Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Santiago Lozano Sandino

lozsandino@gmail.com

lozsandino.github.io

Modelo Poisson autorregresivo semiparamétrico aplicado al hurto de vehículos en Bogotá

Índice

Introducción

- Interés de predecir y entender el delito en áreas de política criminal:
  - Policía Nacional de Colombia
  - Ministerio de Justicia y del Derecho
  - Secretaría Distrital de Seguridad, Convivencia y Justicia
- Algunos delitos son poco frecuentes; en esos casos es mejor usar modelos para datos discretos.
- Modelo Poisson Autorregresivo (Aldor-Noiman *et al.*, 2013):
  - Componente de dependencia temporal.
  - Componente aleatorio.
- El proceso Dirichlet se utiliza para agrupar localidades con comportamiento aleatorio similar.

Introducción

- La inferencia bayesiana utiliza en teorema de Bayes para obtener información sobre los parámetros:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$

- En tal caso los parámetros se tratan como variables aleatorias, que tienen una distribución.
- Para definir el modelo bayesiano se define la relación entre las observaciones y los parámetros.

$$\boldsymbol{y}|\boldsymbol{\theta} \sim f(\boldsymbol{y}|\boldsymbol{\theta})$$
$$\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$$

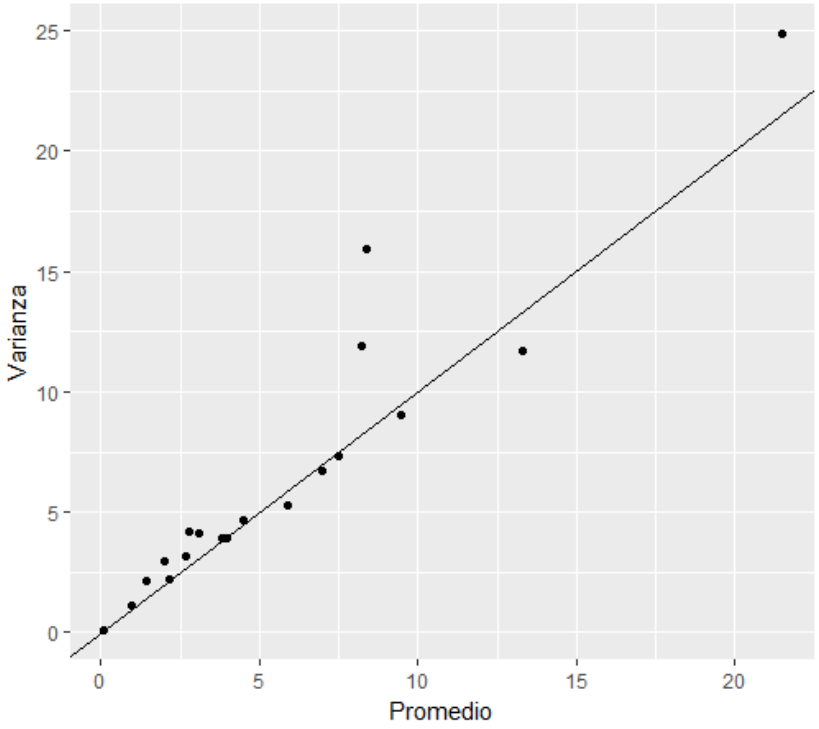
Datos

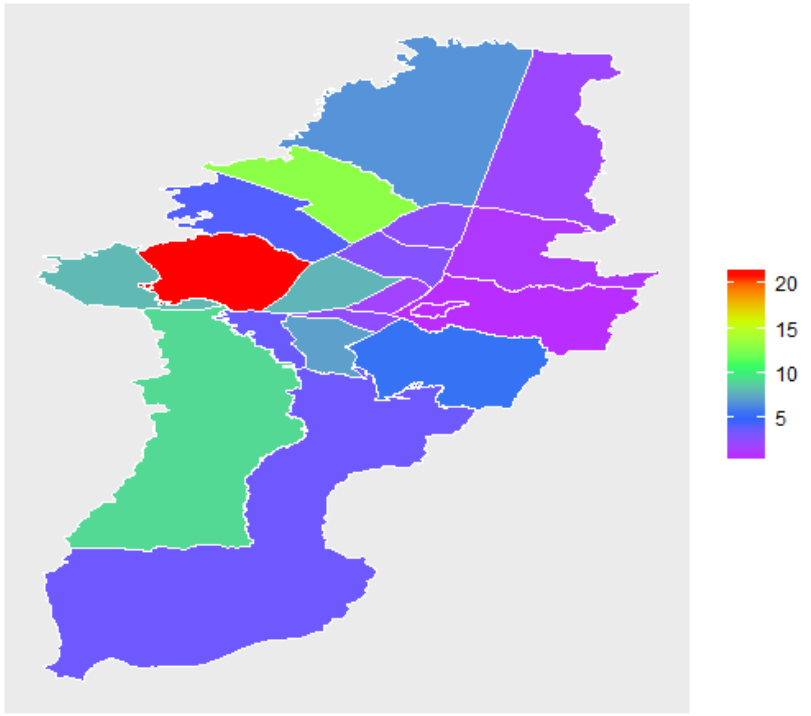
- Base de entrenamiento: Hurto de vehículos semanales, para 19 localidades (1,007 observaciones en total), correspondientes al año 2015.
- Población por localidad.
- Base para evaluar la predicción del modelo: 19 localidades y 52 semanas, correspondientes al año 2016.

