A Semiparametric Poisson Autoregressive Model Applied to Vehicle Theft in Bogotá

Santiago Lozano Sandino[1]
Pontificia Universidad Católica de Chile

Abstract

In crime policy there's a common interest in understanding criminal activity and reducing its incidence. Criminal events tend to be spatial-temporal related; analysing such dependencies is one of the objectives of this application. Additionally, depending on the type of crime and its spatial and temporal level of disaggregation, their occurrence may be relatively infrequent. Under these conditions a low-count model that considers spatial-temporal dependencies may be adequate for both predicting and understanding criminal activity within small windows of time and low level geographical disaggregation. I use a Poisson autoregressive model to study vehicle theft in localities of Bogotá. The model used here also utilizes a Dirichlet process that allows clustering localities with similar behaviour.

Keywords: Crime Policy, Nonparametric and Semiparametric Bayesian Methods, Dirichlet Process and Chinese Restaurant Process.

Introduction

Criminal activity is a topic of great interest in public policy. There are many institutions in Colombia with areas completely dedicated to study crime, with the intention of preventing and understanding it. Just to name a few, institutions such as Colombia's National Police, the Ministry of Justice, the National Institute of Forensic Medicine, among others, have dependencies particularly dedicated to study this type of phenomena. There are many fields that contribute to the understanding of such phenomena, such as sociology, psychology, economics, etc. Nowadays, with the so called *data revolution*, there's a need for governments to use data and statistics to make evidence-based decision. There are many statistical methods that can be used to analyse crime-related data, such as the one used in this study.

I intend to adapt a model proposed by Aldor-Noiman, Brown, Fox & Stine (2013), to study weekly vehicle theft in Bogotá, per locality. The model proposed by Aldor-Noiman *et al.* is an autoregressive first order process for Poisson data that uses a Dirichlet process to group localities with similar behaviour in the random component. These types of models are not exclusive to crime-analysis; for instance it can be used in sales prediction.

[1] Contact: slozano1@uc.cl, lozsandino@gmail.com

The document is divided in six sections: (i) data description, (ii) Bayesian model representation, (iii) procedure for parameter estimation, (iv) configuration and analysis of the simulation, (v) results, and (vi) conclusion and discussion.

I. Data

The data used here comes from the Crime Observatory of Colombia's National Police, which is publicly available. The dataset consists in crimes reported to the police, detailed by type of crime, date and place of occurrence. Additionally, I use the population projection by locality, reported by the District's Planning Office, because this factor may partly account for criminal incidents.

For the estimation of the parameters I used the weekly theft of vehicles occurred in 2015. On the other hand, to check the model's predictive performance I used information of 2016.

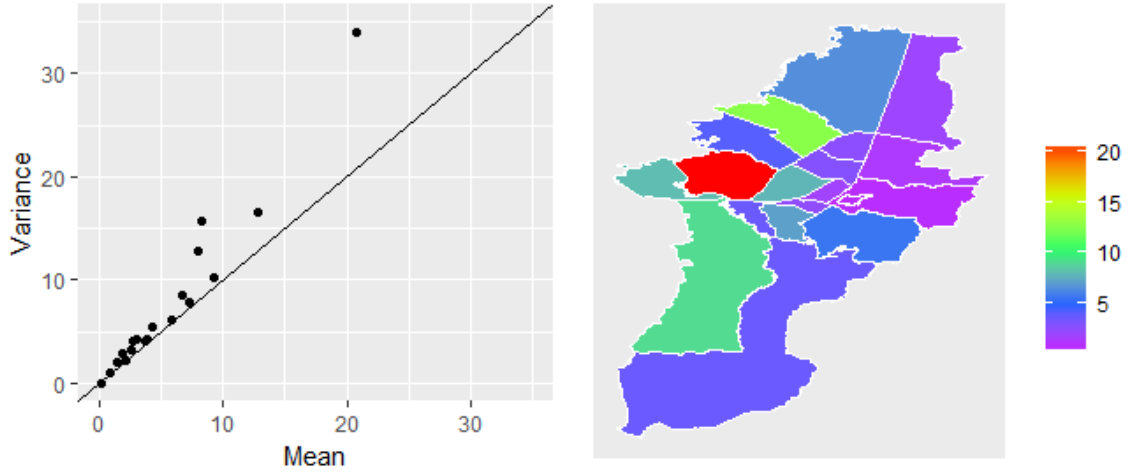

Figure 1. *Left*: average and variance of weekly vehicle theft by locality in 2015. *Right*: average weekly vehicle theft in 2015. Maps generated with geographic information of Unidad Administrativa Especial de Catastro, and using the package *ggmap*.

