 59*, 2303–2333.

Caetano, G. and V. Maheshri (2018). Identifying dynamic spillovers of crime with a causal approach to model selection. *Quantitative Economics 9*(1), 343–394.

Chaudhuri, K., P. Chowdhury, and S. Kumbhakar (2015). Crime in india: specification and estimation of violent crime index. *Journal of Productivity Analysis 43*(1), 13–28.

Chen, Y.-Y., P. Schmidt, and H.-J. Wang (2014). Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics 181*(2), 65–76.

Datos Abiertos Bogotá (2021). Cuadrantes de Policía. Bogotá D.C. Secretaría Distrital de Seguridad, Convivencia y Justicia. Retrieved from: https://datosabiertos.bogota.gov.co/dataset/cuadrantes-de-policia-bogota-d-c. Accessed: 2021-06-06.

Debarsy, N. and C. Ertur (2019). Interaction matrix selection in spatial autoregressive models with an application to growth theory. *Regional Science and Urban Economics 75*, 49–69.

Doğan, O. and S. Taşpınar (2014). Spatial autoregressive models with unknown heteroskedasticity: A comparison of bayesian and robust gmm approach. *Regional Science and Urban Economics 45*, 1–21.

Donohue, J. and S. Levitt (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics 116*(2), 379–420.

Donohue, J. and S. Levitt (2019). The Impact of Legalized Abortion on Crime over the Last Two Decades. Technical report, National Bureau of Economic Research.

Duque, J. C., L. Anselin, and S. J. Rey (2012). The max-p-regions problem. *Journal of Regional Science 52*(3), 397–419.

Ehrlich, I. (1975). The deterrent effect of capital punishment: A question of life and death. *The American Economic Review 65*(3), 397–417.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer Berlin Heidelberg.

Fotheringham, S. and P. Rogerson (Eds.) (2009). *The SAGE handbook of spatial analysis*. Sage Publishing.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996). Crime and social interactions. *The Quarterly journal of economics 111*(2), 507–548.

Glass, A., K. Kenjegalieva, and R. Sickles (2016). A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics 190*(2), 289–300.

Greene, W. (2005). Fixed and random effects in stochastic frontier models. *Journal of productivity analysis 23*(1), 7–32.

Higney, A., N. Hanley, and M. Moro (2022). The lead-crime hypothesis: A meta-analysis. *Regional Science and Urban Economics 97*, 103826.

Hou, Z., S. Zhao, and S. C. Kumbhakar (2023). The gmm estimation of semiparametric spatial stochastic frontier models. *European Journal of Operational Research 305*(3), 1450–1464.

Jondrow, J., C. K. Lovell, I. S. Materov, and P. Schmidt (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of econometrics 19*(2-3), 233–238.

Kumbhakar, S. and K. Lovell (2000). *Stochastic frontier analysis*. Cambridge university press.

Kumbhakar, S. C., C. F. Parmeter, and V. Zelenyuk (2022). Stochastic frontier analysis: Foundations and advances ii. *Handbook of Production Economics, Volume 1, ed. Ray, Chambers and Kumbhakar*, 371–408.

Kumbhakar, S. C., H.-J. Wang, and A. Horncatle (2015). *A Practitioner's Guide to Stochastic Frontier Analysis*. Cambridge University Press.

Kutlu, L. (2018). Estimating efficiency in a spatial autoregressive stochastic frontier model. *Economics Letters 163*, 155–157.

Kutlu, L., K. Tran, and M. Tsionas (2020). A spatial stochastic frontier model with endogenous frontier and environmental variables. *European Journal of Operational Research 286*(1), 389–399.

Lai, H.-p. and K. Tran (2021). Persistent and transient inefficiency in spatialautoregressive panel stochastic frontier model. Technical report.

Lee, L.-f. and J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of econometrics 154*(2), 165–185.

LeSage, J. and R. K. Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Levitt, S. (2004). Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic perspectives 18*(1), 163–190.

Lin, X. and L.-f. Lee (2010). Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics 157*(1), 34–52.

Liu, X. and P. Saraiva (2015). GMM estimation of SAR models with endogenous regressors. *Regional Science and Urban Economics 55*, 68–79.

MacDonald, Z. (2000). The impact of under-reporting on the relationship between unemployment and property crime. *Applied Economics Letters 7*(10), 659–663.

MacDonald, Z. (2001). Revisiting the dark figure: A microeconometric analysis of the under-reporting of property crime and its implications. *British Journal of Criminology 41*(1), 127–149.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics 21*(1), 255–285.

Millimet, D. and C. Parmeter (2021a). Covid-19 severity: A new approach to quantifying global cases and deaths. Technical report, IZA Discussion Paper.

Millimet, D. L. and C. F. Parmeter (2021b). Accounting for Skewed or One-Sided Measurement Error in the Dependent Variable. *Political Analysis*, 1–23.

Millo, G. and G. Piras (2012). splm: Spatial panel data models in r. *Journal of statistical software 47*, 1–38.

Pace, K. and R. Barry (1997). Quick computation of spatial autoregressive estimators. *Geographical analysis 29*(3), 232–247.

Pebesma, E. and R. S. Bivand (2023). *Spatial Data Science With Applications in R*. Chapman & Hall.

Pessino, C., A. Izquierdo, G. Vuletin, et al. (2018). *Better Spending for Better Lives: How Latin America and the Caribbean Can Do More with Less*, Volume 10. Inter-American Development Bank.

Rincke, J. and C. Traxler (2011). Enforcement spillovers. *Review of Economics and Statistics 93*(4), 1224–1234.

Salima, B. A., L. Julie, and V. Lionel (2018). Spatial econometrics on panel data. *in Handbook of Spatial Analysis: Theory and Application with R. . Edited by V. Loonis and MP. de Bellefon.* (7), 179–203.

Sen, A. (2007). Does increased abortion lead to lower crime? Evaluating the relationship between crime, abortion, and fertility. *The BE Journal of Economic Analysis & Policy 7*(1).

Shi, W. and L.-f. Lee (2018). The effects of gun control on crimes: a spatial interactive fixed effects approach. *Empirical economics 55*(1), 233–263.

SIEDCO (2019). Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional - SIEDCO. Secretaría Distrital de Seguridad, Convivencia y Justicia. Alcaldía Mayor de Bogotá D.C. Retrieved from: https://scj.gov.co/en/oficina-oaiee/bi/seguridad_convivencia/siedco. Accessed: 2019-06-06.

Sun, Y. and E. Malikov (2018). Estimation and inference in functional-coefficient spatial autoregressive panel data models with fixed effects. *Journal of Econometrics 203*(2), 359–378.

Weisburd, D. and J. E. Eck (2004). What can police do to reduce crime, disorder, and fear? *The annals of the American academy of political and social science 593*(1), 42–65.