Datos

Código	Localidad	Población	Hurtos 2015	Hurtos 2016
1	Usaquén	494,066	110	67
2	Chapinero	137,870	75	25
3	Santa Fe	110,053	48	16
4	San Cristóbal	406,025	306	118
5	Usme	432,724	199	77
6	Tunjuelito	200,048	201	102
7	Bosa	646,833	434	104
8	Kennedy	1,069,469	1098	703
9	Fontibón	380,453	226	83
10	Engativá	874,755	676	338
11	Suba	1,174,736	352	201
12	Barrios Unidos	240,960	144	62
13	Teusaquillo	151,092	157	75
14	Los Mártires	98,758	101	46
15	Antonio Nariño	108,941	136	75
16	Puente Aranda	258,414	422	276
17	Candelaria	24,096	5	6
18	Rafael Uribe Uribe	375,107	384	148
19	Ciudad Bolívar	687,923	487	127
20	Sumapaz	6,460	0	0

Datos





Izquierda: hurto semanal promedio de vehículos frente a varianza, por localidad con información de 2015. *Derecha:* hurtos semanales promedio, por localidad con información de 2015. Mapas realizados con información geográfica de la Unidad Administrativa Especial de Catastro Distrital y el paquete *ggmap.*

Modelo bayesiano

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

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

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

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

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

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

Modelo bayesiano

El parámetro z_i permite asociar la localidad i con otras localidades con similar comportamiento aleatorio.

El parámetro λ_{z_i} puede ser interpretado como la tasa promedio de incidencia de hurto de vehículos, independiente de la población de la localidad, X_i.

El parámetro α_i puede interpretarse como una medida de correlación positiva (pues toma valores entre 0 y 1) entre la observación actual y la anterior.

Procedimiento para la estimación

1. Simular $\epsilon_{i,t} \sim p\left(\epsilon_{i,t} | y_{i,t}, y_{i,t-1}, X_i, \alpha_i, \lambda_{z_i}\right) \propto$

$$p\left(y_{i,t} - \epsilon_{i,t} | y_{i,t-1}, \alpha_i\right) \times p\left(\epsilon_{i,t} | X_i, \lambda_{z_i}\right),$$

Lo anterior no resulta en una distribución conocida, sin embargo se puede calcular la probabilidad para cada uno de los valores de $\epsilon_{i,t}$, t.q. $\max\left\{0, y_{i,t} - y_{i,t-1}\right\} \leq \epsilon_{i,t} \leq y_{i,t}$, y se simula a partir de ahí.

Procedimiento para la estimación

2. Simular z_i:

$$p(z_i = k | S_i, A_k, X_i, W_k, \gamma_1, \gamma_2, \tau)$$

$$\propto \begin{cases} \tau\, p(S_i | X_i, \gamma_1, \gamma_2), & \text{si } k = 0 \\ n_k p(S_i | X_i, A_k W_k, \gamma_1, \gamma_2), & \text{si } k = 1, \dots, K \end{cases}'$$

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

La simulación se realiza con distribución categórica, con las probabilidades de los eventos dadas por $\tau\, p(S_i | X_i, \gamma_1, \gamma_2)$ y $n_k p(S_i | X_i, A_k W_k, \gamma_1, \gamma_2)$.