Figure 1 presents some of the characteristics of the dataset. The left panel shows that some localities present overdispersion. However, this overdispersion may occur due to mismeasurement; vehicle theft of the first and the last weeks of the year seem dubious compared to the values of the next and previous weeks. However, even if data doesn't have a Poisson-like behaviour, there should be no inconvenient because the model uses binomial thinning component.

II. Bayesian Model

The use of a Poisson autoregressive model is justified by the nature of the data used; weekly vehicle theft by locality is rather infrequent, and thus it's appropriated to use a model for discrete data. As it was show previously, the data used here has a Poisson-like behaviour. Though I used a first order autoregressive model, just as Aldor-Noiman *et al.*, it can easily be generalised to obtain a higher order autoregressive model.

The Poisson autoregressive model consists in two independent components: (i) the autoregressive component that measures the relationship between the current and the previous observations, and (ii) the random component.

The model used in this study is a modification of the one proposed by Aldor-Noiman *et al*. (2013), because here I omit a parameter that reflects the seasonal occurrence rate. There are three reasons behind this omission: (i) the estimation of the parameters is done using information of only one year, (ii) including this parameter in the model may increase the computational cost of the estimation, and (iii) its omission implies a simpler model, which serves the purpose of explaining how such models may be useful in the context of crime policy.

The Bayesian model is defined as:

$$y_{i,t} = \alpha_i \circ y_{i,t-1} + \epsilon_{i,t}, \qquad i = 1, \dots, L, \qquad t = 1, \dots, T$$

$$y_{i,t} - \epsilon_{i,t} | y_{i,t-1}, \alpha_i \overset{ind}{\sim} Bin(y_{i,t-1}, \alpha_i)$$

$$\alpha_i \overset{iid}{\sim} Beta(\eta_1, \eta_2)$$

$$\epsilon_{i,t} | X_i, \lambda_{z_i} \overset{ind}{\sim} Pois(X_i \lambda_{z_i})$$

$$\lambda_{z_i} \sim G$$

$$G \sim DP(\tau, G_0), \qquad G_0 = Gamma(\gamma_1, \gamma_2)$$

Where $\circ$ corresponds to the binomial thinning operator, $\alpha_i \circ y_{i,t-1} = \sum_j^{y_{i,t-1}} \phi_j(\alpha_i)$, with $\phi_j(\alpha_i)$ independent and identically distributed variables that follow a Bernoulli distribution with a probability of success of α_i. The parameter α_i is interpreted as a measure of positive correlation (because its possible values lie between 0 and 1). The $\alpha_i \circ y_{i,t-1}$ corresponds to the autoregressive component, while $\epsilon_{i,t}$ is the random component. The model assumes that there's independence between these two components.

The covariable X_i corresponds to the population of the locality, measured in 1,000 inhabitants. The parameter λ_{z_i} can be interpreted as the average rate of incidence, corrected by the population factor X_i. The value z_i is a parameter that allows associating locality i with other localities with similar random behaviour. For instance, localities i and h are said to have the same random behaviour if $z_i = z_h$. The parameter τ corresponds to the concentration parameter of the Dirichlet process, and is responsible for opening new clusters when performing the estimation.