Procedimiento para la estimación

El propósito de marginalizar las expresiones respecto al parámetro de tasa es simplificar la expresión de tal manera que no sea necesario simular la actualización de λ_k cada vez que se simula z_i. Es decir:

$$\tau\, p(S_i|X_i,\gamma_1,\gamma_2) = \tau \int_{\mathbb{R}_+} p(S_i|\lambda_0,X_i,\gamma_1,\gamma_2) G_0 d\lambda_0\,,$$

$$n_k p(S_i|X_i,A_kW_k,\gamma_1,\gamma_2) = n_k \int_{\mathbb{R}_+} p(S_i|\lambda_k,X_i,A_kW_k,\gamma_1,\gamma_2) G_0 d\lambda_k$$

Procedimiento para la estimación

3. Simular $\lambda_k|B_k, V_k, \gamma_1, \gamma_2 \sim Gamma(B_k + \gamma_1, TV_k + \gamma_2)$,

$$B_k = \sum_{h:z_h=k} S_h \, , V_k = \sum_{h:z_h=k} X_h$$

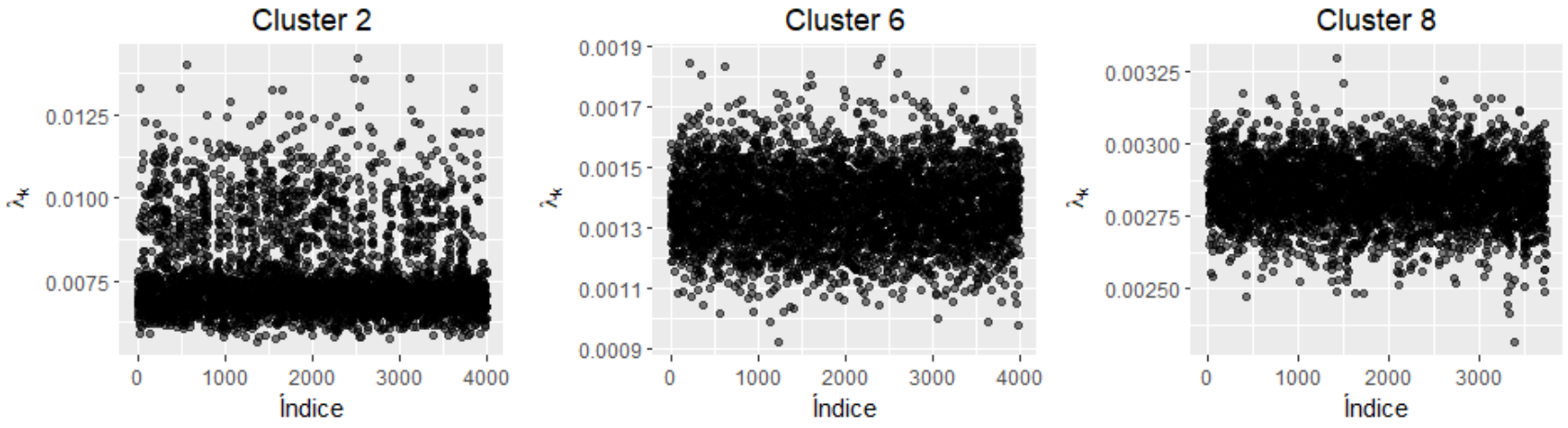
4. Simular $\alpha_i \sim Beta\left(\sum_{t=2}^{T}\left(y_{i,t} - \epsilon_{i,t}\right) + \eta_1, \sum_{t=2}^{T}\left(y_{i,t-1} - y_{i,t} + \epsilon_{i,t}\right) + \eta_2\right)$.

5. Simular $\tilde{y}_{i,t'}|y_{i,t'-1}, X_i, \alpha_i, \lambda_{z_i} \sim Pois\left(X_i \lambda_{z_i}\right)$, $\tilde{y}_{i,t'} \geq \alpha_i \circ y_{i,t'-1}$, luego de simular $\alpha_i \circ y_{i,t'-1} \sim Bin\left(y_{i,t'-1}, \alpha_i\right)$. Si el valor simulado $\tilde{y}_{i,t'} < \alpha_i \circ y_{i,t'-1}$, entonces $\tilde{y}_{i,t'} = \alpha_i \circ y_{i,t'-1}$.

Resumen de la simulación

- Los hiperparámetros usados fueron $\gamma_1 = 1, \gamma_2 = 1, \eta_1 = 1, \eta_2 = 1,$ y $\tau = 1$.
- Una cadena de 5,000 iteraciones, quemando las 1,000 primeras.
- Procedimiento *ex post* para las etiquetas: cluster jerárquico aglomerativo de las particiones, con mínima distancia al centroide, condicionado a que no se unieran particiones que pertenecieran a la misma iteración y priorizando el orden del enlace por tamaño del cluster.

Resumen de la simulación



Traceplot de los clusters 2, 6 y 8, para el análisis de convergencia.

Resultados

Resumen de la simulación de λ_k para cada uno de los clusters estimados, para $k = 1, \dots, K$.

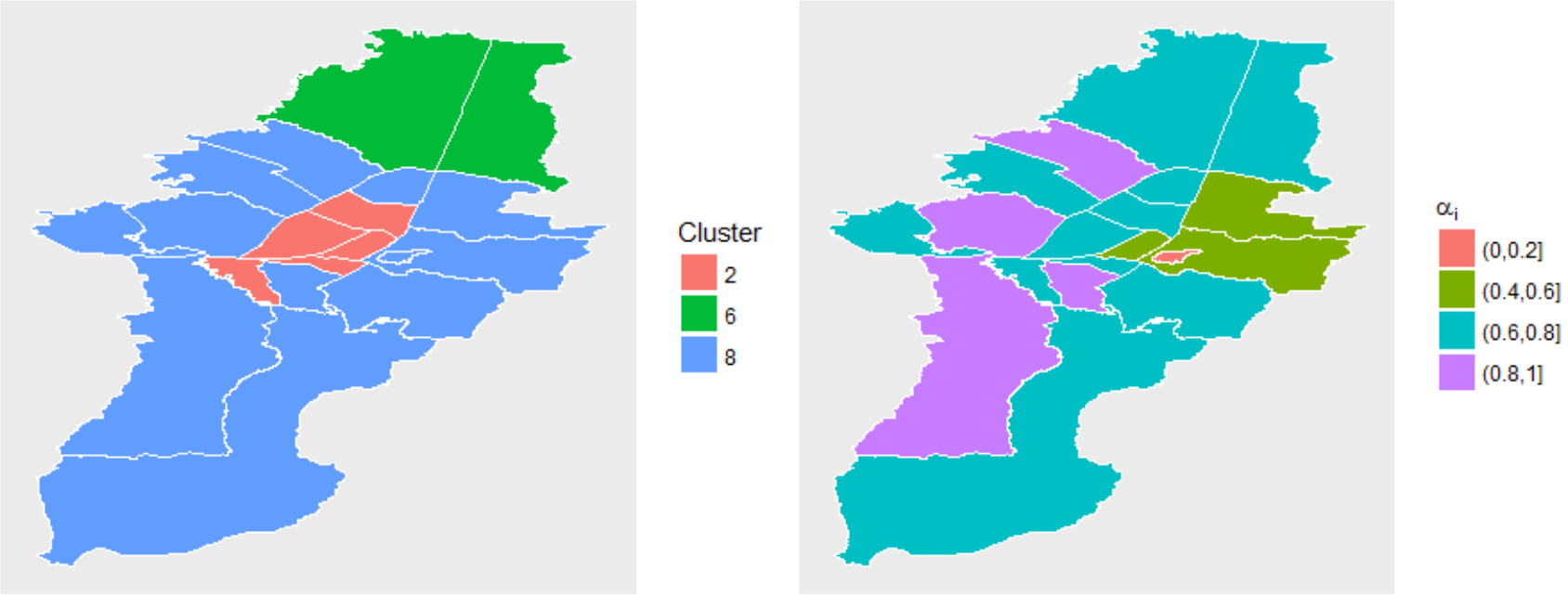
Cluster	Percentil 2.5	Mediana	Percentil 97.5	Promedio	Tamaño de muestra
1	0.00272	0.00649	0.01115	0.00632	56
2	**0.00618**	**0.00714**	**0.01121**	**0.00763**	**3,988**
3	0.00484	0.00596	0.00706	0.00597	1,027
4	0.00333	0.00417	0.00536	0.00426	162
5	0.00294	0.00336	0.0039	0.00338	289
6	**0.00114**	**0.00137**	**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**

Resultados

Resumen de la simulación de $\alpha_i$para cada una de las localidades y el cluster al que fueron asignados.