III. Parameter Estimation

Aldor-Noiman *et al.* (2013) proposed a Markov Chain Monte Carlo algorithm to estimate the parameters. Here I present the MCMC used in the estimation of the parameters:

1. Sample the random component $\epsilon_{i,t}$, for $i = 1, \dots, L$ and $t = 1, \dots, T$:

   The simulated values of the random component are sampled form the posterior distribution of $\epsilon_{i,t}$, i.e.:

$$p\left(\epsilon_{i,t} | y_{i,t}, y_{i,t-1}, X_i, \alpha_i, \lambda_{z_i}\right) \propto \binom{y_{i,t-1}}{y_{i,t} - \epsilon_{i,t}} \alpha_i^{y_{i,t} - \epsilon_{i,t}} (1 - \alpha_i)^{y_{i,t-1} - (y_{i,t} - \epsilon_{i,t})} \times \frac{\left(X_i \lambda_{z_i}\right)^{\epsilon_{i,t}} \exp\{-X_i \lambda_{z_i}\}}{\epsilon_{i,t}!}$$

   This is not a recognisable distribution. However one can simulate the probability of each of the possible values of $\epsilon_{i,t}$, considering that $\max\{0, y_{i,t} - y_{i,t-1}\} \leq \epsilon_{i,t} \leq y_{i,t}$, and sample a value using these estimated probabilities.

2. Sample the parameter for cluster indicator z_i, for $i = 1, \dots, L$:

   This is done using the Dirichlet process. I used a Chinese restaurant process, which is a common representation of Dirichlet process. Also, I used a collapsed sampling, meaning that the posterior was marginalised with respect to the parameter λ_{z_i}. This gives a computational advantage, because it suppresses the need to simulate a value of λ_k each time the members of cluster k changes.

   To sample the value of z_i we can use the posterior distribution:

$$p(z_i = k | S_i, A_k, X_i, W_k) \propto \begin{cases} \tau \int_{\mathbb{R}_+} p(S_i | \lambda_0, X_i) G_0 d\lambda_0 \,, & \text{si } k = 0 \\ n_k \int_{\mathbb{R}_+} p(S_i | \lambda_k, X_i, A_k, W_k) G_0 d\lambda_k \,, & \text{si } k = 1, \dots, K \end{cases}$$

   Where,

$$S_i = \sum_{t=1}^{T} \epsilon_{i,t} \,, \qquad A_k = \sum_{h \neq i : z_h = k} S_h \,, \qquad W_k = \sum_{h \neq i : z_h = k} X_h$$

$$\int_{\mathbb{R}_+} p(S_i | \lambda_0, X_i) G_0 d\lambda_0 = \frac{\Gamma(S_i + \gamma_1)}{\Gamma(\gamma_1) S_i!} \left(\frac{TX_i}{TX_i + \gamma_2}\right)^{S_i} \left(\frac{\gamma_2}{TX_i + \gamma_2}\right)^{\gamma_1}$$

$$\int_{\mathbb{R}_+} p(S_i | \lambda_k, X_i, A_k W_k) G_0 d\lambda_k = \frac{\Gamma(S_i + A_k + \gamma_1)}{\Gamma(\gamma_1 + A_k) S_i!} \left(\frac{TX_i}{TX_i + TW_k + \gamma_2}\right)^{S_i} \left(\frac{\gamma_2 + TW_k}{TX_i + TW_k + \gamma_2}\right)^{\gamma_1 + A_k}$$

The marginalised posterior distributions are binomial negative. If $z_i = 0$, then the observation i is assigned to a new cluster, and the value of K is increased by one.

3. Sample the rate parameters λ_k:

   Once the clusters are assigned, one can simulate the values of λ_k, for $k = 1, \dots, K$, from the posterior distribution:

$$p(\lambda_k|B_k, V_k) \propto \frac{(TV_k\lambda_k)^{B_k}\exp\{-TV_k\lambda_k\}}{B_k!} \times \frac{\gamma_2^{\gamma_1}}{\Gamma(\gamma_1)}\lambda_k^{\gamma_1-1}\exp\{-\gamma_2\lambda_k\}$$
$$\propto \lambda_k^{B_k+\gamma_1-1}\exp\{-(TV_k+\gamma_2)\lambda_k\}$$

   One can easily identify the posterior, as $\lambda_k|B_k, V_k \sim Gamma(B_k+\gamma_1, TV_k+\gamma_2)$, where $B_k = \sum_{h:z_h=k} S_h$ and $V_k = \sum_{h:z_h=k} X_h$.