ID	Localidad	α_i				Cluster
		Perc. 2.5	Median	Perc. 97.5	Prom.	
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

Resultados

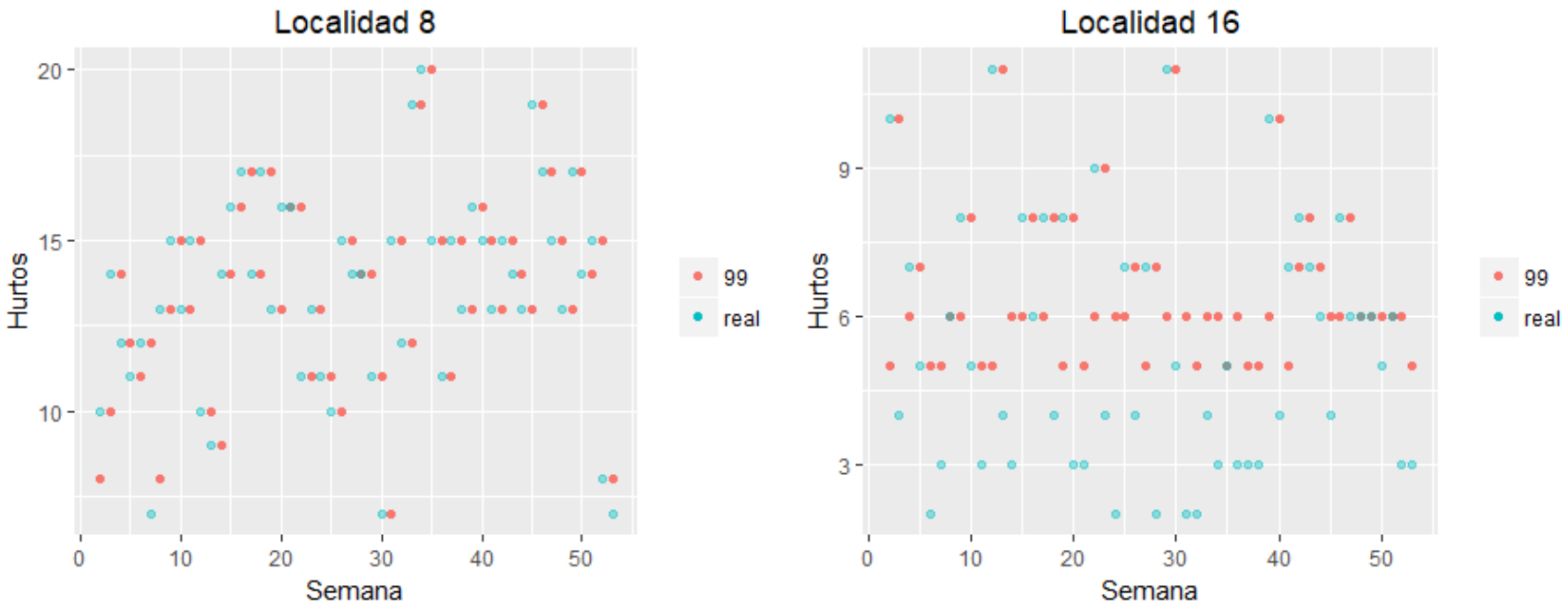


Izquierda: moda a posteriori del cluster asignado a la localidad. En detalle aquí.
Derecha: promedio a posteriori del parámetro α_i. En detalle aquí.

Resultados

- Se puede utilizar el percentil 99 a posteriori para hacer predicciones.
- En general 92% de las observaciones de dentro del rango del percentil 99.
- En algunas localidades, el desempeño de la predicción era bastante más bajo: Kennedy y Puente Aranda, la proporción observaciones dentro del rango del percentil 99 es 57.69% y 71.15%, respectivamente

Resultados



Hurtos semanales en 2016 y el percentil 99 predictivo a posteriori, para las localidades de Kennedy y Puente Aranda. En detalle aquí.

Conclusiones y discusión

- El modelo es de fácil interpretación y permite hacer predicciones.
- En general tiene buen desempeño predictivo, excepto en algunas localidades.
- Limitaciones:
  - Solo considera correlación positiva.
  - No considera autocorrelación espacial.
  - Autorregresivo de primer orden.
- Es necesario comparar el desempeño del modelo con otros modelos más sencillos.

Referencias

- Aldor-Noiman, S., Brown, L.D., Fox, E.B., & Stine, R.A. (2013). *Spatio-Temporal Low Count Processes with Application to Violent Crime Events*. Recuperado el 27 de septiembre de 2017 de la página web: 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 (s.f.). *Estudio Criminológico*. Recuperados el 27 de octubre de 2017 de la página web: https://www.policia.gov.co/observatorio/estudio_criminologia
- Secretaría Distrital de Planeación – Alcaldía Mayor de Bogotá (s.f.). *Reloj de población*. Recuperado el 27 de octubre de 2017 de la página web: 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 (s.f.). *Mapa de referencia: localidad*. Recuperado el 27 de octubre de 2017 de la página web: https://www.ideca.gov.co/es/servicios/mapa-de-referencia/tabla-mapa-referencia

¡Muchas gracias!