4. Sample α_i from $Beta\left(\sum_{t=2}^{T}(y_{i,t} - \epsilon_{i,t}) + \eta_1, \sum_{t=2}^{T}(y_{i,t-1} - y_{i,t} + \epsilon_{i,t}) + \eta_2\right)$, for $i = 1, \dots, L$.
5. Simulate the predicted value $\tilde{y}_{i,t'}$:

   Considering that there's conditional independence between observations in t', the posterior distribution of $\tilde{y}_{i,t'}$ is:

$$p(\tilde{y}_{i,t'}|y_{i,t'-1}, X_i, \alpha_i, \lambda_{z_i}) = \frac{\left(X_i\lambda_{z_i}\right)^{\tilde{y}_{i,t'}-\alpha_i\circ y_{i,t'-1}}\exp\{-X_i\lambda_{z_i}\}}{(\tilde{y}_{i,t'} - \alpha_i \circ y_{i,t'-1})!}, \qquad \tilde{y}_{i,t'} \geq \alpha_i \circ y_{i,t'-1}$$

   To do this, in each iteration a value $\alpha_i \circ y_{i,t'-1}$ is sampled from a $Bin(y_{i,t'-1}, \alpha_i)$, using the values of α_i and λ_{z_i}. In case that $y_{i,t'-1} = 0$, then $\alpha_i \circ y_{i,t'-1} = 0$. Else if the simulated value is $\tilde{y}_{i,t'} < \alpha_i \circ y_{i,t'-1}$, then $\tilde{y}_{i,t'} = \alpha_i \circ y_{i,t'-1}$.

IV. Simulation

The hyperparameters used for the simulation were $\gamma_1 = 1, \gamma_2 = 1, \eta_1 = 1$, y $\eta_2 = 1$. While the values of η_1 and η_2 don't offer much challenge, setting the values of γ_1 and γ_2 is much trickier. They need to be set considering that G_0 must not be a flat prior, because it would be a very strong assumption in these cases (Gelman *et al.*, 2014, p. 551). Also, it can be sensitive to the values of X_i; here I defined X_i as 1,000 inhabitants and $\gamma_1 = 1, \gamma_2 = 1$ work well with this definition, but it may change if X_i is defined as 100,000 inhabitants.

The parameter of concentration of the Dirichlet process was fixed at $\tau = 1$. This is a common practice in Dirichlet process models that favours a small number of clusters, while another practice is to assign a distribution to τ (Gelman *et al.*, 2014, p. 551).

The initial number of clusters was set as $K = L$, the same number of localities. The initial values of λ_i were set their empirical values, the weekly average vehicle theft, corrected by the factor X_i. The initial values of $\alpha_i$were selected randomly, sampling from $Unif(0, 1)$. A single chain of 5,000 iterations was created, burning out the first 1,000.

I'm interested in doing cluster analysis, so I need to apply a relabelling algorithm to alleviate the label switching problem. Gelman *et al*. (p. 543, 2014) list a handful of algorithms dedicated to this purpose. However, I opted to use an *ex post* hierarchical clustering on the partitions generated by the Chinese restaurant process sampling of each iteration. I'll give a brief summary about this algorithm, though I intend to show the details in another article.

The relabelling algorithm used in this study is an agglomerative conditional hierarchical clustering that uses centroid linkage and prioritises by the size of the cluster. It's conditional in the sense that no partition must be assigned to a cluster if that cluster already contains partitions belonging to the same iteration.

It's important to note that the clustering mentioned here is not the same as the mentioned previously, because this one is used for relabelling. The Chinese restaurant process generates a partition of the observations in each iteration (i.e. it generates clusters with these observations), while the hierarchical clustering algorithm is used to identify similar partitions across iterations (i.e. relabelling the clusters).

The analysis of this simulation showed that there no problems with the simulations of α_i; there were no autocorrelation problems and the traceplots showed good convergence.

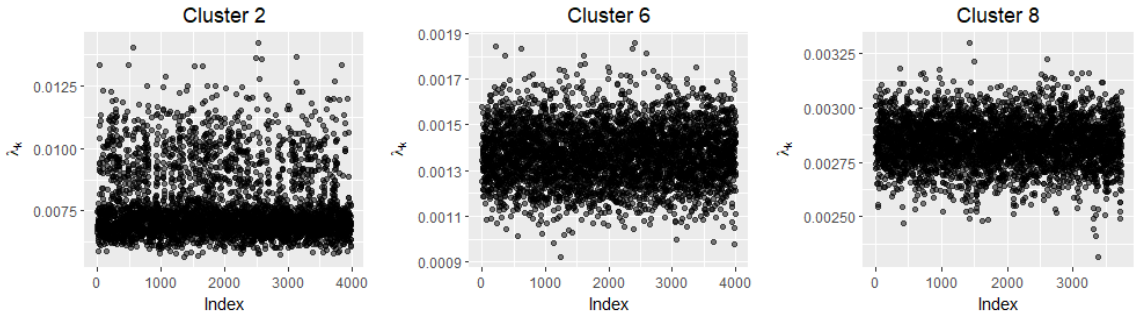

Figure 2. Traceplot of clusters 2, 6 and 8.

On the other hand, the convergence and autocorrelation of λ_k was analysed after the relabelling. There were no problems of autocorrelation and the traceplots showed good convergence. However, the analysis of convergence and autocorrelation of λ_k is unnecessary, because the sampling of these parameters don't depend on previous simulations. In other words, it can be interpreted as an independence chain, which doesn't suffer from autocorrelation nor convergence problems. The results of the simulation of α_i and λ_k are detailed in the annex section.

V. Results

Using relabelling method eight different clusters were identified. However, only three of these clusters appear as the posterior mode. The left panel of figure 3 shows the localities and its assigned cluster label.

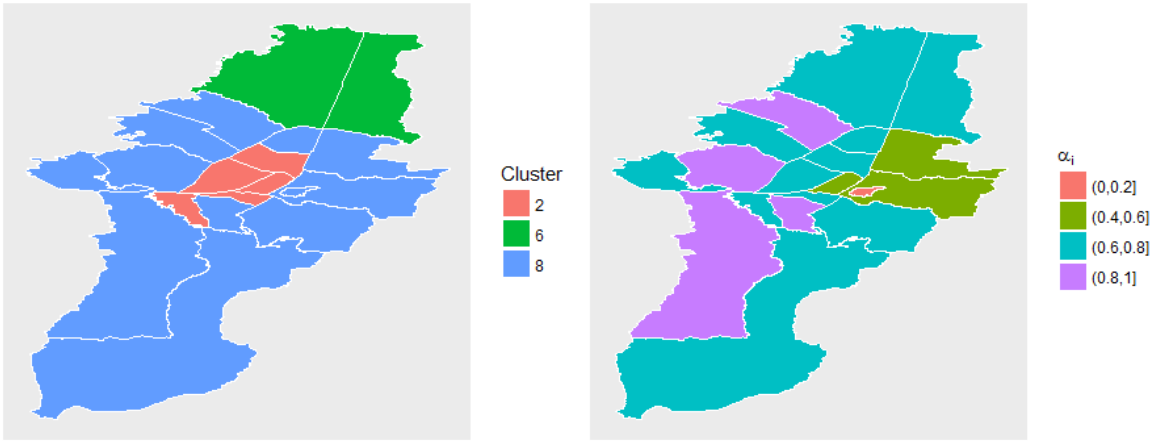

Figura 1. *Left*: posterior mode of the cluster assigned to the locality. *Right*: posterior mean of parameter $\boldsymbol{\alpha_i}$. Maps generated using *ggmap*.

The results of the posterior mode of the clusters indicate that localities in the north of Bogotá, Suba and Usaquén, may have similar behaviour, assigning these localities to cluster 6. On the other hand, localities in situated in the city centre were clustered together, assigning them to cluster 2. Also, analysing the mean posterior values of λ_2 and λ_6, this clustering produces intuitive results. The mean of the rate parameter of cluster 2 is $\lambda_2 = 0.00763$, in contrast with the estimated value of cluster 6, which is $\lambda_6 = 0.00137$. In other words, the incidence of vehicle theft is higher for localities in the city centre.

The simulation of α_i allows inferring about the level of correlation between the current and previous observations. Though most of the localities show high levels of correlation (higher that 0.6), some localities situated in the westernmost side of the city show the highest correlation.

One advantage of using a Bayesian approach is that, since we can obtain a posterior distribution via simulation, it's easy to check any summary statistics. Given the nature of the problem, one can use the 95 and 99 posterior predictive percentiles. The reason behind using such summary statistics is that institutions dedicated to preventing crime are commonly interested in predicting the worst case scenario. There's obviously a trade-off, since these statistics may produce overestimated values.

One way to check the model's predictive performance is calculating the proportion of observations that fall under the values of the 95 and 99 posterior predictive percentiles: 86.13% and 92% of the vehicle thefts occurred in 2016 fell below (or were equal to) the 95 and 99 posterior predictive percentile, respectively.

Generally the predictive performance of this model is quite good. Nevertheless there are some localities such as Kennedy and Puente Aranda, whose proportion of observations that fall under the 99 posterior predictive percentile is 57.69% and 71.15%.

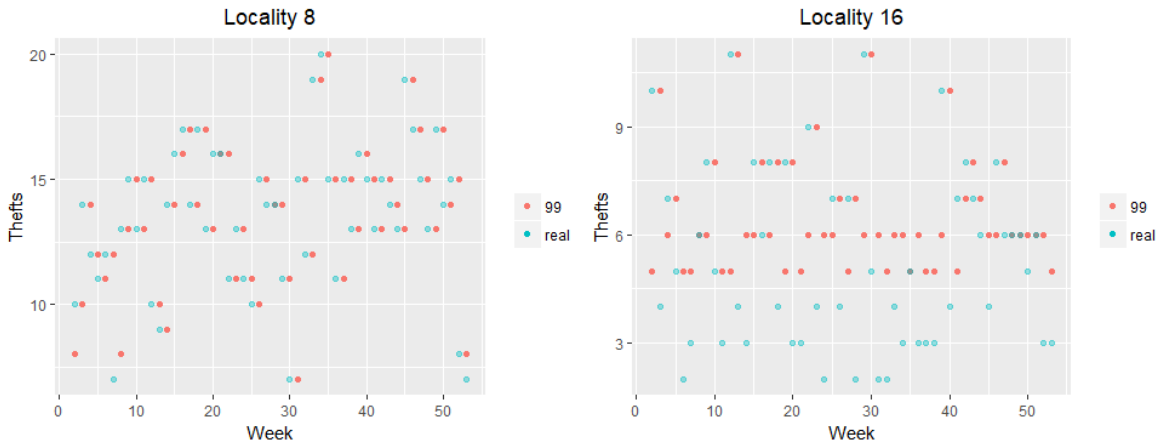

Figure 4. Weekly vehicle theft occurred in 2016 and 99 posterior predictive percentile, in Kennedy and Puente Aranda.

VI. Conclusions

One of the major advantages of this model is it's easy to interpret: it allows measuring temporal correlation and clustering localities with similar random behaviour. Without any further input and given the rest of the model's configuration, the Dirichlet process adjusts number of clusters that best fits the data and groups the localities. The parameter of λ_k can be interpreted as the estimated rate of vehicle theft per 1,000 inhabitants for localities belonging to cluster k.

Also, the model can be used to make predictions. Exploiting the fact that it's a Bayesian model, one can estimate the 99 percentile of the posterior predictive distribution and use it as the worst case scenario. The model shows a fair predictive performance; more often than not the real values do not surpass the predicted value.

However, one drawback of the model is that the parameter α_i only measures positive temporal correlation. This assumption may be quite restrictive in some cases. For instance, consider the case in which a new policy is implemented in a set of localities that diminish the number of thefts. One may believe that if there might exist a *balloon effect*, there would be displacement of criminal activity to other localities not affected by this policy. But another configuration of the model should be proposed to account for this kind of relationship between observations.

The model can be altered for other type of relationships between observations. For example, if one wants to account for an autoregressive process of a higher order, the following model can be estimated:

$$y_{i,t} = \alpha_i \circ y_{i,t-1} + \beta_i \circ y_{i,t-2} + \cdots + \omega_i \circ y_{i,t-d} + \epsilon_{i,t}$$

This method can also be used to estimate the correlation between other localities. Nevertheless, increasing the number of parameters of the model increases the computational cost, because it increases the size of the parameter space that the MCMC needs to explore.

Plus, the initial model configuration is flexible enough: each locality has its own correlation parameter, α_i.

As a final recommendation, this Poisson autoregressive model works well with low-count data. What's considered as low-count data depends on the level of spatial-temporal disaggregation. To make a similar analysis on more frequent events, it would be necessary to modify the model accordingly.

References

Aldor-Noiman, S., Brown, L.D., Fox, E.B., & Stine, R.A. (2013). *Spatio-Temporal Low Count Processes with Application to Violent Crime Events*. Retrieved September 27 of 2017, from the URL: https://arxiv.org/abs/1304.5642

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehatri, A., & Rubin, D.B. (2014). Bayesian Data Analysis. Third edition. CRC Press, Taylor & Francis Group.

Kahle, D., & Wickham, H. *ggmap: Spatial Visualization with ggplot2*. The R Journal, 5(1), 144-161. URL: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Observatorio del Delito – Policía Nacional de Colombia (n.d.). *Estudio Criminológico*. Retreived October 27 of 2017, from the URL: https://www.policia.gov.co/observatorio/estudio_criminologia

Secretaría Distrital de Planeación – Alcaldía Mayor de Bogotá (n.d.). *Reloj de población*. Retreived October 27 of 2017, from the URL: http://www.sdp.gov.co/portal/page/portal/PortalSDP/InformacionTomaDecisiones/Estadisticas/ProyeccionPoblacion:Proyecciones%20de%20Poblaci%F3n

Unidad Administrativa Especial de Catastro Distrital – Alcaldía Mayor de Bogotá (n.d.). *Mapa de referencia: localidad*. Retreived October 27 of 2017, from the URL: https://www.ideca.gov.co/es/servicios/mapa-de-referencia/tabla-mapa-referencia

Annex

Table 1. Summary of the simulation of λ_k, for $k = 1, \ldots, K$.

Cluster	Percentile 2.5	Median	Percentile 97.5	Mean	Size of the sample
1	0.00272	0.00649	0.01115	0.00632	56
2	**0.00618**	**0.00714**	**0.01121**	**0.00763**	**3,988**
3	0.00484	0.00596	0.00706	0.00597	1,027
4	0.00333	0.00417	0.00536	0.00426	162
5	0.00294	0.00336	0.00390	0.00338	289
6	**0.00114**	**0.00137**	**0.00162**	**0.00137**	**4,000**
7	0.00219	0.00249	0.00281	0.00249	261
8	**0.00262**	**0.00284**	**0.00305**	**0.00284**	**3,742**

Table 2. Summary of the simulation of α_i and the clusters assigned to each locality using the posterior mode, for $i = 1, \ldots, L$.

ID	Locality	α_i				Cluster
		Perc. 2.5	Median	Perc. 97.5	Mean	
1	Usaquén	0.54	0.64	0.72	0.63	6
2	Chapinero	0.47	0.58	0.69	0.58	8
3	Santa Fe	0.37	0.50	0.63	0.50	8
4	San Cristóbal	0.72	0.77	0.82	0.77	8
5	Usme	0.63	0.69	0.75	0.69	8
6	Tunjuelito	0.65	0.72	0.78	0.72	2
7	Bosa	0.70	0.75	0.78	0.75	8
8	Kennedy	0.86	0.88	0.90	0.88	8
9	Fontibón	0.68	0.74	0.80	0.74	8
10	Engativá	0.80	0.83	0.86	0.83	8
11	Suba	0.73	0.77	0.82	0.77	6
12	Barrios Unidos	0.65	0.72	0.79	0.72	8
13	Teusaquillo	0.57	0.65	0.72	0.65	2
14	Los Mártires	0.38	0.48	0.57	0.48	2
15	Antonio Nariño	0.54	0.63	0.71	0.63	2
16	Puente Aranda	0.76	0.80	0.83	0.80	2
17	Candelaria	0.00	0.11	0.46	0.14	8
18	Rafael Uribe Uribe	0.77	0.81	0.84	0.81	8
19	Ciudad Bolívar	0.79	0.82	0.86	0.82	8